Universal Source Coding in the Non-Asymptotic Regime

by

Nematollah Iri

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved January 2018 by the
Graduate Supervisory Committee:

Oliver Kosut, Chair
Daniel Bliss
Lalitha Sankar
Junshan Zhang

ARIZONA STATE UNIVERSITY

May 2018

ABSTRACT

Fundamental limits of fixed-to-variable (F-V) and variable-to-fixed (V-F) length universal source coding at short blocklengths is characterized. For F-V length coding, the Type Size (TS) code has previously been shown to be optimal up to the third-order rate for universal compression of all memoryless sources over finite alphabets. The TS code assigns sequences ordered based on their type class sizes to binary strings ordered lexicographically.

Universal F-V coding problem for the class of first-order stationary, irreducible and aperiodic Markov sources is first considered. Third-order coding rate of the TS code for the Markov class is derived. A converse on the third-order coding rate for the general class of F-V codes is presented which shows the optimality of the TS code for such Markov sources.

This type class approach is then generalized for compression of the parametric sources. A natural scheme is to define two sequences to be in the same type class if and only if they are equiprobable under any model in the parametric class. This natural approach, however, is shown to be suboptimal. A variation of the Type Size code is introduced, where type classes are defined based on neighborhoods of minimal sufficient statistics. Asymptotics of the overflow rate of this variation is derived and a converse result establishes its optimality up to the third-order term. These results are derived for parametric families of i.i.d. sources as well as Markov sources.

Finally, universal V-F length coding of the class of parametric sources is considered in the short blocklengths regime. The proposed dictionary which is used to parse the source output stream, consists of sequences in the boundaries of transition from low to high quantized type complexity, hence the name Type Complexity (TC) code. For large enough dictionary, the ϵ-coding rate of the TC code is derived and a converse result is derived showing its optimality up to the third-order term.

TABLE OF CONTENTS

CHAPTER Page

Chapter 1

INTRODUCTION

Data compression is intrinsic to human desire to describe the world without and hence closely related to the very beginning question of the mankind: "Having observed a phenomenon, describe it". Jorma Rissanen's minimum description length principle Rissanen (1986) (which is a formalization of Occam's razor principle), measures the quality of the description based on the code length with which the data can be described. The appealing intuition is that, the more regularities in the data permit shorter code lengths.

Given statistics of the underlying source generating the data, Shannon Shannon (1948) showed that the entropy of the source is the fundamental asymptotic limit (as the blocklength approaches infinity) for the code length. In consequence of this, a new theory arose: *the source coding theory*. Apart from its very own bearing, the theory has relevance to many other articulations such as the estimation problems Rissanen (1984); Weinberger *et al.* (1994); Ryabko (2009), prediction problems Rissanen (1984); Merhav and Feder (1998); Ryabko (2009, 1988); Ziv (2002), detection problems Merhav (2014); Heydari *et al.* (2016a), inference problems Wainwright and Jordan (2008); Heydari and Tajer (2017); Heydari *et al.* (2016b), classification problems Ziv (2008), etc.

In the traditional source coding doctrine, performance of algorithms are characterized in the limit of large blocklengths. In some modern applications, however, data is continuously generated and updated, making them highly delay-sensitive. Therefore, it is vital to characterize the overheads associated with operation in the short blocklength regime.

To evaluate the performance of source coding for blocklengths at which the law of large numbers does not apply, we need a more refined metric than expected length. Thus, we use ϵ-coding rate, the minimum rate such that the corresponding overflow probability is less than ϵ. Fundamental limits of ϵ-coding rate for fixed-to-variable (F-V) lossless data compression in the non-universal setup are derived in Kontoyiannis and Verdú (2014), both for *i.i.d.* as well as Markov sources. In most applications, however, the statistics of the source are unknown or arduous to estimate, especially at short blocklengths, where there are constraints on the available data for the inference task. In the universal setup, a class of models is given, however the true model in the class that generates the data is unknown. From an algorithmic angle, the aim of universal source coding is to propose a compression algorithm in which the encoding process is ignorant of the underlying unknown parameters, yet achieving the performance criteria.

Analysis of the finite blocklength behavior as well as fine asymptotics of universal source coding have been considered in Kosut and Sankar (2013, 2014b,a); Tan (2014) for the class of *i.i.d.* sources. Similar to the aforementioned works, the universal source coding scheme in this paper compresses the whole file. Therefore, we relax the prefix condition Szpankowski and Verdu (2011), and hence the coding scheme is called the one-to-one code. Imposing the prefix free condition, the ϵ-coding rate of the Two Stage code Kosut and Sankar (2014b,a) and that of the Bayes code Saito *et al.* (2014, 2015) are also considered in the literature.

The Type Size code (TS code) is introduced in Kosut and Sankar (2013) for compression of the class of *all* stationary memoryless sources, in which sequences are encoded in increasing order of type class size. It is shown that the resulting third-order term is $\frac{|\mathcal{X}|-3}{2}\log n$ bits, where $|\mathcal{X}|$ is the alphabet size. Its optimality is shown in Kosut and Sankar (2014b). Subsequently, a converse bound is derived in

Beirami and Fekri (2014) for one-to-one average minimax (and maximin) redundancy of memoryless sources, which consequently shows that the TS code is optimal up to $o(\log n)$ for universal one-to-one compression of *all* memoryless sources, considering expected length as the performance metric Beirami and Fekri (2014). However, an achievable scheme for universal one-to-one compression of parametric sources with more *structure* is not provided.

We first consider the universal F-V length compression problem in the finite block-length regime for the class of first-order stationary, irreducible and aperiodic Markov sources. We provide performance guarantees for the Type Size code for this model class. Using the Type Size code, we show that the minimal number of bits required to compress a length n sequence with probability $1-\epsilon$ is at most

$$nH(X_2|X_1)+\sigma\sqrt{n}Q^{-1}(\epsilon)+\left(\frac{|\mathcal{X}|(|\mathcal{X}|-1)}{2}-1\right)\log n+\mathcal{O}(1) \tag{1.1}$$

where, $H(X_2|X_1)$ is the conditional entropy of the source, σ is a generalization of the varentropy to the Markov case, $Q(\cdot)$ is the tail of the standard normal distribution and $|\mathcal{X}|$ is the alphabet size. The first two terms in (1.1) are the same as the non-universal case Kontoyiannis and Verdú (2014), while the third-order $\log n$ term represents the cost of universality. We further provide a converse, showing that ϵ-coding rate of any universal code is lower bounded by (1.1), thus proving that the Type Size code does at least as well as any universal code for compressing the class of Markov sources. The proof involves two new ideas compared to the *i.i.d.* case: (i) We develop tight bounds on the size of a Markov type class by relating it to Eulerian cycles on a directed graph and using the BEST theorem (see Lemma 2). (ii) We use a Markov version of the multidimensional Berry-Esseen theorem Lapinskas (1974), to derive upper and lower bounds on the tail of the distribution of empirical entropy.

We next consider compression of more *structured* parametric sources. Type classes

in Kosut and Sankar (2013); Iri and Kosut (2015); Beirami and Fekri (2014) are based on the empirical probability mass function (EPMF). In particular, two sequences are in the same (elementary) type class if they have the same EPMF. Elementary type classes do not exploit the inherited structure in the model class. To generalize the notion of a type to richer model classes, we define the *point* type class as the set of sequences equiprobable under any model in the class. The size of the point type class structure is analyzed in Merhav and Weinberger (2004). This natural characterization of type classes is based on the philosophy that the sequences with the same probability (under any model in the class) are "*indistinguishable*". Such a philosophy has been employed before in the relevant applications, e.g. the universal simulation Merhav and Weinberger (2004) and the universal random number generation Seroussi and Weinberger (2015) problems. Perhaps surprisingly, we show that this natural approach is suboptimal for the universal source coding problem. In this thesis, we characterize the structure of the type classes in a new fashion for the sake of optimally compressing exponential families of distributions. We refer to this new approach as *quantized* types. We divide the convex hull of the set of minimal sufficient statistics into cuboids. Two sequences are in the same quantized type class if their minimal sufficient statistics belong to the same cuboid. Therefore, we show that *approximate* indistinguishability leads to optimality for the source coding problem.

We consider F-V length codes for a d-dimensional exponential family of distributions over a finite alphabet $\mathcal{X}$. For ease of exposition, we first assume, data generated by the unknown true model in this family is independent and identically distributed ($i.i.d.$). We subsequently extend the results to Markov data generation mechanisms. We provide performance guarantees for the Type Size code for these model classes. Using the Type Size code, we show that the minimal number of bits required to

compress a length-n sequence with probability $1-\epsilon$ is at most

$$nH + \sigma\sqrt{n}Q^{-1}(\epsilon) + \left(\frac{d}{2} - 1\right)\log n + \mathcal{O}(1) \tag{1.2}$$

where H and σ^2 are the entropy and varentropy of the underlying source respectively, $Q(\cdot)$ is the tail of the standard normal distribution and d is the dimension of the model class. The second-order term is the payoff due to operation in the short blocklengths while the third-order $\log n$ term represents the cost of universality. Precise bounds on the fourth-order $\mathcal{O}(1)$ term is beyond the scope of this thesis. However, analyzing the fourth-order term is considered in the literature for the related source coding problems. For example, it is shown in Szpankowski (2008) that the fourth-order term is either a constant or has fluctuating behavior for average codelength of a binary memoryless source.

Finally we consider universal variable-to-fixed (V-F) length compression of the parametric class in the short blocklength regime. A V-F length code consists of a dictionary of pre-specified size. Elements of the dictionary (segments) are used to parse the infinite sequence emitted from the source. Segments may have variable length, however they are encoded to the fixed-length binary representation of their indices within the dictionary. In order to be able to uniquely parse any infinite sequence into the segments, we assume the dictionary to be complete (i.e. every infinite length sequence has a prefix within the dictionary) and proper (i.e. no segment is a prefix of another segment).

For a given memoryless source, Tunstall Tunstall (1967) provided an average-case optimal algorithm to maximize average segment length. A central limit theorem for the Tunstall algorithm's code length has been derived in Drmota *et al.* (2010). Universal V-F length codes are studied in e.g. Krichevsky and Trofimov (1981); Lawrence (1977); Tjalkens and Willems (1992); Visweswariah *et al.* (2001). Upper

and lower bounds for the *redundancy* of a universal code for the class of *all* memoryless sources is derived in Krichevsky and Trofimov (1981). Universal V-F length coding for the class of all binary memoryless sources is then considered in Lawrence (1977); Tjalkens and Willems (1992), where Tjalkens and Willems (1992) provides an asymptotically average sense optimal [1] algorithm. Later, optimal redundancy for V-F length compression of the class of all Markov sources is derived in Visweswariah *et al.* (2001). In an attempt to compare performance of V-F length codes with F-V length codes for compression of the class of all Markov sources, a dictionary construction that asymptotically achieves the optimal error exponent is proposed in Merhav and Neuhoff (1992). All previous works, consider model classes that include all distributions within a simplex. However, universal V-F length coding for a *more structured* model classes has not been considered in the literature. In this thesis, we characterize asymptotics of the ϵ-coding rate for large enough dictionaries. We provide an achievable scheme for compressing d-dimensional exponential family of distributions as the parametric model class. We then provide a converse result, showing that our proposed scheme is optimal up to the third-order ϵ-coding rate.

In previous universal V-F length codes, one can define a notion of complexity for sequences. In Krichevsky and Trofimov (1981); Lawrence (1977); Tjalkens and Willems (1992); Visweswariah *et al.* (2001), a sequence with high complexity has low probability under a certain composite or mixture source. While in Merhav and Neuhoff (1992), high complexity sequences have high scaled (by sequence length) empirical entropy. The dictionary of such algorithms then consists of sequences in the boundaries of transition from low complexity to high complexity. We follow similar complexity theme to design the dictionary. The sequence complexity in our

[1] Throughout, "optimality" of an algorithm is considered only up to the model cost term in the coding rate.

proposed coding scheme is synonymous to scaled type class size, hence we call it *Type complexity* (TC) code. Scaled empirical entropy Merhav and Neuhoff (1992) is ignorant of the underlying structure of the model class. Therefore, in order to *fully* exploit the inherited structure of the model class, we characterize type classes based on quantized types Iri and Kosut (2016a,b).

We provide performance guarantee for V-F length compression of the exponential family using our proposed Type Complexity code. We upper bound the ϵ-coding rate of the quantized type implementation of the Type Complexity code by

$$H + \sigma\sqrt{\frac{H}{\log M}}Q^{-1}(\epsilon) + H\frac{d}{2}\frac{\log\log M}{\log M} + \mathcal{O}\left(\frac{1}{\log M}\right)$$

where H, σ^2 are the entropy and the varentropy of the underlying source, respectively, M is the pre-specified dictionary size, $Q(\cdot)$ is the tail of the standard normal distribution, and d is the dimension of the model class. We then provide a converse result showing that this rate is optimal up to the third-order term. Our converse proof relies on the connection between V-F length and F-V length codes observed in Merhav and Neuhoff (1992), along with a converse result for F-V length prefix codes Kosut and Sankar (2014a).

The rest of the thesis is organized as follows: In Chapter 2, we consider the universal F-V length coding of the class of Markov sources. We introduce the finite-length lossless source coding problem and the related definitions in Sec. 2.1. In Sec. 2.2 we present the main theorem of the Chapter. We present a lemma bounding the size of a Markov type class in Sec. 2.3. The bounds on the tail of the empirical entropy along with other preliminary results are also provided in Sec. 2.3. We sketch the proof of main theorem in Sec. 2.4. We next consider the compression problem for the parametric family in Chapter 3. We introduce the exponential family and related

definitions in Section 3.1. In Section 3.2, we describe quantized type classes and the variation of the TS code used for optimal performance. In Section 3.3, we present the main theorem of the chapter, which characterizes the performance of the TS code using quantized type classes up to third order. We present preliminary results including a lemma bounding the size of a type class in Section 3.4. We provide the proof of main theorem in Section 3.5. Extensions to the Markov case is considered in Section 3.6. We show the suboptimality of the approach based on point type classes in Section 3.7. Finally the V-F length coding is considered in Chapter 4. In Sec. 4.1, we introduce the exponential family, V-F length coding and related definitions. Type complexity algorithm is presented in Sec. 4.2. Main result of the chapter is stated in Sec. 4.3. We present preliminary results in Sec. 4.4. The achievability and the converse results are proved in Sec.'s 4.5 and 4.6, respectively. We conclude, in Chapter 5. A number of proofs are given in the appendices.

Chapter 2

MARKOV SOURCE

2.1 Preliminaries

Notation: We use $\mathbb{P}$ (resp. $\mathbb{E}$) to denote probability (resp. expectation) with respect to the Markov source parameterized by $\hat{p}$. All logarithms are with respect to base 2. Let $\mathcal{M}^n$ to be the set of first-order stationary, irreducible and aperiodic Markov distributions on n-length sequences over the finite alphabet $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$. Let $\hat{\mathcal{P}}$ be the set of transition distributions $\hat{p}(\cdot|\cdot)$ on $\mathcal{X}$, which parametrize the class of Markov sources assuming they are stationary, irreducible and aperiodic. Let p be the stationary distribution (one letter marginal) of $\hat{p}$ and we denote two-letter marginal as $\tilde{p}(x, y) = p(y) \cdot \hat{p}(x|y),$ for $x, y \in \mathcal{X}$. We consider a universal source coding problem in which a single code must compress a sequence X^n that is an output of a Markov source in such a class. The true model $\hat{p}(\cdot|\cdot) \in \hat{\mathcal{P}}$ that generates the data is unknown. For a sequence $x^n \in \mathcal{X}^n$, we use the following notations: $n_i = |\{s \in [n] : x_s = i\}|$ is the number of occurrences of the symbol $i \in \mathcal{X}$ in the string $x^n = x_1 \ldots x_n$, where $[n] = \{1, \ldots, n\}$. Likewise,

$$n_{ij} = |\{s \in [n] : (x_s, x_{s+1}) = (i, j)\}|$$

is the number of occurrences of consecutive (i, j) in x^n. Note that we adopt cyclic convention, i.e. $x_{n+1} = x_1$. Let q_{x^n} be the type of x^n. Hence we have $q_{x^n}(i) = \frac{n_i}{n}$. We also define the joint type $\tilde{q}_{x^n}$ over $\mathcal{X}^2$, as follows: $\tilde{q}_{x^n}(i, j) = \frac{n_{ij}}{n}$. We frequently omit the subscript x^n, whenever it is understood from the context. Let T_q and $T_{\tilde{q}}$, be the type class of q and $\tilde{q}$, respectively, i.e. $T_q = \{x^n \in \mathcal{X}^n : q_{x^n} = q\}$ and

$T_{\tilde{q}} = \{x^n \in \mathcal{X}^n : \tilde{q}_{x^n} = \tilde{q}\}$. Define conditional type as: $\hat{q}(i|j) = \frac{n_{ij}}{n_j} = \frac{\tilde{q}(i,j)}{q(j)}$. We denote empirical entropies as

$$H(q) = -\sum_{j \in \mathcal{X}} \frac{n_j}{n} \log \frac{n_j}{n} \tag{2.1}$$

and

$$H(\tilde{q}) = -\sum_{(i,j) \in \mathcal{X}^2} \frac{n_{ij}}{n} \log \frac{n_{ij}}{n}. \tag{2.2}$$

For notational convenience with a slight abuse of notation we denote $H(\tilde{q}|q) = H(\tilde{q}) - H(q)$.

We consider a fixed-to-variable code that encodes an n-length sequence from the Markov source to a variable-length bit string via a coding function:

$$\phi : \mathcal{X}^n \to \{0,1\}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, ...\}.$$

We do not make the assumption that the code is prefix-free. Let $l(\phi(x^n))$ be the number of bits in the compressed binary string when x^n is the source sequence. The figure of merit in this setting is the probability under the true model that the length of compressed string exceeds a given integer k, given by

$$\epsilon_n(k, \phi, \hat{p}) = \mathbb{P}[l(\phi(X^n)) \geq k].$$

We gauge the performance of algorithms through the ϵ-coding rate at blocklength n given by:

$$R_n(\epsilon, \phi, \hat{p}) = \min\left\{\frac{k}{n} : \epsilon_n(k, \phi, \hat{p}) \leq \epsilon\right\}.$$

2.2 Main Result for the Markov Source

For the class of *all* memoryless sources over finite alphabets, the fixed-to-variable Type Size (TS) code is introduced in Kosut and Sankar (2013), which sorts sequences based on the size of the elementary type class from smallest to largest and then encodes sequences to variable-length bit-strings in this order. More precisely, define

the support set of a sequence as the set of observed symbols in it. The output of the encoder consists of a header that encodes the support of a sequence and a string that maps sequences to binary strings based on the size of the type class of X^n, among all sequences with the support set indicated in the header. That is, if two sequences x^n and y^n have the same support and $|T_{q(x^n)}| \leq |T_{q(y^n)}|$, then $l(\phi(x^n)) \leq l(\phi(y^n))$, where $q(x^n)$ is the type of x^n and $T_{q(x^n)}$ is the type class of x^n. The third-order coding rate of the TS code is derived in Kosut and Sankar (2013) and its optimality is shown in Kosut and Sankar (2014b). In this section, we assume a stationary Markov model is in force and derive the coding rate of the Type Size code for this generalized source model. The following theorem gives the optimal rate up to third order for the class of first-order stationary, irreducible and aperiodic Markov sources, and states that the Type Size code achieves this rate.

Theorem 1. *Assume a first-order stationary, irreducible and aperiodic Markov source as the underlying model. For a given $\epsilon > 0$, we have*

$$\inf_{\phi} \sup_{\hat{p} \in \hat{\mathcal{P}}} \Big[R_n(\epsilon, \phi, \hat{p}) - H(X_2|X_1) - \frac{\sigma}{\sqrt{n}} Q^{-1}(\epsilon) \Big] = \left(\frac{|\mathcal{X}|(|\mathcal{X}|-1)}{2} - 1 \right) \frac{\log n}{n} + \mathcal{O}\Big(\frac{1}{n}\Big) \tag{2.3}$$

where $H(X_2|X_1)$ is the conditional entropy of $\hat{p}$ and

$$\sigma^2 = \lim_{n \to \infty} \mathbb{E}\Big[\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(-\log \hat{p}(X_{i+1}|X_i) - H(X_2|X_1) \big) \Big)^2 \Big]. \tag{2.4}$$

Moreover, the Type Size code achieves the infimum in (2.3) for all distributions $\hat{p}$.

2.3 Preliminary Results

2.3.1 Variance of Occurrence

The following lemma will be needed to prove our main result. The lemma derives the growth rate of variance of number of occurrences of single letters.

Lemma 1. *Let 1_A be the indicator function of the event A, i.e. $1_A(x) = 1$ if $x \in A$ and 0 otherwise. Then*

$$Var(n_j) = Var\Big(\sum_{i\in[n]}(1_{X_i=j})\Big) = \Theta(n).$$

Proof. Let X_i be i'th source outcome. Let $\mathbf{Z}_i$ be the indicator vector of X_i, i.e., if $X_i = j \in \mathcal{X}$, then $\mathbf{Z}_i = \mathbf{e}_j$, where $\mathbf{e}_j$ is a vector of size $|\mathcal{X}|$ with 1 in the j-th position and zero elsewhere. Since X_i follows a Markov chain process, so does $\mathbf{Z}_i$. Let $\mathbb{P}(X_i = j) = p_j$, $j = 1, ..., |\mathcal{X}|$ and $\boldsymbol{p} = (p_i)$ be the stationary distribution of the Markov source. Let $\mathbf{W} = (w_{ij})$ be the transition probability matrix of X_i. It is shown in Anderson (1989) that

$$\mathrm{Cov}\left(\mathbf{Z}_i, \mathbf{Z}_{i-s}^T\right) = \left(\mathbf{W}^T\right)^s \mathbf{E}, \quad s = 0, 1, ...$$

where $\mathbf{E} = \mathbf{D}_{\boldsymbol{p}} - \boldsymbol{p}\boldsymbol{p}^T$, in which $\mathbf{D}_{\boldsymbol{p}}$ is a diagonal matrix with i-th diagonal element being p_i. Let $\mathbf{N} = \sum_{i=1}^n \mathbf{Z}_i$. We have

$$\begin{aligned}
\mathrm{Var}\left(\mathbf{N}\right) &= \mathrm{Var}\left(\sum_{i=1}^n \mathbf{Z}_i\right) \\
&= \sum_{i=1}^n \mathrm{Var}(\mathbf{Z}_i) + 2\sum_{i<j} \mathrm{Cov}\left(\mathbf{Z}_i, \mathbf{Z}_j^T\right) \\
&= \sum_{i=1}^n \mathrm{Var}\left(\mathbf{Z}_i\right) + 2\sum_{i<j} \left(\mathbf{W}^T\right)^{j-i} \mathbf{E}. \qquad (2.5)
\end{aligned}$$

The first term in (2.5) is $\Theta(n)$. We now find the asymptotics of the second term. We have

$$\begin{aligned}
\sum_{i<j} (\mathbf{W}^T)^{j-i}\mathbf{E} &= \left(\sum_{i=1}^{n-1}\sum_{s=1}^{n-i} \left(\mathbf{W}^T\right)^s\right) \mathbf{E} \\
&= \left(\sum_{i=1}^{n-1}\sum_{s=1}^{n-i} \left(\sum_{j=1}^{|\mathcal{X}|} \sigma_j \mathbf{u}_j \mathbf{v}_j^T\right)^s\right) \mathbf{E} \qquad (2.6) \\
&= \left(\sum_{i=1}^{n-1}\sum_{s=1}^{n-i} \left(\mathbf{u}_1\mathbf{v}_1^T + \sum_{j=2}^{|\mathcal{X}|} \sigma_j^s \mathbf{u}_j \mathbf{v}_j^T\right)\right) \mathbf{E} \qquad (2.7)
\end{aligned}$$

where in (2.6), the singular value decomposition of $\mathbf{W}^T$ is employed, with $\mathbf{u}_j$ and $\mathbf{v}_j$ as the right and left eigenvectors of $\mathbf{W}^T$ (resp.) corresponding to singular value σ_j. Since the Markov source is irreducible and aperiodic, therefore, $\sigma_1 = 1$ and $\sigma_j < 1$ for $j > 1$. Note that the stationary distribution of the Markov chain is $\mathbf{u}_1 = \boldsymbol{p} = (p_1, \cdots, p_n)$, and $\mathbf{v}_1 = \mathbf{e} = (1, ..., 1)$ is the all-ones vector. Since

$$(\mathbf{u}_1\mathbf{v}_1^T)\mathbf{E} = (\boldsymbol{p}\mathbf{e}^T)(\mathbf{D}_{\boldsymbol{p}} - \boldsymbol{p}\boldsymbol{p}^T) = \boldsymbol{p}\boldsymbol{p}^T - \boldsymbol{p}\boldsymbol{p}^T = 0$$

therefore, (2.6) vanishes exponentially. Lemma then follows. □

2.3.2 Markov Type Class Size

The BEST theorem van Aardenne-Ehrenfest and de Bruijn (1987) counts the number of Eulerian circuits in a simple graph based on the number of arborescences in the graph. Based on BEST theorem's proof and a counting argument, Davisson *et. al.* Davisson *et al.* (1981) derived a lower bound for the size of a Markov type class, where their corresponding graphs are not necessarily simple. We use the same counting argument to derive a more refined result. The following lemma upper and lower bounds the size of a Markov type class. It gives a tighter bound than Davisson's result in certain regimes.

Lemma 2. *If there exists a constant $\gamma > 0$, such that $n_{ij} \geq \gamma n_i$, for all $i, j \in \mathcal{X}$ with $n_{ij} > 0$, then size of the Markov type class of type $\tilde{q}$ is bounded as*

$$nr(\tilde{q}) - \sum_{j=1}^{|\mathcal{X}|}\sum_{i=1}^{|\mathcal{X}|}\frac{1}{12n_{ij}} + \frac{1}{12}\sum_{j=1}^{|\mathcal{X}|}\frac{1}{n_j+1} - |\mathcal{X}|\log\gamma \leq \log|T_{\tilde{q}}|$$
$$\leq nr(\tilde{q}) - \sum_{j=1}^{|\mathcal{X}|}\sum_{i=1}^{|\mathcal{X}|}\frac{1}{12(n_{ij}+1)} + \frac{1}{12}\sum_{j=1}^{|\mathcal{X}|}\frac{1}{n_j} + \log|\mathcal{X}|$$

where

$$r(\tilde{q}) = H(\tilde{q}|q) + \frac{1}{2n}\sum_{j=1}^{|\mathcal{X}|}\log n_j - \frac{1}{2n}\sum_{j=1}^{|\mathcal{X}|}\sum_{i=1}^{|\mathcal{X}|}\log n_{ij} - \frac{|\mathcal{X}|(|\mathcal{X}|-1)}{2n}\log(2\pi) \quad (2.8)$$

is the common part of the lower and upper bounds.

Proof. ***Lower Bound:*** Let $x^n = x_1 x_2 \cdots x_n$ be the observed sequence. We associate a directed multi-graph $G = (V, E)$ to x^n as follows. The vertices V of G correspond to the distinct symbols in x^n. We assume n is large enough such that we see all the alphabet letters in the observed sequence, hence $V = \mathcal{X}$. Given type $\tilde{q}$ with $n_{xy} = n\tilde{q}_{xy}$, the set of arcs E in G contains n_{xy} arcs from node x to y for each pair $(x, y) \in \mathcal{X}^2$. We say y is parent of x. Note that each sequence $x^n \in T_{\tilde{q}}$ corresponds to a Eulerian path on all n edges in G. Let $\mathrm{Eul}(G)$ be the set of Eulerian circuits. Since Eulerian circuits are invariant to cyclic permutations of the edges while sequences are not, the number of elements of the type class is not exactly the same as the number of Eulerian circuits, but it is certainly true that $|\mathrm{Eul}(G)| \leq |T_{\tilde{q}}|$.

Note that, unlike in the BEST theorem, we allow parallel edges, so we cannot use the proof of the BEST theorem exactly. We first convert G to a simple graph (allowing self loops) G', by merging all parallel edges into a single edge. Now, following essential argument of the BEST theorem, we fix a vertex $u \in V(G')$ and one of its outgoing edges, say e_1. We can think of this step as fixing the starting point of the Eulerian circuit. We then find a spanning tree, S, of G' rooted at u (in graph terminology this is an arborescence of G' rooted at u). For any vertex v with out-degree d_v, we label the outgoing edges of v arbitrarily from $1, \cdots, d_v$, while if $v \neq u$ we reserve number d_v to the only arc belonging to S, and for $v = u$ we reserve number 1 to e_1. As shown in van Aardenne-Ehrenfest and de Bruijn (1987); Davisson *et al.* (1981), any distinct labeling corresponds to an Eulerian circuit in G, by beginning with e_1 and always traveling along the outgoing edge with lowest remaining label from the current vertex. For notational convenience let $x_{j,S}$, be the parent of node j in S. Note that for each vertex $j \in \mathcal{X}$, since the label of the outgoing edge of j in S is fixed, we can label other outgoing edges of j through $(n_j - 1)!$ arrangements. However, For any of the node j's parents, say i, in any such labeling the $n_{ji}!$ permutations of labelings of

the parallel edges between j and i result in a same Eulerian circuit. The same, for all the $(n_{jx_{j,S}} - 1)!$ permutations of the edges (which are not in S) between j and $x_{j,S}$. Therefore, the number of Eulerian circuits correspond to the spanning tree S and this specific labeling is given by

$$\prod_{j=1}^{|\mathcal{X}|} \frac{(n_j - 1)!}{\{\prod_{\substack{i=1 \\ i \neq x_{j,S}}} (n_{ji}!)\}(n_{jx_{j,S}} - 1)!}. \tag{2.9}$$

Since there is at least one spanning tree S, for any S, (2.9) constitutes a lower bound on $|T_{\tilde{q}}|$. Using Stirling's formula, we may further lower bound (2.9) in a manner independent of S by

$$\begin{aligned}
\prod_{j=1}^{|\mathcal{X}|} \frac{(n_j - 1)!}{\{\prod_{\substack{i=1 \\ i \neq x_j}} (n_{ji}!)\}(n_{jx_{j,S}} - 1)!} &= \prod_{j=1}^{|\mathcal{X}|} \Big(\frac{n_j!}{\prod_{i=1} n_{ij}!} \cdot n_{jx_{j,S}} \cdot n_j^{-1} \Big) \\
&\geq \prod_{j=1}^{|\mathcal{X}|} \Big(\frac{\sqrt{2\pi} n_j^{n_j + \frac{1}{2}} 2^{-n_j + \frac{1}{12(n_j+1)}}}{\prod_{i=1}^{|\mathcal{X}|} \sqrt{2\pi} n_{ij}^{n_{ij} + \frac{1}{2}} 2^{-n_{ij} + \frac{1}{12 n_{ij}}}} \cdot n_{jx_{j,S}} \cdot n_j^{-1} \Big) \\
&\geq \frac{2^{nH(\tilde{q}|q) + \sum_j \left(\frac{1}{2} \log n_j + \frac{1}{12(n_j+1)} \right) - \sum_{i,j} \left(\frac{1}{2} \log n_{ij} + \frac{1}{12 n_{ij}} \right) + |\mathcal{X}| \log \gamma}}{(2\pi)^{\frac{|\mathcal{X}|(|\mathcal{X}|-1)}{2}}}
\end{aligned} \tag{2.10}$$

where we used the assumption that $n_{jx_{j,S}} \geq \gamma n_j$ for all $j \in \mathcal{X}$. The lower bound on the type class size follows.

Upper Bound: It is shown in Davisson *et al.* (1981) that $|T_{\tilde{q}}| \leq |\mathcal{X}| \prod_{j=1}^{|\mathcal{X}|} \frac{n_j!}{\prod_{i=1}^{|\mathcal{X}|} n_{ji}!}$. Using the same approach as in deriving the lower bound, one may bound the factorials using the Stirling formula Csiszár and Körner (1982), to derive the upper bound on the type class size as follows:

$$\begin{aligned}
|\mathcal{X}| \Pi_{j=1}^{|\mathcal{X}|} \frac{n_j!}{\Pi_{i=1}^{|\mathcal{X}|} n_{ji}!} &\leq |\mathcal{X}| \cdot \Pi_{j=1}^{|\mathcal{X}|} \frac{\sqrt{2\pi} n_j^{n_j + \frac{1}{2}} 2^{-n_j + \frac{1}{12 n_j}}}{\Pi_{i=1}^{|\mathcal{X}|} \sqrt{2\pi} n_{ji}^{n_{ji} + \frac{1}{2}} 2^{-n_{ji} + \frac{1}{12(n_{ji}+1)}}} \\
&= \frac{|\mathcal{X}|}{(\sqrt{2\pi})^{|\mathcal{X}| \cdot (|\mathcal{X}|-1)}} \cdot \frac{2^{-nH(q) + n \log n + \frac{1}{2} \sum_{j=1}^{|\mathcal{X}|} \log n_j - n + \sum_{j=1}^{|\mathcal{X}|} \frac{1}{12 n_j}}}{2^{-nH(\tilde{q}) + n \log n + \frac{1}{2} \sum_{j=1}^{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \log n_{ij} - n + \sum_{j=1}^{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \frac{1}{12(n_{ij}+1)}}}.
\end{aligned}$$

Lemma follows. □

Corollary 1. *If we sort sequences based on the empirical entropy and encode them in this order, we obtain the same result.*

2.3.3 Bounds on Empirical Entropy

We first show the following refined version of delta method Cramér (1999) which is a Markov version of Proposition 1 in MolavianJazi and Laneman (2013).

Proposition 1. *Let* $\mathbf{U}_i$ *for* $i = 1, \ldots, n$ *be zero-mean, first-order, stationary, irreducible and aperiodic Markov chain of random vectors in* $\mathbb{R}^m$ *with finite third moment* $\mathbb{E}(\|\mathbf{U}_i\|^3)$*. Let* $f : \mathbb{R}^m \to \mathbb{R}$ *be a function with uniformly bounded second-order partial derivatives in a* m*-hypercube neighborhood of* $\mathbf{0}$ *with side length at least* $\frac{1}{\sqrt[4]{n}}$*. Let* $\boldsymbol{j}$ *be the vector of first-order partial derivatives of* f *at* $\mathbf{0}$*, i.e.* $j_r = \frac{\partial f(\mathbf{u})}{\partial u_r}\Big|_{\mathbf{u}=\mathbf{0}}$*. Let* $\boldsymbol{V}_n = Cov(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{U}_i)$ *and* $\sigma_n^2 = \boldsymbol{j}^T \boldsymbol{V}_n \boldsymbol{j}$*. Let* $\boldsymbol{V}_\infty = \lim_{n\to\infty} \boldsymbol{V}_n$ *and* $\sigma_\infty^2 = \lim_{n\to\infty} \sigma_n^2$*. Fix a positive constant* α*. There exists* B *such that for any* δ *such that* $|\delta| \le \alpha$*,*

$$\Big|\mathbb{P}\Big[f\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{U}_i\Big) \ge f(\mathbf{0}) + \frac{\sigma_\infty}{\sqrt{n}}\delta\Big] - Q(\delta)\Big| \le \frac{B}{\sqrt{n}}. \tag{2.11}$$

Proof. We first normalize the function f as follows. Let $\mathbf{R}$ be the Cholesky decomposition matrix of $\mathbf{V}_n$, so that $\mathbf{R}$ is upper triangular and $\mathbf{V}_n = \mathbf{R}^T\mathbf{R}$. Moreover $\mathbf{R}^{-T}\mathbf{V}_n\mathbf{R}^{-1} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Note that $\|\frac{\mathbf{R}\mathbf{j}}{\sigma_n}\|^2 = \frac{\mathbf{j}^T\mathbf{V}_n\mathbf{j}}{\sigma_n^2} = 1$. Thus we may define an orthogonal matrix $\mathbf{A}$ whose first column is $\mathbf{R}\mathbf{j}/\sigma_n$. Now let $\tilde{f}(\mathbf{w}) = \frac{1}{\sigma_\infty}[f(\mathbf{R}^T\mathbf{A}\mathbf{w}) - f(\mathbf{0})]$. Thus we may write $f(\mathbf{u}) = f(\mathbf{0}) + \sigma_\infty \tilde{f}(\mathbf{A}^{-1}\mathbf{R}^{-T}\mathbf{u})$. In particular

$$\mathbb{P}\Big[f\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{U}_i\Big) \ge f(\mathbf{0}) + \frac{\sigma_\infty}{\sqrt{n}}\delta\Big] = \mathbb{P}\Big[\tilde{f}\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{W}_i\Big) \ge \frac{\delta}{\sqrt{n}}\Big]$$

where we have defined $\mathbf{W}_i = \mathbf{A}^{-1}\mathbf{R}^{-T}\mathbf{U}_i$. Defining $\bar{\mathbf{W}} = \frac{1}{n}\sum_{i=1}^n \mathbf{W}_i$ we note that $\mathrm{Cov}(\bar{\mathbf{W}}) = \frac{1}{n}\mathbf{I}$. Moreover, by our choice of $\mathbf{A}$, we have $\nabla\tilde{f}(\mathbf{w}) = \frac{1}{\sigma_\infty}\mathbf{A}^T\mathbf{R}\mathbf{j} = \frac{\sigma_n}{\sigma_\infty}\mathbf{e}_1$ where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$. Let $r(\mathbf{w}) = \tilde{f}(\mathbf{w}) - \frac{\sigma_n}{\sigma_\infty}w_1$. Let $\mathcal{B} = \Big\{\mathbf{w} : \|\mathbf{R}^T\mathbf{A}\mathbf{w}\|_\infty \le$

$\frac{1}{\sqrt[4]{n}}\Big\}$. By Taylor's theorem, for all $\mathbf{w} \in \mathcal{B}$ we have $|r(\mathbf{w})| \leq c_0\|\mathbf{w}\|^2$, where $c_0 = \sup_{\mathbf{w}\in\mathcal{B}} \max_{l\in[m]} \sum_{p=1}^{m} \frac{1}{2}\Big|\frac{\partial^2 \tilde{f}(\mathbf{w})}{\partial w_l \partial w_p}\Big|$. By the assumption of uniform boundedness of the second-order partial derivatives of f, c_0 is finite. Thus for all $\mathbf{w} \in \mathcal{B}$ we have $\tilde{f}(\mathbf{w}) \geq \frac{\sigma_n}{\sigma_\infty} w_1 - c_0\|\mathbf{w}\|^2$. We may write

$$\begin{aligned}
\mathbb{P}(\tilde{f}(\bar{\mathbf{W}}) \geq \delta/\sqrt{n}) &\geq \mathbb{P}(\tilde{f}(\bar{\mathbf{W}}) \geq \delta/\sqrt{n}, \bar{\mathbf{W}} \in \mathcal{B}) \\
&\geq \mathbb{P}(\frac{\sigma_n}{\sigma_\infty}\bar{\mathrm{W}}_1 - c_0\|\bar{\mathbf{W}}\|^2 \geq \delta/\sqrt{n}, \bar{\mathbf{W}} \in \mathcal{B}) \\
&\geq \mathbb{P}(\frac{\sigma_n}{\sigma_\infty}\bar{\mathrm{W}}_1 - c_0\|\bar{\mathbf{W}}\|^2 \geq \delta/\sqrt{n}) \\
&\quad - \mathbb{P}(\bar{\mathbf{W}} \notin \mathcal{B}) \\
&= \mathbb{P}(\bar{\mathbf{W}} \in \mathcal{C}) - \mathbb{P}(\bar{\mathbf{W}} \notin \mathcal{B}) \qquad (2.12)
\end{aligned}$$

where we have defined

$$\mathcal{C} = \Big\{\mathbf{w} \in \mathbb{R}^m : \frac{\sigma_n}{\sigma_\infty}\mathrm{w}_1 - c_0\|\mathbf{w}\|^2 \geq \delta/\sqrt{n}\Big\}. \qquad (2.13)$$

We may bound the second term in (2.12) using Chebyshev's inequality by

$$\begin{aligned}
\mathbb{P}\Big(\bar{\mathbf{W}} \notin \mathcal{B}\Big) &= \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{W}_i \notin \mathcal{B}\Big) \\
&= \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}^{-1}\mathbf{R}^{-T}\mathbf{U}_i \notin \mathcal{B}\Big) \\
&= \mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{U}_i\Big\|_\infty > \frac{1}{\sqrt[4]{n}}\Big) \\
&\leq \sum_{s=1}^{m}\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{U}_i(s) > \frac{1}{\sqrt[4]{n}}\Big) \\
&\leq \sum_{s=1}^{m}\frac{\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{U}_i(s)\Big)}{\frac{1}{\sqrt{n}}} = \mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big). \qquad (2.14)
\end{aligned}$$

where for the last line we show in AppendixA (Claim 1) that $\mathrm{Var}\Big(\frac{1}{n}\big(\sum_{i=1}^{n}\mathbf{U}_i(s)\big)\Big) = \mathcal{O}(\frac{1}{n})$ for all $s = 0, \ldots, m-1$, where $\mathbf{U}_i(s)$ is the sth component of $\mathbf{U}_i$. Now consider

the first term in (2.12). Using Lapinskas' Lapinskas (1974) version of the Berry-Esseen theorem, and defining $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, since $\mathrm{Cov}(\sqrt{n}\bar{\mathbf{W}}) = I$ and $\mathcal{C}$ is convex, we have

$$\begin{aligned}
\mathbb{P}\Big(\bar{\mathbf{W}} \in \mathcal{C}\Big) &\geq \mathbb{P}\Big(\mathbf{Z}/\sqrt{n} \in \mathcal{C}\Big) - \frac{B'}{\sqrt{n}} = \mathbb{P}\Big(\frac{\sigma_n}{\sigma_\infty}\frac{\mathrm{Z}_1}{\sqrt{n}} - \frac{c_0}{n}\|\mathbf{Z}\|^2 \geq \frac{\delta}{\sqrt{n}}\Big) - \frac{B'}{\sqrt{n}} \\
&= \mathbb{P}\Big(\frac{\sigma_n}{\sigma_\infty}Z_1 - \frac{c_0}{\sqrt{n}}\|\mathbf{Z}\|^2 \geq \delta\Big) - \frac{B'}{\sqrt{n}} \\
&= \mathbb{P}\Big(\frac{\sigma_n}{\sigma_\infty}Z_1 \geq \delta\Big) - \mathbb{P}\Big(\delta < \frac{\sigma_n}{\sigma_\infty}Z_1 \leq \delta + \frac{c_0}{\sqrt{n}}\|\mathbf{Z}\|^2\Big) - \frac{B'}{\sqrt{n}} \\
&= Q\Big(\frac{\sigma_\infty}{\sigma_n}\delta\Big) - \mathbb{P}\Big(\frac{\sigma_\infty}{\sigma_n}\delta < Z_1 \leq \frac{\sigma_\infty}{\sigma_n}\delta + \frac{\sigma_\infty}{\sigma_n}\frac{c_0}{\sqrt{n}}\|\mathbf{Z}\|^2\Big) - \frac{B'}{\sqrt{n}} \qquad (2.15)
\end{aligned}$$

where $\frac{B'}{\sqrt{n}}$ is as in Equation (3) in Theorem of Lapinskas (1974). We show that the second term in (2.15) is $\mathcal{O}(\frac{1}{\sqrt{n}})$. Let p_S be the distribution of a chi-squared distribution with m degrees of freedom. Thus $\|\mathbf{Z}\|^2 \sim p_S$. We may write the second term in (2.15) as

$$\int_0^\infty \mathbb{P}\Big(\frac{\sigma_\infty}{\sigma_n}\delta < Z_1 \leq \frac{\sigma_\infty}{\sigma_n}\delta + \frac{\sigma_\infty}{\sigma_n}\frac{c_0 s}{\sqrt{n}}\Big|\|\mathbf{Z}\|^2 = s\Big)dp_S(s). \qquad (2.16)$$

For any i, let $S_i = \frac{2\pi^{(i+1)/2}}{\Gamma((i+1)/2)}$, so that $S_i r^i$ is the surface area of the i-sphere of radius r (note that i denotes the dimension of the manifold of the sphere, not the dimension that the sphere sits in). Conditioning on $\|\mathbf{Z}\|^2 = s$, $\mathbf{Z}$ is uniformly distributed on the $(m-1)$-sphere of radius $\sqrt{s}$. For any $0 \leq a \leq b$, we may write

$$\mathbb{P}\big(a < Z_1 \leq b \,\big|\, \|\mathbf{Z}\|^2 = s\big) = \frac{\mathrm{Vol}(\mathbf{z} : a < z_1 \leq b, \|\mathbf{z}\|^2 = s)}{S_{m-1}s^{\frac{m-1}{2}}} \qquad (2.17)$$

where Vol denotes the $(m-1)$-dimensional Lebesgue measure. The volume in (2.17) may be upper bounded by

$$
\begin{aligned}
&\mathrm{Vol}\Big(\big\{\mathbf{z} : a \le z_1 \le b, a^2 + z_2^2 + \cdots + z_m^2 = s\big\} \\
&\qquad\qquad \cup \big\{\mathbf{z} : z_1 = b, s - b^2 \le z_2^2 + \cdots + z_m^2 \le s - a^2\big\} \Big) \\
&= (b-a) S_{m-2} (s-a^2)^{\frac{m-2}{2}} + \int_{\sqrt{s-b^2}}^{\sqrt{s-a^2}} S_{m-2} t^{m-2} dt \\
&\le (b-a) S_{m-2} s^{\frac{m-2}{2}} + \int_{\sqrt{s-b^2}}^{\sqrt{s-a^2}} S_{m-2} (\sqrt{s})^{m-2} dt \\
&= \left[b - a + \sqrt{s-a^2} - \sqrt{s-b^2} \right] S_{m-2} s^{\frac{m-2}{2}}.
\end{aligned}
$$

Thus

$$
\mathbb{P}\big(a < Z_1 \le b \,\big|\, \|\mathbf{Z}\|^2 = s\big) \le \left[b - a + \sqrt{s-a^2} - \sqrt{s-b^2} \right] \frac{S_{m-2}}{S_{m-1}} \sqrt{s}.
$$

Note that for the probability in (2.16), $a = \frac{\sigma_\infty}{\sigma_n}\delta$ and $b = \frac{\sigma_\infty}{\sigma_n}\delta + \frac{\sigma_\infty}{\sigma_n}\frac{c_0 s}{\sqrt{n}}$, so $b - a = \frac{\sigma_\infty}{\sigma_n}\frac{c_0 s}{\sqrt{n}}$ and $\sqrt{s-a^2} - \sqrt{s-b^2} \le \frac{\sigma_\infty}{\sigma_n}\frac{c_0 s}{\sqrt{n}}$ for sufficiently large n. Thus (2.16) may be upper bounded by

$$
\begin{aligned}
\int_0^\infty dp_S(s) \frac{\sigma_\infty}{\sigma_n} \frac{2c_0}{\sqrt{n}} \frac{S_{m-2}}{S_{m-1}} \sqrt{s} &= \frac{\sigma_\infty}{\sigma_n} \frac{2c_0 S_{m-2}}{\sqrt{n} S_{m-1}} \mathbb{E}(\|\mathbf{Z}\|) \\
&= \frac{\sigma_\infty}{\sigma_n} \frac{2c_0 S_{m-2}}{\sqrt{n} S_{m-1}} \sqrt{2} \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})} \\
&= \frac{\sigma_\infty}{\sigma_n} \frac{\sqrt{n} c_0 (m-1)}{\sqrt{\pi n}} \\
&= \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}
$$

where we have used the mean of a chi distribution and the definition of S_i. We show in Appendix A (Claim 3) that $\frac{\sigma_\infty - \sigma_n}{\sigma_n} = \mathcal{O}(\frac{1}{\sqrt{n}})$, hence using Taylor series expansion, we have

$$
Q\big(\frac{\sigma_\infty}{\sigma_n}\delta\big) = Q\big(\delta + \frac{\sigma_\infty - \sigma_n}{\sigma_n}\delta\big) = Q\big(\delta\big) + \frac{C}{\sqrt{n}}
$$

for a constant C. Repeating the entire argument with $-f$ playing the role of f gives the opposite inequality. □

Lemma 3. *Fix positive constant $\wp$. For any distribution in $\hat{\mathcal{P}}$,*

$$\left|\mathbb{P}\Big(H(\tilde{q}|q) > \wp\Big) - Q\Big(\frac{\sqrt{n}}{\sigma}(\wp - H(X_2|X_1))\Big)\right| \leq \frac{A}{\sqrt{n}} \tag{2.18}$$

where A is a constant not depending on n.

Proof. We use Proposition 1. We define $\mathbf{U}_i \in \mathbb{R}^{|\mathcal{X}|^2}$ for $i = 1, \ldots, n$ as follows. We index $\mathbf{U}_i$'s by tuples, where each component of $\mathbf{U}_i$ corresponds to an element of $\mathcal{X}^2$. Now let $\mathbf{U}_i(x, y) = 1 - \tilde{p}(x, y)$, if $(X_i, X_{i+1}) = (x, y)$ and $-\tilde{p}(x, y)$ otherwise. Since the Markov chain $\{X_i\}$ is stationary, irreducible and aperiodic, so is $\{\mathbf{U}_i\}$. Clearly $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$ for $i = 1, \ldots, n$. Let us denote $\bar{\mathbf{U}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{U}_i$. Let us define $f : \mathbb{R}^{|\mathcal{X}|^2} \to \mathbb{R}$ as follows:

$$f(\mathbf{u}) = \sum_{(x,y)\in\mathcal{X}^2} \Big(\mathbf{u}(x, y) + \tilde{p}(x, y)\Big) \cdot \log \frac{\mathbf{u}(x, y) + \tilde{p}(x, y)}{\sum_{z\in\mathcal{X}} \Big(\mathbf{u}(z, y) + \tilde{p}(z, y)\Big)}.$$

Observe that $H(\tilde{q}|q) = f(\bar{\mathbf{U}})$. Finally, in Appendix B (Claim 4) we show that for this choice of $\mathbf{U}_i$'s and $f(\cdot)$, $\sigma_\infty^2 = \sigma^2$. Applying Proposition 1 now completes the proof. □

2.3.4 Laplace's Approximation

We have the following theorem for the integral of manifolds. We refer the reader to (Wong, 1989, Chap. 9, Th. 3) for a detailed proof.

Theorem 2. *Wong (1989) Let D be a d-dimensional differentiable manifold embedded in $\mathbb{R}^m$ and f and g be functions that are infinitely differentiable on D. Let*

$$J(n) = \int_D g(x) \exp^{-nf(x)}\, dx. \tag{2.19}$$

Assume that: (i) the integral in (2.19) converges absolutely for all $n \geq n_0$; (ii) there exists a point x^* in the interior of D such that for every $\epsilon > 0$, $\rho(\epsilon) > 0$ where

$$\rho(\epsilon) = \inf\{f(x) - f(x^*) : x \in D \text{ and } |x - x^*| \geq \epsilon\}$$

and (iii) the Hessian matrix $A = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)\Big|_{x=x^*}$ *is positive definite. Let* $F \in \mathbb{R}^{m\times d}$ *be an orthonormal basis for the tangent space to* D *at* x^**. Then*

$$J(n) = \exp^{-nf(x^*)} \left(\frac{2\pi}{n}\right)^{\frac{d}{2}} g(x^*)|F^T A F|^{-\frac{1}{2}} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

2.4 Proof Sketch of Main Result for Markov Source

Proof of Theorem 1 is similar to its *i.i.d.* version Kosut and Sankar (2013, 2014b,a) which deviates from it due to differences in the Lemmas 2 and 3 with their counterparts in Kosut and Sankar (2013) and Kosut and Sankar (2014b), respectively. We omit details of the proof for space.

Achievability: First note that by Lemma 1, one can show that the probability that the assumption of Lemma 2, does not hold vanishes as $\mathcal{O}(\frac{1}{n})$. Now, Lemmas 2 and 3 allow us to bound the CDF of the size of the empirical type class $T_{\tilde{q}}$. In particular, we can show that $\mathbb{P}\left[\frac{\log|T_{\tilde{q}}|}{n} > \tau\right] \leq \epsilon$, where

$$\tau = H(X_2|X_1) + \frac{\sigma}{\sqrt{n}} Q^{-1}(\epsilon) + \Big(\frac{|\mathcal{X}|}{2} - \frac{|\mathcal{X}|^2}{2}\Big)\frac{\log n}{n} + \mathcal{O}\left(\frac{1}{n}\right).$$

Thus the rate achieved by the Type Size code can be upper bounded by calculating the total number of sequences in type classes with size no larger than $2^{n\tau}$. The log of this number of sequences can be upper bounded by

$$n\tau + \big(|\mathcal{X}|^2 - |\mathcal{X}| - 1\big)\log n + \mathcal{O}(1).$$

Substituting for τ yields the desired achievable bound.

Converse: The converse is proved using a uniform mixture $\bar{p}$ of all Markov distributions with fixed entropy rate. An application of a finite blocklength converse from Kontoyiannis and Verdú (2014) can be used to bound the rate in terms of the CDF of $-\log \bar{p}(X^n)$ with respect to the distribution $\bar{p}$. Applying Laplace's approximation allows us to bound this CDF in terms of the CDF of the empirical entropy, which in turn is bounded using Lemma 3.

2.5 Two Stage Code

Two-Stage code Cover and Thomas (2006) is a well-known approach to encode strings from an unknown source. In the first stage, type of the sequence is encoded. Subsequently, in the second stage the index of the sequence, among all sequences with the described type class in the first stage, is encoded. For the *i.i.d.* data generation mechanism, third-order coding rate of a two-stage code wherein the first stage is a fixed-length code and the second stage is an optimal variable-length code for each type class is derived in Kosut and Sankar (2014b,a). Denote $\Phi^{2\mathcal{S}}$ as the n-length Two-Stage code. Let $\mathcal{P}_n(\mathcal{X})$ be the number of types with alphabet $\mathcal{X}$. By construction, the ϵ-rate achieved by this code is given by Kosut and Sankar (2014a)

$$R_n(\Phi^{2\mathcal{S}}, \epsilon, p_{\theta^*}) = \frac{1}{n}\Big(\lceil \log |\mathcal{P}_n(\mathcal{X})| \rceil + k^*(\epsilon)\Big) \tag{2.20}$$

where

$$k^*(\epsilon) = \min\left\{k : \sum_{\boldsymbol{\tau}_c \in \mathcal{T}_c} \mathbb{P}(T_{\boldsymbol{\tau}_c}) \left|1 - \frac{2^{k+1}-1}{|T_{\boldsymbol{\tau}_c}|}\right|^+ \leq \epsilon\right\}. \tag{2.21}$$

Following the same steps as in Kosut and Sankar (2014b,a), we derive the third-order coding rate of the Two-Stage code for the Markov case as stated in the following Theorem.

Theorem 3. *For the Two-Stage code with fixed-length first stage,*

$$R(\Phi_n^{2\mathcal{S}}; \epsilon, p) = H(X_2|X_1) + \frac{\sigma}{\sqrt{n}} Q^{-1}(\epsilon) + \frac{|\mathcal{X}|(|\mathcal{X}|-1)}{2}\frac{\log n}{n} + \mathcal{O}\left(\frac{1}{n}\right). \tag{2.22}$$

Proof. The proof is the same as in Kosut and Sankar (2014a), and we omit it due to similarity. □

Chapter 3

PARAMETRIC SOURCE

3.1 Problem Statement

Let Θ be a compact subset of $\mathbb{R}^d$ with non-empty interior. Probability distributions in an exponential family can be expressed in the form Merhav and Weinberger (2004)

$$p_\theta(x) = 2^{\langle\theta,\boldsymbol{\tau}(x)\rangle-\psi(\theta)} \tag{3.1}$$

where $\theta \in \Theta$ is the d-dimensional parameter vector, $\boldsymbol{\tau}(x) : \mathcal{X} \to \mathbb{R}^d$ — the crux of our parametric approach — is the vector of sufficient statistics and $\psi(\theta)$ is the normalizing factor. Let the model class $\mathcal{P} = \{p_\theta, \theta \in \Theta\}$, be the exponential family of distributions over the finite alphabet $\mathcal{X} = \{1, \cdots, |\mathcal{X}|\}$, parameterized by $\theta \in \Theta \subset \mathbb{R}^d$, where d is the degrees of freedom in the minimal description of $p_\theta \in \mathcal{P}$ in the sense that no smaller dimensional family can capture the same model class. The degrees of freedom turns out to characterize the richness of the model class in our context. Compactness of Θ implies existence of a constant upper bound $\wp$ on the norm of the parameter vectors, namely $\|\theta\| \leq \wp$ for all $\theta \in \Theta$. We denote the (unknown) true model in force as p_{θ^*}. $\mathbb{P}_\theta$, $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ denote probability, expectation and variance with respect to p_θ, respectively. All logarithms are in base 2. Instead of introducing different indices for every new constant $C_1, C_2, ...$, the same letter C is used to denote different constants whose precise values are irrelevant.

From (3.1), the probability of a sequence x^n drawn *i.i.d.* from a model p_θ in the exponential family takes the form Merhav and Weinberger (2004)

$$
\begin{aligned}
p_\theta(x^n) &= \prod_{i=1}^{n} p_\theta(x_i) \\
&= \prod_{i=1}^{n} 2^{\left\langle \theta, \boldsymbol{\tau}(x_i) \right\rangle - \psi(\theta)} \\
&= 2^{\left\{ n \left[\left\langle \theta, \boldsymbol{\tau}(x^n) \right\rangle - \psi(\theta) \right] \right\}}
\end{aligned} \tag{3.2}
$$

where

$$
\boldsymbol{\tau}(x^n) = \frac{\sum_{i=1}^{n} \boldsymbol{\tau}(x_i)}{n} \in \mathbb{R}^d \tag{3.3}
$$

is a minimal sufficient statistic Merhav and Weinberger (2004). Note that $\boldsymbol{\tau}(x)$ and $\boldsymbol{\tau}(x^n)$ are distinguished based upon their arguments.

We consider a fixed-to-variable code that encodes an n-length sequence from the parametric source to a variable-length bit string via a coding function

$$
\phi : \mathcal{X}^n \to \{0,1\}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, \cdots\}.
$$

We do not make the assumption that the code is prefix-free. Let $l(\phi(x^n))$ be the number of bits in the compressed binary string when x^n is the source sequence. We gauge the performance of algorithms through the ϵ-coding rate at blocklength n given by

$$
R_n(\epsilon, \phi, p_{\theta^*}) := \min \left\{ \frac{k}{n} : \mathbb{P}_{\theta^*} \Big[l(\phi(X^n)) \geq k \Big] \leq \epsilon \right\}.
$$

3.2 Type Size Code

For the class of all memoryless sources over a finite alphabet $\mathcal{X}$, the fixed-to-variable TS code is introduced in Kosut and Sankar (2013), which sorts sequences based on the size of the elementary type class from smallest to largest and then encodes sequences to variable-length bit-strings in this order. More precisely, define the support of a sequence as the set of observed symbols in it. The output of the

encoder consists of a header that encodes the support of the sequence and a body that maps sequences to binary strings based on the size of their type class, among all sequences with the support set indicated in the header. That is, if two sequences x^n and y^n have the same support and $|T_{x^n}| \leq |T_{y^n}|$, then $l\left(\phi(x^n)\right) \leq l\left(\phi(y^n)\right)$, where T_{x^n} is the type class of x^n.

We borrow the spirit of the TS code, yet our approach for parametric sources departs from that of Kosut and Sankar (2013) in two ways

1. Rather than defining type classes based on the EPMFs, we use quantized type classes, which are based on the neighborhoods of the minimal sufficient statistics.

2. We omit the header encoding the support of the observed sequence. This header is unnecessary given the assumption that Θ is compact, because under this assumption, for any distribution in $\mathcal{P}$, each letter $x \in \mathcal{X}$ occurs with some probability bounded away from zero. Thus, all letters are likely to be observed for even moderate blocklengths.

We first define quantized type classes for the purpose of compressing the exponential family. We cover the convex hull of the set of minimal sufficient statistics $\mathcal{T} = \operatorname{conv}\left\{\boldsymbol{\tau}(x^n) : x^n \in \mathcal{X}^n\right\}$, into d-dimensional cubic grids — cuboids — of side length $\frac{s}{n}$, where $s > 0$ is a constant. The union of such disjoint cuboids should cover $\mathcal{T}$. The position of these cuboids is arbitrary, however once we cover the space, the covering is fixed throughout. We represent each d-dimensional cuboid by its geometrical *center*. Denote $G(\boldsymbol{\tau}_0)$ as the cuboid with center $\boldsymbol{\tau}_0$, more precisely

$$G(\boldsymbol{\tau}_0) := \left\{\boldsymbol{z} + \boldsymbol{\tau}_0 \in \mathbb{R}^d : -\frac{s}{2n} < z_i \leq \frac{s}{2n} \text{ for } 1 \leq i \leq d\right\} \tag{3.4}$$

where z_i is the i-th component of the d-dimensional vector $\boldsymbol{z}$. Let $\boldsymbol{\tau}_c(x^n)$ be the center

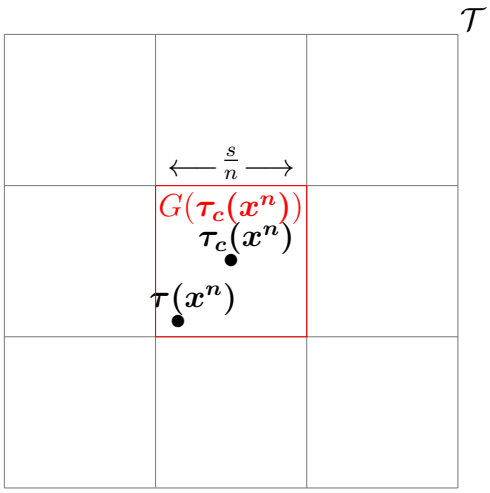

Figure 3.1: Quantized Types

of the cuboid that contains $\boldsymbol{\tau}(x^n)$. Let us denote $\mathcal{T}_c$ as the set of cuboid centers, i.e., $\mathcal{T}_c = \{\boldsymbol{\tau}_c(x^n) : x^n \in \mathcal{X}^n\}$.

We then define the quantized type class of x^n as

$$T_{x^n} := \{y^n \in \mathcal{X}^n : \boldsymbol{\tau}(y^n) \in G\left(\boldsymbol{\tau}_c(x^n)\right)\} \tag{3.5}$$

the set of all sequences y^n with minimal sufficient statistic belonging to the very same cuboid containing the minimal sufficient statistic of x^n (See Figure 3.1).

Since quantized type classes are represented by the cuboids and consequently the cuboid centers, we may interchangeably use $T_{\boldsymbol{\tau}_0}$ as the type class with corresponding cuboid center $\boldsymbol{\tau}_0$. Hence, $T_{\boldsymbol{\tau}_c(x^n)}$ is the same as T_{x^n}.

Two sequences within the given type class are indistinguishable from the coding perspective. The sequence indistinguishability introduced in this paper is reminiscent of the Balasubramanian's model indistinguishability Balasubramanian (2005). In contrast to the sequence indistinguishability approach where the space of minimal sufficient statistics is partitioned into cuboids, in a model indistinguishability approach one may partition the source space. Asymptotics of the model indistin-

guishability approach is derived in Rissanen (1996), where the maximum likelihood estimate is quantized to some precision by being the center of a cuboid. However, in their setup, the quantized code has the same logarithmic term as the maximum likelihood code with no quantization (See also Rissanen (2001)). For parametric TS code, the type class structure in Merhav and Weinberger (2004), corresponds to the point type approach, wherein no quantization is done; i.e. $s = 0$. In this limit, the size of the type class in Merhav and Weinberger (2004) depends on the dimension d' of the derived lattice space (Merhav and Weinberger, 2004, Eq.A3) rather than the model parameter dimension d. We return to this issue in Section 3.7, wherein we show that using point types, the TS code achieves a third-order rate of $\left(\frac{d'}{2} - 1\right) \log n$, which is not tight enough for our purposes due to the fact that d' is in general larger than d.

As a direct consequence of our TS code construction, we have the following finite blocklength achievable bound; it constitutes a modification of Theorem 3 in Kosut and Sankar (2013).

Theorem 4. *Kosut and Sankar (2013) For the TS code*

$$R_n(\epsilon, \phi, p_{\theta^*}) = \frac{1}{n} \lceil \log M(\epsilon) \rceil \tag{3.6}$$

where

$$M(\epsilon) = \inf_{\gamma : \mathbb{P}_{\theta^*}\left(\frac{1}{n} \log |T_{\boldsymbol{\tau}_c(X^n)}| > \gamma\right) \leq \epsilon} \sum_{\substack{\boldsymbol{\tau}_c \in \mathcal{T}_c: \\ \frac{1}{n} \log |T_{\boldsymbol{\tau}_c}| \leq \gamma}} |T_{\boldsymbol{\tau}_c}|. \tag{3.7}$$

3.3 Main Result

Let $H(p_\theta) = \mathbb{E}_\theta \left(\log \frac{1}{p_\theta(X)} \right)$ and $\sigma^2(p_\theta) = \mathbb{V}_\theta \left(\log \frac{1}{p_\theta(X)} \right)$ be the entropy and the varentropy of p_θ. The following theorem exactly characterizes achievable ϵ-rates up to third-order term, as well as asserting that this rate is achievable by the TS code.

Theorem 5. *For any stationary memoryless exponential family of distributions parameterized by* Θ*,*

$$\inf_{\phi} \sup_{\theta \in \Theta} \left[R_n(\epsilon, \phi, p_\theta) - H(p_\theta) - \frac{\sigma(p_\theta)}{\sqrt{n}} Q^{-1}(\epsilon) \right] = \left(\frac{d}{2} - 1 \right) \frac{\log n}{n} + \mathcal{O}\left(\frac{1}{n} \right) \tag{3.8}$$

where the infimum is achieved by the TS code using quantized types.

Example 1. *For the class of all i.i.d. distributions* $d = |\mathcal{X}| - 1$*, and Theorem 5 reduces to the result in Kosut and Sankar (2013).*

3.4 Auxiliary Results

Define

$$\hat{\theta}\left(\boldsymbol{\tau}\right) = \arg\max_{\theta \in \Theta} \left(\langle \theta, \boldsymbol{\tau} \rangle - \psi(\theta) \right). \tag{3.9}$$

Note that since the Hessian matrix of $\psi(\theta)$, $\boldsymbol{\nabla}^2\left(\psi(\theta)\right) = \mathrm{Cov}_\theta\left(\boldsymbol{\tau}(X)\right)$ is positive definite, the log-likelihood function is strictly concave and hence the maximum likelihood $\hat{\theta}(\boldsymbol{\tau})$ is unique.

For notational convenience, we may omit the dependencies on $\boldsymbol{\tau}$ and $\boldsymbol{\tau}_c$ in $\hat{\theta}\left(\boldsymbol{\tau}(x^n)\right)$ and $\hat{\theta}\left(\boldsymbol{\tau}_c(x^n)\right)$, and simply denote them by $\hat{\theta}(x^n)$ and $\hat{\theta}_c(x^n)$, respectively.

The next lemma provides tight upper and lower bounds on the type class size. Beside its exclusive bearing, it is a main component of the achievability proof.

Lemma 4 (Type Class Size)**.** *Let* $\kappa = \wp \frac{\sqrt{d}}{2}$*. For large enough* n*, the size of the type class of* x^n *is bounded as*

$$r(x^n) - 2\kappa s + C' \leq \log |T_{x^n}| \leq r(x^n) + 2\kappa s + C$$

where

$$r(x^n) = -\log p_{\hat{\theta}_c(x^n)}(x^n) - \frac{d}{2} \log n + d \log s$$

is the common part of the upper and lower bounds and C, C' *are constants independent of* n*.*

Proof. For notational convenience, when it is clear from the context, we may suppress the arguments in $\boldsymbol{\tau}_c(x^n)$ and $G(\boldsymbol{\tau}_c(x^n))$ and denote them simply as $\boldsymbol{\tau}_c$ and $G(\boldsymbol{\tau}_c)$.

Motivated by (Merhav and Weinberger, 2004, Eq. A2), we bound $|T_{x^n}|$ as follows:

$$\frac{\mathbb{P}_{\hat{\theta}_c(x^n)}\left\{\boldsymbol{\tau}(X^n) \in G\left(\boldsymbol{\tau}_c(x^n)\right)\right\}}{\max\limits_{\substack{y^n:\\ \boldsymbol{\tau}(y^n)\in G(\boldsymbol{\tau}_c(x^n))}} \mathbb{P}_{\hat{\theta}_c(x^n)}(y^n)} \leq |T_{x^n}| \leq \frac{\mathbb{P}_{\hat{\theta}_c(x^n)}\left\{\boldsymbol{\tau}(X^n) \in G\left(\boldsymbol{\tau}_c(x^n)\right)\right\}}{\min\limits_{\substack{y^n:\\ \boldsymbol{\tau}(y^n)\in G(\boldsymbol{\tau}_c(x^n))}} \mathbb{P}_{\hat{\theta}_c(x^n)}(y^n)}. \tag{3.10}$$

Let $nG(\boldsymbol{\tau}_c) = \{n\mathbf{z} : \mathbf{z} \in G(\boldsymbol{\tau}_c)\}$. It is clear that

$$\mathbb{P}_{\hat{\theta}_c(x^n)}\left\{\boldsymbol{\tau}(X^n) \in G(\boldsymbol{\tau}_c)\right\} = \mathbb{P}_{\hat{\theta}_c(x^n)}\left\{n\boldsymbol{\tau}(X^n) \in nG(\boldsymbol{\tau}_c)\right\}.$$

Exploiting the result in (Stone, 1967, Corollary 1), we have

$$\mathbb{P}_{\hat{\theta}_c(x^n)}\left\{n\boldsymbol{\tau}(X^n) \in nG(\boldsymbol{\tau}_c)\right\} = \frac{s^d}{(2\pi n)^{\frac{d}{2}}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{(n\boldsymbol{\tau}_c - n\boldsymbol{\mu}_c)\cdot\boldsymbol{\Sigma}^{-1}\cdot(n\boldsymbol{\tau}_c - n\boldsymbol{\mu}_c)}{2n}} + o\left(n^{-\frac{d}{2}}\right) \tag{3.11}$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}$ are the mean and the covariance (resp.) of $\boldsymbol{\tau}(X)$ under $\hat{\theta}_c(x^n)$. To proceed, we show that $\boldsymbol{\mu}_c = \boldsymbol{\tau}_c$. We have

$$\begin{aligned}
\hat{\theta}_c(x^n) &= \arg\min_{\theta\in\Theta}\Big(D(p_{\hat{\theta}_c(x^n)}\|p_\theta) + H(p_{\hat{\theta}_c(x^n)})\Big) \\
&= \arg\max_{\theta\in\Theta}\mathbb{E}_{\hat{\theta}_c(x^n)}\Big(\log p_\theta(X)\Big) \\
&= \arg\max_{\theta\in\Theta}\mathbb{E}_{\hat{\theta}_c(x^n)}\Big(\langle\theta, \boldsymbol{\tau}(X)\rangle - \psi(\theta)\Big) \\
&= \arg\max_{\theta\in\Theta}\langle\theta, \boldsymbol{\mu}_c\rangle - \psi(\theta).
\end{aligned}$$

That is, $\hat{\theta}_c(x^n)$ is the maximum likelihood estimate for $\boldsymbol{\mu}_c$ and (by definition (3.9)) $\boldsymbol{\tau}_c$. However, in order to be the maximum likelihood estimate, it must be that the derivative of the log-likelihood function is 0, hence $\nabla\psi(\hat{\theta}_c(x^n)) = \boldsymbol{\mu}_c$ and $\nabla\psi(\hat{\theta}_c(x^n)) = \boldsymbol{\tau}_c$. Therefore $\boldsymbol{\mu}_c$ and $\boldsymbol{\tau}_c$ are equal. Due to (3.11) and $\boldsymbol{\mu}_c = \boldsymbol{\tau}_c$, there exist constants C, C' such that, for large enough n,

$$d\log s - \frac{d}{2}\log n + C' \leq \log p_{\hat{\theta}_c(x^n)}\{\boldsymbol{\tau}(X^n) \in G(\boldsymbol{\tau}_c)\} \leq d\log s - \frac{d}{2}\log n + C. \tag{3.12}$$

On the other hand

$$\log p_{\hat{\theta}_c(x^n)}(x^n) = n\left[\langle \hat{\theta}_c(x^n), \boldsymbol{\tau}(x^n)\rangle - \psi\left(\hat{\theta}_c(x^n)\right)\right].$$

Therefore

$$\max_{\substack{y^n: \\ \boldsymbol{\tau}(y^n)\in G(\boldsymbol{\tau}_c(x^n))}} \log p_{\hat{\theta}_c(x^n)}(y^n) \le \log p_{\hat{\theta}_c(x^n)}(x^n) + 2\kappa s \tag{3.13}$$

and

$$\min_{\substack{y^n: \\ \boldsymbol{\tau}(y^n)\in G(\boldsymbol{\tau}_c(x^n))}} \log p_{\hat{\theta}_c(x^n)}(y^n) \ge \log p_{\hat{\theta}_c(x^n)}(x^n) - 2\kappa s \tag{3.14}$$

where we used $\|\hat{\theta}_c(x^n)\| \le \wp$ and the fact that if $\boldsymbol{\tau}(x^n)$ and $\boldsymbol{\tau}(y^n)$ belong to the same cuboid, then $\|\boldsymbol{\tau}(x^n) - \boldsymbol{\tau}(y^n)\| < \frac{s\sqrt{d}}{n}$. Plugging (3.12,3.13,3.14) in (3.10), the lemma follows. □

Corollary 2. *For large enough n, the size of the type class of x^n with corresponding cuboid center $\boldsymbol{\tau}_c$ is bounded as*

$$nf(\boldsymbol{\tau}_c) - 6\kappa s - C'' \le \log|T_{\boldsymbol{\tau}_c}| \le nf(\boldsymbol{\tau}_c)$$

where, $C'' = C - C'$ and

$$f(\boldsymbol{\tau}) = -\langle \hat{\theta}(\boldsymbol{\tau}), \boldsymbol{\tau}\rangle + \psi\left(\hat{\theta}(\boldsymbol{\tau})\right) - \frac{d}{2n}\log n + \frac{d\log s}{n} + \frac{3\kappa s}{n} + \frac{C}{n}. \tag{3.15}$$

We appeal to the following normal approximation result in order to bound the CDF of the type class size (in the achievability proof) and further CDF of the mixture distribution (in the converse proof) with that of the normal distribution.

Lemma 5 (Asymptotic Normality of Information)**.** *Fix a positive constant α. For a stationary memoryless source, there exists a finite positive constant A, such that for all $n \ge 1$ and z such that $|z| \le \alpha$,*

$$\left|\mathbb{P}_{\theta^*}\left\{\frac{-\log p_{\hat{\theta}(X^n)}(X^n) - nH}{\sqrt{n}\sigma} > z\right\} - Q(z)\right| \le \frac{A}{\sqrt{n}} \tag{3.16}$$

where $H := H(p_{\theta^*})$ *and* $\sigma^2 := \sigma^2(p_{\theta^*})$*, are the entropy and varentropy of the true model* p_{θ^*}*, respectively.*

Proof. See Appendix C □

The following lemma provides a guarantee in approximation of $p_{\hat{\theta}(x^n)}(x^n)$ with $p_{\hat{\theta}_c(x^n)}(x^n)$, which allows us to use the Lemma 5 in the achievability proof.

Lemma 6 (Maximum Likelihood Approximation)**.** *Let* κ *be defined as in Lemma 4. We have*

$$\log p_{\hat{\theta}(x^n)}(x^n) - \log p_{\hat{\theta}_c(x^n)}(x^n) \leq 2\kappa s.$$

Proof. See Appendix D. □

We need the following machinery lemmas for the achievability proof.

Lemma 7. *There exists a Lipschitz constant* K_0 *independent of* n*, so that for any minimal sufficient statistics* $\boldsymbol{\tau}_1$ *and* $\boldsymbol{\tau}_2$*,*

$$|f(\boldsymbol{\tau}_1) - f(\boldsymbol{\tau}_2)| \leq K_0 \|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\|. \tag{3.17}$$

Proof. See Appendix E. □

Let $\omega = \frac{\log|\mathcal{X}| - H}{5}$. Without loss of generality, we may assume that the true model is non-uniform distribution, otherwise TS code (like any other rational code) is obviously optimal. Therefore, $\omega > 0$. Let $0 \leq \lambda < H + \omega$, and $\rho(\lambda) = \mathrm{Vol}\{\boldsymbol{\tau} : f(\boldsymbol{\tau}) \leq \lambda\}$ be the volume of the sub-level sets.

Lemma 8. *There exists a Lipschitz constant* K_1 *so that for all* $0 \leq a, b < H + \omega$*,*

$$|\rho(a) - \rho(b)| \leq K_1 |a - b|.$$

Proof. See Appendix F. □

For our converse proof, we will need the regular value theorem (Balasko, 2009, Prop. 2.3.2) from manifold theory (see also (Robbin and Salamon., 2011, Theorem 9)), stated as follows.

Theorem 6. *Let M and N be smooth manifolds of dimensions m_1, m_2 with $m_1 \geq m_2$. Let $\eta_0 : M \longrightarrow N$ and $b \in N$ be such that for any $a \in \eta_0^{-1}(b)$, the Jacobian matrix of η_0 at a is a surjective map from M to N. Then, $\eta_0^{-1}(b)$ is a $(m_1 - m_2)$-dimensional manifold.*

We have the following Laplace's approximation theorem for the integral of manifolds. We refer the reader to (Wong, 1989, Chap. 9, Th. 3) for a detailed proof. In the converse proof, we use the Laplace's approximation to bound the self information of the mixture distribution.

Theorem 7 (Laplace's Approximation)**.** *Kosut and Sankar (2014b) Let D be a $\tilde{d}-$dimensional differentiable manifold embedded in $\mathbb{R}^m$ and $\eta_1(\cdot)$ and $\eta_2(\cdot)$ be functions that are infinitely differentiable on D. Let*

$$Z(n) = \int_D \eta_2(x) e^{-n\eta_1(x)} dx \tag{3.18}$$

Assume that: (i) the integral in (3.18) converges absolutely for all $n \geq n_0$; (ii) there exists a point x^* in the interior of D such that for every $\epsilon > 0$, $\xi(\epsilon) > 0$ where

$$\xi(\epsilon) = \inf \{\eta_1(x) - \eta_1(x^*) : x \in D \text{ and } |x - x^*| \geq \epsilon\}$$

and (iii) the Hessian matrix $\mathcal{E} = \left(\frac{\partial^2 \eta_1(x)}{\partial x_i \partial x_j}\right)\Big|_{x=x^*}$ is positive definite. Let $F \in \mathbb{R}^{m \times \tilde{d}}$ be an orthonormal basis for the tangent space to D at x^*. Then

$$Z(n) = e^{-n\eta_1(x^*)} \left(\frac{2\pi}{n}\right)^{\frac{\tilde{d}}{2}} \eta_2(x^*) \left|F^T \mathcal{E} F\right|^{-\frac{1}{2}} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

3.5 Proof of Theorem 5

3.5.1 Achievability

In this subsection we bound the third-order coding rate of the quantized implementation of the TS code. We continue from the finite blocklength result in Theorem 4, and evaluate its asymptotic performance.

For the constants C and A in Lemmas 4 and 5, let

$$\gamma = H + \frac{\sigma}{\sqrt{n}} Q^{-1}\Big(\epsilon - \frac{A}{\sqrt{n}}\Big) - \frac{d}{2n}\log n + \frac{d}{n}\log s + \frac{4\kappa s}{n} + \frac{C}{n}. \tag{3.19}$$

Denote

$$\begin{aligned} p_\gamma &:= \mathbb{P}_{\theta^*}\Big[\log|T_{X^n}| > n\gamma\Big] \\ &= \mathbb{P}_{\theta^*}\Big[\log|T_{\boldsymbol{\tau}_c(X^n)}| > n\gamma\Big]. \end{aligned} \tag{3.20}$$

We have

$$\begin{aligned} p_\gamma &\le \mathbb{P}_{\theta^*}\Big[-\log p_{\hat{\theta}_c(x^n)}(X^n) > nH + \sqrt{n}\sigma Q^{-1}\Big(\epsilon - \frac{A}{\sqrt{n}}\Big) + 2\kappa s\Big] && (3.21)\\ &\le \mathbb{P}_{\theta^*}\Big[\frac{-\log p_{\hat{\theta}(x^n)}(X^n) - nH}{\sqrt{n}\sigma} > Q^{-1}\Big(\epsilon - \frac{A}{\sqrt{n}}\Big)\Big] && (3.22)\\ &\le Q\Big(Q^{-1}\big(\epsilon - \frac{A}{\sqrt{n}}\big)\Big) + \frac{A}{\sqrt{n}} && (3.23)\\ &= \epsilon \end{aligned}$$

where (3.21) follows from Lemma 4 and (3.19), (3.22) is from Lemma 6, and (3.23) is a consequence of Lemma 5. Since for γ in (3.19), we have $p_\gamma \le \epsilon$, therefore it satisfies the constraint of (3.7). We can therefore, bound $M(\epsilon)$ defined in (3.7), with this choice of γ. Fixing $\Delta = \frac{1}{n}$, we have

$$
\begin{aligned}
M(\epsilon) &\leq \sum_{\substack{\boldsymbol{\tau}_c \in \mathcal{T}_c: \\ \frac{1}{n}\log|T_{\boldsymbol{\tau}_c}| \leq \gamma}} |T_{\boldsymbol{\tau}_c}| \\
&\leq \sum_{\substack{\boldsymbol{\tau}_c \in \mathcal{T}_c: \\ f(\boldsymbol{\tau}_c) - \frac{6\kappa s + C''}{n} \leq \gamma}} 2^{nf(\boldsymbol{\tau}_c)} \qquad (3.24) \\
&= \sum_{i=0}^{\infty} \sum_{\substack{\boldsymbol{\tau}_c \in \mathcal{T}_c: \\ f(\boldsymbol{\tau}_c) \in \mathcal{A}_i}} 2^{nf(\boldsymbol{\tau}_c)} \\
&\leq \sum_{i=0}^{\infty} |\{\boldsymbol{\tau}_c \in \mathcal{T}_c : f(\boldsymbol{\tau}_c) \in \mathcal{A}_i\}| \cdot 2^{n\gamma + 6\kappa s + C'' - ni\Delta} \qquad (3.25)
\end{aligned}
$$

where (3.24) is from Corollary 2 and $\mathcal{A}_i = \left(\gamma + \frac{6\kappa s + C''}{n} - (i+1)\Delta, \gamma + \frac{6\kappa s + C''}{n} - i\Delta\right]$. The rest of the proof is similar to Kosut and Sankar (2013), however we continue the proof for completeness. We have

$$
\begin{aligned}
|\{\boldsymbol{\tau}_c \in \mathcal{T}_c : f(\boldsymbol{\tau}_c) \in \mathcal{A}_i\}| &= \sum_{\substack{\boldsymbol{\tau}_c \in \mathcal{T}_c: \\ f(\boldsymbol{\tau}_c) \in \mathcal{A}_i}} \frac{\mathrm{Vol}\left(G(\boldsymbol{\tau}_c)\right)}{\left(\frac{s}{n}\right)^d} \qquad (3.26) \\
&= \frac{1}{\left(\frac{s}{n}\right)^d} \mathrm{Vol}\left(\bigcup_{\substack{\boldsymbol{\tau}_c \in \mathcal{T}_c: \\ f(\boldsymbol{\tau}_c) \in \mathcal{A}_i}} G(\boldsymbol{\tau}_c)\right) \qquad (3.27) \\
&\leq \frac{1}{\left(\frac{s}{n}\right)^d} \mathrm{Vol}\left(\bigcup_{\boldsymbol{\tau} \in \mathcal{T}: f(\boldsymbol{\tau}) \in \mathcal{A}_i} G(\boldsymbol{\tau})\right)
\end{aligned}
$$

where (3.26) results from $\mathrm{Vol}\left(G(\boldsymbol{\tau}_c)\right) = \left(\frac{s}{n}\right)^d$, (3.27) follows from disjointness of the cuboids. If $\boldsymbol{\tau} \in G(\boldsymbol{\tau}_c)$, then $\|\boldsymbol{\tau} - \boldsymbol{\tau}_c\| \leq \frac{s\sqrt{d}}{2n}$ and consequently by Lemma 7

$$
|f(\boldsymbol{\tau}) - f(\boldsymbol{\tau}_c)| \leq K_0 \cdot \frac{s\sqrt{d}}{2n} := K_2 \frac{s}{n} \qquad (3.28)
$$

where $K_2 = K_0\frac{\sqrt{d}}{2}$. Therefore, for $a = \gamma + \frac{6\kappa s + C''}{n} - (i+1)\Delta$,

$$\begin{aligned}
|\{\boldsymbol{\tau}_c \in \mathcal{T}_c : f(\boldsymbol{\tau}_c) \in \mathcal{A}_i\}| &\leq \frac{1}{(\frac{s}{n})^d} \cdot \mathrm{Vol}\Big(\bigcup_{a < f(\boldsymbol{\tau}) \leq a+\Delta} G(\boldsymbol{\tau}) \Big) \\
&\leq \frac{1}{(\frac{s}{n})^d} \mathrm{Vol}\left(\left\{ \boldsymbol{\tau} : f(\boldsymbol{\tau}) \in \left(a - K_2\frac{s}{n}, a + \Delta + K_2\frac{s}{n} \right] \right\} \right) \qquad (3.29) \\
&= \frac{1}{\left(\frac{s}{n}\right)^d} \left[\rho\left(a + \Delta + K_2\frac{s}{n} \right) - \rho\left(a - K_2\frac{s}{n} \right) \right] \qquad (3.30)
\end{aligned}$$

where (3.29) is from (3.28). In order to continue from (3.30), recall $\omega = \frac{\log|\mathcal{X}| - H}{5}$. Observe that by (3.19) , $a + K_2\frac{s}{n} + \Delta \leq H + \frac{C}{\sqrt{n}}$, for a positive constant C. Since $\omega > 0$, $H + \frac{C}{\sqrt{n}} < H + \omega$ for large enough n. Similar argument shows that $0 \leq a - K_2\frac{s}{n} < H + \omega$. Therefore boundary conditions of Lemma 8 are satisfied. Continuing from (3.30) and using Lemma 8, we then have

$$|\{\boldsymbol{\tau}_c \in \mathcal{T}_c : f(\boldsymbol{\tau}_c) \in \mathcal{A}_i\}| \leq \frac{K_1}{\left(\frac{s}{n}\right)^d} \cdot \left[\Delta + 2K_2\frac{s}{n} \right]. \qquad (3.31)$$

Applying (3.31) to (3.25), we obtain

$$\begin{aligned}
M(\epsilon) &\leq \sum_{i=0}^{\infty} \frac{K_1}{(\frac{s}{n})^d} \cdot \left[\Delta + 2K_2\frac{s}{n} \right] \cdot 2^{n\gamma + 6\kappa s + C'' - ni\Delta} \\
&= \frac{n^d}{s^d} \cdot \left[\Delta + 2K_2\frac{s}{n} \right] \cdot 2^{n\gamma + 6\kappa s + C''} \cdot \frac{K_1}{1 - 2^{-n\Delta}}.
\end{aligned}$$

From (3.19) and since $s > 0$ is a constant and $\Delta = \frac{1}{n}$, we obtain

$$\log M(\epsilon) \leq nH + \sigma\sqrt{n}Q^{-1}(\epsilon) + \left(\frac{d}{2} - 1 \right) \log n + \mathcal{O}(1).$$

3.5.2 Converse

For a parameter vector $\theta \in \Theta$, define $J(\theta) = nH(p_\theta) + \sigma(p_\theta)\sqrt{n}Q^{-1}(\epsilon)$. We first rewrite the entropy function as follows:

$$\begin{aligned} H(p_\theta) &= -\sum_{x\in\mathcal{X}} p_\theta(x)\log p_\theta(x) \\ &= -\sum_{x\in\mathcal{X}} p_\theta(x)\left(\langle\theta, \boldsymbol{\tau}(x)\rangle - \psi(\theta)\right) \qquad (3.32) \\ &= -\langle\theta, \mathbb{E}_\theta(\boldsymbol{\tau}(x))\rangle + \psi(\theta) \\ &= -\langle\theta, \nabla\psi(\theta)\rangle + \psi(\theta) \qquad (3.33) \end{aligned}$$

where (3.32) is from (3.1) and (3.33) is from $\mathbb{E}_\theta(\boldsymbol{\tau}(x)) = \nabla\psi(\theta)$ Jordan (2003). Taking derivative of (3.33) with respect to θ, we obtain

$$\nabla H(p_\theta) = -\theta\nabla^2\psi(\theta). \qquad (3.34)$$

Since $\nabla^2\psi(\theta) = \mathrm{Cov}(\boldsymbol{\tau}(X))$ is positive definite, (3.34) vanishes only at the uniform distribution $\theta_u = (0, \cdots, 0)$. Since Θ has nonempty interior, let θ_0 be a point in the interior of Θ with $J(\theta_0) \neq J(\theta_u)$. Define

$$\Theta_0 := \{\theta \in \Theta : J(\theta) = J(\theta_0)\}.$$

As $\theta_u \notin \Theta_0$, $\nabla H(p_\theta)$ is nonzero for all parameters $\theta \in \Theta_0$. Therefore, for large enough n, $\nabla J(\theta)$ is also nonzero for all $\theta \in \Theta_0$. Hence, the Jacobian of $J(\cdot)$ at any point in the set $J^{-1}(J(\theta_0))$ is a surjective map from Θ_0 to $\mathbb{R}$. Theorem 6 then implies that Θ_0 is a $(d-1)$-dimensional manifold.

In order to prove the converse, it suffices to show that

$$\sup_{\theta\in\Theta_0} R_n(\epsilon, \phi, p_\theta) \geq \frac{J(\theta_0)}{n} + \left(\frac{d}{2} - 1\right)\frac{\log n}{n} - \mathcal{O}\left(\frac{1}{n}\right).$$

Let $\overline{p}(x^n)$ be the mixture distribution with uniform prior among n-length *i.i.d.* distributions with marginals parametrized by Θ_0, i.e.

$$\overline{p}(x^n) = \frac{1}{\mathrm{Vol}(\Theta_0)} \int_{\theta \in \Theta_0} p_\theta(x^n) d\theta \tag{3.35}$$

where $\mathrm{Vol}(\cdot)$ is the d-dimensional volume. For any $\gamma > 0$, applying Theorem 3 in Kosut and Sankar (2014b) gives

$$\epsilon + 2^{-\gamma} \geq \inf_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\iota_{\overline{p}}(X^n) \geq k + \gamma \right) \tag{3.36}$$

where $\iota_{\overline{p}}(X^n) := -\log \overline{p}(X^n)$ is the self information of the mixture distribution. We then provide a lower bound for the self information. We may rewrite (3.35) as

$$\overline{p}(x^n) = \frac{1}{\mathrm{Vol}(\Theta_0)} \int_{\theta \in \Theta_0} 2^{-g(\theta)} d\theta$$

where $g(\theta) := -\log p_\theta(x^n)$. Since Θ_0 is a $(d-1)$-dimensional manifold, application of the Laplace's approximation of integrals (Theorem 7) yields

$$\overline{p}(x^n) = \frac{1}{\mathrm{Vol}(\Theta_0)} 2^{-g(\hat{\theta})} \left(\frac{2\pi}{n} \right)^{\frac{d-1}{2}} \left| F^T \mathcal{E} F \right|^{-\frac{1}{2}} \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right) \tag{3.37}$$

where $\hat{\theta} := \hat{\theta}(x^n)$ is the maximum likelihood estimate of θ for x^n. Continuing from (3.36) for a constant $C > 0$, we obtain

$$\begin{aligned}
&\epsilon + 2^{-\gamma} \\
&\geq \inf_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\iota_{\overline{p}}(X^n) \geq k + \gamma \right) \\
&\geq \inf_{\theta \in \Theta_0} \mathbb{P}_\theta \left(-\log p_{\hat{\theta}}(X^n) + \frac{d-1}{2} \log n + C \geq k + \gamma \right) \qquad (3.38) \\
&= \inf_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\frac{-\log p_{\hat{\theta}}(X^n) - nH(p_\theta)}{\sqrt{n}\sigma} \geq \frac{k + \gamma - \frac{d-1}{2} \log n - C - nH(p_\theta)}{\sqrt{n}\sigma} \right) \\
&\geq Q\left(\frac{k + \gamma - \frac{d-1}{2} \log n - C - nH(p_\theta)}{\sqrt{n}\sigma} \right) - \frac{A}{\sqrt{n}} \qquad (3.39)
\end{aligned}$$

where (3.38) is due to (3.37) and the definition of $g(\cdot)$, while (3.39) is from Lemma 5. Setting $\gamma = \frac{1}{2}\log n$ and rearranging gives

$$\frac{k}{n} \geq \inf_{\theta \in \Theta_0} H(p_\theta) + \frac{\sigma(p_\theta)}{\sqrt{n}} Q^{-1}\left(\epsilon + \frac{A+1}{\sqrt{n}}\right) + \left(\frac{d}{2} - 1\right)\frac{\log n}{n} + \frac{C}{n}.$$

Recalling that $H(p_\theta) + \frac{\sigma(p_\theta)}{\sqrt{n}} Q^{-1}(\epsilon)$ is fixed at $\frac{J(\theta_0)}{n}$ for all $\theta \in \Theta_0$ and that $\frac{k}{n} = \max_{\theta \in \Theta_0} R_n(\epsilon, \phi, p_\theta)$, theorem follows.

3.6 Parametric Markov Class

We now consider extensions to the class of parametric Markov models. Let $\mathcal{M}$ be the exponential family of first-order, stationary, irreducible and aperiodic Markov sources, parametrized by a d-dimensional parameter vector $\theta \in \Theta_{\mathcal{M}} \subset \mathbb{R}^d$. Transition probabilities of the distribution $p_\theta \in \mathcal{M}$ has the following exponential structure

$$p_\theta(x_i|x_{i-1}) = 2^{\langle \theta, \boldsymbol{\tau}(x_{i-1}, x_i)\rangle - \psi(\theta)} \tag{3.40}$$

where $\boldsymbol{\tau} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the vector of sufficient statistics.

Similar to Merhav and Neuhoff (1992), we assume that the initial source symbol x_0 is fixed and known to both the encoder and the decoder. From (3.40), the probability of a sequence x^n drawn according to the first-order Markov source $p_\theta \in \mathcal{M}$ in the exponential family takes the form

$$\begin{aligned} p_\theta(x^n) &= \prod_{i=1}^{n} p_\theta\left(x_i|x_{i-1}\right) \\ &= \prod_{i=1}^{n} 2^{\langle \theta, \boldsymbol{\tau}(x_{i-1}, x_i)\rangle - \psi(\theta)} \\ &= 2^{n[\langle \theta, \boldsymbol{\tau}(x^n)\rangle - \psi(\theta)]} \end{aligned}$$

where $\boldsymbol{\tau}(x^n) = \frac{\sum_{i=1}^{n} \boldsymbol{\tau}(x_{i-1}, x_i)}{n} \in \mathbb{R}^d$ is a minimal sufficient statistic. Through the same approach as in Section 3.2, we partition the convex hull of the space of minimal suffi-

cient statistics into cuboids of side length $\frac{s}{n}$ defined as in (3.4). We then characterize quantized type classes as in (3.5).

Let

$$H(p_\theta) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_\theta \left[\log \frac{1}{p_\theta(X^n)} \right] \tag{3.41}$$

and

$$\sigma^2(p_\theta) = \lim_{n\to\infty} \frac{1}{n} \mathbb{V}_\theta \left[\log \frac{1}{p_\theta(X^n)} \right] \tag{3.42}$$

be the entropy and the varentropy rate of the Markov process parametrized by θ, respectively. The following theorem characterizes the fundamental limits of universal one-to-one compression of parametric Markov sources, as well as asserting that the TS code is optimal up to the third-order term.

Theorem 8. *For any first-order, stationary, irreducible and aperiodic Markov exponential model class parametrized by* $\Theta_{\mathcal{M}}$

$$\inf_{\phi} \sup_{\theta\in\Theta_{\mathcal{M}}} \left[R_n(\epsilon, \phi, p_\theta) - H(p_\theta) - \frac{\sigma(p_\theta)}{\sqrt{n}} Q^{-1}(\epsilon) \right] = \left(\frac{d}{2} - 1 \right) \frac{\log n}{n} + \mathcal{O}\left(\frac{1}{n} \right).$$

where the infimum is achieved by the quantized type class implementation of the TS code.

Proof. Let $Y_i = (X_{i-1}, X_i)$ be a random vector defined by overlapping blocks of $\{X_n\}$. Since X_n form a Markov chain, so does $\{Y_n\}$. The proof follows the same lines as those in the proof of the parametric *i.i.d.* class $\mathcal{P}$, with $\boldsymbol{\tau}(Y_n)$ playing the role of $\boldsymbol{\tau}(X_n)$. The only deviations from the memoryless proof occur in lines (3.11), (3.23) and (3.39). As a counterpart of the *i.i.d.* ratio limit theorem of (3.11) for a Markov sources, we may use Theorem 8 of Korshunov (2001), which states that

$$p_{\hat{\theta}_c(x^n)} \left\{ n\boldsymbol{\tau}(Y^n) \in nG(\boldsymbol{\tau}_c) \right\} = \frac{s^d}{(2\pi n)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{\left\langle (x-n\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}, x-n\boldsymbol{\mu} \right\rangle}{2n}} + o\left(n^{-\frac{d}{2}} \right)$$

where $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are the covariance and mean of the stationary distribution of the Markov chain, respectively. Finally (3.23) and (3.39) can be derived from the Markov

version of the normal approximation inequality stated below. The proof is the same as in Appendix C.

Lemma 9 (Asymptotic Normality of Information)**.** *Fix a positive constant α. For a first-order, stationary, irreducible and aperiodic Markov source, there exists a finite positive constant A' such that for all $n \geq 1$ and z such that $|z| \leq \alpha$,*

$$\left| \mathbb{P}_{\theta^*} \left\{ \frac{-\log p_{\hat{\theta}(X^n)}(X^n) - nH}{\sqrt{n}\sigma} > z \right\} - Q(z) \right| \leq \frac{A'}{\sqrt{n}} \tag{3.43}$$

where $H := H(p_{\theta^*})$ and $\sigma^2 := \sigma^2(p_{\theta^*})$, are the entropy and varentropy rate of the true model, p_{θ^*}, respectively.

The rest of the proof is the same as the *i.i.d.* case and we omit it due to similarity.

□

Example 2. *For the class of all first-order stationary, irreducible and aperiodic Markov sources $d = |\mathcal{X}|(|\mathcal{X}| - 1)$, and Theorem 8 reduces to the result in Iri and Kosut (2015).*

3.7 Type Size Code with Point Type Classes

In this section we analyze the performance of the point type class implementation of the TS code. For a sequence $x^n \in \mathcal{X}^n$, define the point type class containing x^n as

$$T_{x^n} = \{y^n \in \mathcal{X}^n : p_\theta(x^n) = p_\theta(y^n) \text{ for all } \theta \in \Theta\} \tag{3.44}$$

the set of all n-length sequences $y^n \in \mathcal{X}^n$ equiprobable with x^n, simultaneously under all models in $\mathcal{P}$. Consequently, (3.2) enforces two sequences to be in the same type class if and only if their minimal sufficient statistics are equal. Hence, from a geometric perspective, point type classes correspond to zero sidelength $s = 0$ in Figure 3.1, i.e. type classes are points in the space of minimal sufficient statistics. We first review

the derivation of the size of a point type class from Merhav and Weinberger (2004). We then provide upper and lower bounds for the asymptotic rate of the TS code with point type class implementation, showing that the TS code performs strictly worse for $s = 0$ in terms of third-order coding rate.

Let $\boldsymbol{\tau}(x)[j]$, $j = 1, \cdots, d$, be the j-th component of the d-dimensional vector $\boldsymbol{\tau}(x)$. For any index $j = 1, \cdots, d$, there exists a fixed real number $\beta[j][0]$ and r_j pairwise incommensurable real numbers $\beta[j][t]$, $t = 1, \cdots, r_j$, such that regardless of the observed sample $x \in \mathcal{X}$, $\boldsymbol{\tau}(x)[j]$ can be uniquely decomposed as Merhav and Weinberger (2004)

$$\boldsymbol{\tau}(x)[j] = \beta[j][0] + \sum_{t=1}^{r_j} \beta[j][t] \tilde{L}(x)[j][t] \tag{3.45}$$

where $\tilde{L}(x)[j][t]$, $t = 1, \cdots, r_j$, are integers depending on the sample x through $\boldsymbol{\tau}(x)[j]$. The decomposition (3.45) defines a unique one-to-one mapping between the real-valued $\boldsymbol{\tau}(x)[j]$ and r_j integers $\tilde{L}(x)[j][t]$. Concatenating the corresponding unique integers $\tilde{L}(x)[\cdot][\cdot]$, each d-dimensional vector $\boldsymbol{\tau}(x)$ corresponds to a unique integer-valued vector $\tilde{\boldsymbol{L}}(x) \in \mathbb{Z}^{\sum_{j=1}^{d} r_j}$. For all $j = 1 \cdots d$, we may choose without loss of generality $\beta[j][0] = \boldsymbol{\tau}(1)[j]$. With this choice we always have $\tilde{\boldsymbol{L}}(1) = (0, \cdots, 0)^T$. Let d', which is called the dimensionality of the type class in Merhav and Weinberger (2004), be the rank of the matrix $\tilde{\mathbb{L}} = \left[\tilde{\boldsymbol{L}}(2) - \tilde{\boldsymbol{L}}(1) \quad \cdots \quad \tilde{\boldsymbol{L}}(|\mathcal{X}|) - \tilde{\boldsymbol{L}}(1)\right]$. Therefore, there are d' linearly independent rows in $\tilde{\mathbb{L}}$. Let the indices of the linearly independent rows be $i_1, \cdots, i_{d'}$. For any $x \in \mathcal{X}$, define d'-dimensional vector $\boldsymbol{L}(x)$ as $\boldsymbol{L}(x)[j] = \tilde{\boldsymbol{L}}(x)[i_j]$ for $j = 1 \cdots d'$. Since the other rows are linear combination of the independent rows, we can denote this transformation as $\tilde{\mathbb{L}} = \boldsymbol{R}\mathbb{L}$, where $\boldsymbol{R}$ is a $\sum_{j=1}^{d} r_j \times d'$ matrix and $\mathbb{L}$ is a full-rank $d' \times (|\mathcal{X}| - 1)$ dimensional matrix $\mathbb{L} = \left[\boldsymbol{L}(2) - \boldsymbol{L}(1) \quad \cdots \quad \boldsymbol{L}(|\mathcal{X}|) - \boldsymbol{L}(1)\right]$. Since $\tilde{\boldsymbol{L}}(1) = \boldsymbol{L}(1) = \boldsymbol{0}$, there is a one to one correspondence between $\boldsymbol{L}(x)$ and $\tilde{\boldsymbol{L}}(x)$ and consequently between $\boldsymbol{L}(x)$ and $\boldsymbol{\tau}(x)$.

Note that $d' \geq d$, and in many cases the inequality is strict. The main finding of this section is that d' is the critical dimension for the behavior of the TS code under point type classes, rather than d. Since d' may be larger than d, the performance of the TS code with point type classes may be strictly worse than that with quantized type classes.

Let $\mathbf{b}$ be a $d \times 1$ column vector containing $\beta[j][0]$'s for $j = 1, \cdots, d$ and $\mathbb{A}$ is a $d \times \sum_{j=1}^{d} r_j$ block diagonal matrix containing $\beta[j][t]$'s in (3.45). For real-valued vector $\ell \in \mathbb{R}^{d'}$, let $\boldsymbol{\tau}(\ell) = \mathbf{b} + \mathbb{A}\boldsymbol{R}\ell$. For a constant $C > 0$ to be defined later, define $f_0(\ell)$ as follows:

$$f_0(\ell) = -\frac{1}{n}\left(\left\langle \hat{\theta}(\boldsymbol{\tau}(\ell)), \boldsymbol{\tau}(\ell) \right\rangle - \psi\left(\hat{\theta}(\boldsymbol{\tau}(\ell))\right)\right) - \frac{d'}{2n}\log 2\pi n + \frac{C}{n} \tag{3.46}$$

$$= -\frac{1}{n}\left(\left\langle \hat{\theta}(\mathbf{b} + \mathbb{A}\boldsymbol{R}\ell), \mathbf{b} + \mathbb{A}\boldsymbol{R}\ell \right\rangle - \psi\left(\hat{\theta}\left(\mathbf{b} + \mathbb{A}\boldsymbol{R}\ell\right)\right)\right) - \frac{d'}{2n}\log 2\pi n + \frac{C}{n}. \tag{3.47}$$

For a sequence x^n, define $\boldsymbol{L}(x^n)$ similar to (3.3) as

$$\boldsymbol{L}(x^n) = \frac{\sum_{i=1}^{n} \boldsymbol{L}(x_i)}{n} \tag{3.48}$$

and let $\mathcal{L} = \{\boldsymbol{L}(x^n) : x^n \in \mathcal{X}^n\}$ be the set of lattice points. Throughout, $\boldsymbol{L} \in \mathbb{Z}^{d'}$ denotes an integer-valued lattice point, while $\ell \in \mathbb{R}^{d'}$ denotes real-valued d'-dimensional vector.

The size of a point type class is derived in Merhav and Weinberger (2004), which we reproduce it in Appendix G for completeness. Moreover, we show that the third-order term in their result is a constant to obtain the following lemma.

Lemma 10. *For large enough n, the size of the point type class containing x^n with $\boldsymbol{L}(x^n) = \boldsymbol{L}$, is bounded as*

$$nf_0(\boldsymbol{L}) - 2C \leq \log |T_{x^n}| \leq nf_0(\boldsymbol{L}) \tag{3.49}$$

where C is the constant in (3.46, 3.47).

Proof. See Appendix G. □

The following is our main theorem for this section, characterizing the exact performance of the TS code with point type classes up to third-order.

Theorem 9. *Let ϕ_0 be the point type class implementation of the TS code. The ϵ-coding rate of ϕ_0, for all $\theta \in \Theta$ is given by*

$$R_n(\epsilon, \phi_0, p_\theta) = H(p_\theta) + \frac{\sigma(p_\theta)}{\sqrt{n}} Q^{-1}(\epsilon) + \left(\frac{d'}{2} - 1\right) \frac{\log n}{n} + \mathcal{O}\left(\frac{1}{n}\right). \tag{3.50}$$

Proof. The achievability proof is similar to Section 3.5, hence we only highlight the differences. Again for simplicity, we denote $H = H(p_{\theta^*})$ and $\sigma = \sigma(p_{\theta^*})$ as the entropy and the varentropy of the underlying model p_{θ^*}, respectively. Let

$$\gamma' = H + \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\epsilon - \frac{A}{\sqrt{n}}\right) - \frac{d'}{2n} \log(2\pi n) + \frac{C}{n}. \tag{3.51}$$

We now show that for this choice of γ', $p_{\gamma'} \leq \epsilon$, where $p_{\gamma'}$ is defined as in (3.20). We have

$$\begin{aligned} p_{\gamma'} &= \mathbb{P}_{\theta^*}\left[\log |T_{X^n}| > n\gamma'\right] \\ &= \mathbb{P}_{\theta^*}\left[\frac{-\log p_{\hat{\theta}}(x^n) - nH}{\sigma\sqrt{n}} > Q^{-1}\left(\epsilon - \frac{A}{\sqrt{n}}\right)\right] && (3.52) \\ &\leq Q\left(Q^{-1}\left(\epsilon - \frac{A}{\sqrt{n}}\right)\right) + \frac{A}{\sqrt{n}} && (3.53) \\ &= \epsilon \end{aligned}$$

where (3.52) follows from (3.49, 3.46, 3.51) by noticing that

$$f_0(\boldsymbol{L}) = -\frac{1}{n} \log p_{\hat{\theta}(x^n)}(x^n) - \frac{d'}{2n} \log(2\pi n) + \frac{C}{n}$$

for any x^n with $\boldsymbol{L}(x^n) = \boldsymbol{L}$, and (3.53) is an application of Lemma 5.

Recall that there is a one-to-one correspondence between T_{x^n} and $\boldsymbol{L}(x^n)$, hence we can denote T_{x^n} as $T_{\boldsymbol{L}(x^n)}$. Furthermore, once x^n is understood from the context,

we simplify $T_{\boldsymbol{L}(x^n)}$ and rewrite it as $T_{\boldsymbol{L}}$. We can then reformulate the equation for $M(\epsilon)$ in (3.7) for point type classes. We can achieve this, simply by replacing $\boldsymbol{\tau}_c(X^n)$ with $\boldsymbol{L}(x^n)$ as the representative of the type class.

We then bound $M(\epsilon)$ in (3.7) with the choice of γ' in (3.51). Through the same approach as in Subsection 3.5.1, one can show that

$$M(\epsilon) \leq \sum_{i=0}^{\infty} |\{\boldsymbol{L} \in \mathcal{L} : f_0(\boldsymbol{L}) \in \mathcal{A}_i'\}| \cdot 2^{\{n\gamma' + 2C - ni\Delta\}} \tag{3.54}$$

where $\mathcal{A}_i' = \left(\gamma' + \frac{2C}{n} - (i+1)\Delta, \gamma' + \frac{2C}{n} - i\Delta\right]$ and C is the constant in (3.49). We now evaluate $|\{\boldsymbol{L} \in \mathcal{L} : f_0(\boldsymbol{L}) \in \mathcal{A}_i'\}|$. Define a 2-norm ball of radius r around a point $\ell_0 \in \mathbb{R}^{d'}$ as

$$B_r(\ell_0) = \left\{\ell \in \mathbb{R}^{d'} : \|\ell - \ell_0\| < r\right\}. \tag{3.55}$$

In the sequel we use $\boldsymbol{L}$ as the lattice points in $\mathcal{L}$, while we reserve the notation ℓ for points in the convex hull of $\mathcal{L}$ which we denote by $\mathfrak{L} = \text{conv}(\mathcal{L})$. Observe that for any two different points $\boldsymbol{L}_1, \boldsymbol{L}_2 \in \mathcal{L}$, $\|\boldsymbol{L}_1 - \boldsymbol{L}_2\| \geq \frac{1}{n}$, and therefore, $B_{\frac{1}{2n}}(\boldsymbol{L}_1)$ and $B_{\frac{1}{2n}}(\boldsymbol{L}_2)$ are disjoint. Since the convex hull $\mathfrak{L}$ is a d'-dimensional space, there exists a constant $C > 0$ (its precise value is $\frac{\pi^{\frac{d'}{2}}}{2^{d'}\Gamma(\frac{d'}{2}+1)}$ Ren (1994)) such that

$$\text{Vol}\left(B_{\frac{1}{2n}}(\boldsymbol{L})\right) = \frac{C}{n^{d'}}. \tag{3.56}$$

Therefore

$$\begin{aligned} |\{\boldsymbol{L} \in \mathcal{L} : f_0(\boldsymbol{L}) \in \mathcal{A}_i'\}| &= \sum_{\substack{\boldsymbol{L}\in\mathcal{L} \\ f_0(\boldsymbol{L})\in\mathcal{A}_i'}} \frac{n^{d'}}{C} \text{Vol}\left(B_{\frac{1}{2n}}(\boldsymbol{L})\right) \\ &= \frac{n^{d'}}{C} \text{Vol}\left(\bigcup_{\substack{\boldsymbol{L}\in\mathcal{L} \\ f_0(\boldsymbol{L})\in\mathcal{A}_i'}} B_{\frac{1}{2n}}(\boldsymbol{L})\right) \\ &\leq \frac{n^{d'}}{C} \text{Vol}\left(\bigcup_{\substack{\ell\in\mathfrak{L} \\ f_0(\ell)\in\mathcal{A}_i'}} B_{\frac{1}{2n}}(\boldsymbol{L})\right). \end{aligned} \tag{3.57}$$

where (3.57) follows from disjointness of the balls. Proceeding as in Subsection 3.5.1, it is straightforward to show that for a constant $C > 0$

$$| \{ \boldsymbol{L} \in \mathcal{L} : f_0(\boldsymbol{L}) \in \mathcal{A}_i' \} | \leq C n^{d'-1}. \tag{3.58}$$

The rest of the proof is similar to the Subsection 3.5.1, which we omit due to similarity.

We now provide a converse for the performance of the Type Size code with point type classes. We can rewrite the corresponding finite blocklength result (3.7) for point type classes as

$$M(\epsilon) = \inf_{\gamma' : p_{\gamma'} \leq \epsilon} v(\gamma'), \tag{3.59}$$

where $p_{\gamma'}$ is defined as in (3.20) and

$$v(\gamma') = \sum_{\substack{\boldsymbol{L} \in \mathcal{L}: \\ \frac{1}{n} \log |T_{\boldsymbol{L}}| \leq \gamma'}} |T_{\boldsymbol{L}}|. \tag{3.60}$$

Notice that $v(\gamma')$ is non-decreasing function of γ', while $p_{\gamma'}$ is non-increasing function of γ'. Therefore, if for some γ_0', $p_{\gamma_0'} > \epsilon$, then one can conclude that

$$M(\epsilon) \geq v(\gamma_0'). \tag{3.61}$$

We then show that $p_{\gamma_0'} > \epsilon$ for the following choice of γ_0'

$$\gamma_0' = H + \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\epsilon + \frac{A+1}{\sqrt{n}}\right) - \frac{d'}{2n} \log(2\pi n) - \frac{C}{n} \tag{3.62}$$

where A is the constant in Lemma 5 and C is the constant in (3.49). Indeed

$$p_{\gamma_0'} \geq \mathbb{P}_{\theta^*}\left[-\frac{1}{n} \log p_{\hat{\theta}(X^n)}(X^n) - \frac{d'}{2n} \log(2\pi n) - \frac{C}{n} > \gamma_0' \right] \tag{3.63}$$

$$= \mathbb{P}_{\theta^*}\left[\frac{-\log p_{\hat{\theta}(X^n)}(X^n) - nH}{\sigma\sqrt{n}} > Q^{-1}\left(\epsilon + \frac{A+1}{\sqrt{n}}\right) \right] \tag{3.64}$$

$$> \epsilon \tag{3.65}$$

where (3.63) is from the type class size bound (3.49) and the definition of $p_{\gamma_0'}$ in (3.20), (3.64) is from the choice of γ_0' in (3.62), and (3.65) is a consequence of Lemma 5. Continuing from (3.61), we may write

$$\begin{aligned} M(\epsilon) &\geq \sum_{\substack{\boldsymbol{L}\in\mathcal{L} \\ \frac{1}{n}\log|T_{\boldsymbol{L}}|\leq\gamma_0'}} |T_{\boldsymbol{L}}| \\ &\geq \sum_{\substack{\boldsymbol{L}\in\mathcal{L} \\ f_0(\boldsymbol{L})\leq\gamma_0'}} 2^{nf_0(\boldsymbol{L})-2C} \end{aligned} \tag{3.66}$$

where (3.66) exploits the bounds for the type class size (3.49). For $\Delta = \frac{1}{n}$, (3.66) can simply be lower bounded as follows by restricting the summation to $\boldsymbol{L}$ in $\mathcal{A}_0$, where $\mathcal{A}_0 = \{\boldsymbol{L} \in \mathcal{L} : \gamma_0' - \Delta < f_0(\boldsymbol{L}) \leq \gamma_0'\}$

$$M(\epsilon) \geq |\mathcal{A}_0| \cdot 2^{n\gamma_0' - n\Delta - 2C}. \tag{3.67}$$

We now provide a lower bound on $|\mathcal{A}_0|$. Let $\tilde{\mathcal{A}}_0 = \{\ell \in \mathfrak{L} : \gamma_0' - \Delta < f_0(\ell) \leq \gamma_0'\}$.

Lemma 11. *There exists a constant C such that*

$$\frac{\mathrm{Vol}\left(\bigcup_{\ell\in\tilde{\mathcal{A}}_0} B_{\frac{1}{2n}}(\ell)\right)}{\mathrm{Vol}\left(\bigcup_{\boldsymbol{L}\in\mathcal{A}_0} B_{\frac{1}{2n}}(\boldsymbol{L})\right)} \leq C. \tag{3.68}$$

Proof. See Appendix H. □

We then have

$$\begin{aligned} |\mathcal{A}_0| &= \sum_{\boldsymbol{L}\in\mathcal{A}_0} \frac{\mathrm{Vol}\left(B_{\frac{1}{2n}}(\boldsymbol{L})\right)}{\mathrm{Vol}\left(B_{\frac{1}{2n}}(\boldsymbol{L})\right)} \\ &= Cn^{d'} \sum_{\boldsymbol{L}\in\mathcal{A}_0} \mathrm{Vol}\left(B_{\frac{1}{2n}}(\boldsymbol{L})\right) && (3.69) \\ &= Cn^{d'} \mathrm{Vol}\left(\bigcup_{\boldsymbol{L}\in\mathcal{A}_0} B_{\frac{1}{2n}}(\boldsymbol{L})\right) && (3.70) \\ &\geq n^{d'} \frac{\mathrm{Vol}\left(\bigcup_{\ell\in\tilde{\mathcal{A}}_0} B_{\frac{1}{2n}}(\ell)\right)}{C} && (3.71) \end{aligned}$$

where (3.69) follows from (3.56) (recall that the letter C may denote different constants), (3.70) is due to disjointness of the balls, and (3.71) is a consequence of Lemma 11. Define,

$$\rho_0(\lambda) = \mathrm{Vol}\{\ell \in \mathfrak{L} : f_0(\ell) \leq \lambda\}. \tag{3.72}$$

We need the following technical lemma, which we prove in Appendix I.

Lemma 12. *There exists a positive constant K_4, such that for all $\gamma_0' - \Delta \leq \lambda \leq \gamma_0'$ we have*

$$\left| \frac{d}{d\lambda} \rho_0(\lambda) \right| \geq K_4.$$

Recalling the definition of $\tilde{\mathcal{A}}_0$, we may continue from (3.71) and write

$$\begin{aligned}
|\mathcal{A}_0| &\geq \frac{n^{d'}}{C} \mathrm{Vol}\left(\cup_{\ell : \gamma_0' - \Delta < f_0(\ell) \leq \gamma_0'} B_{\frac{1}{2n}}(\ell) \right) \\
&\geq \frac{n^{d'}}{C} \mathrm{Vol}\left(\{\ell : f_0(\ell) \in (\gamma_0' - \Delta, \gamma_0']\} \right) && (3.73) \\
&= \frac{n^{d'}}{C} \left(\rho_0(\gamma_0') - \rho_0(\gamma_0' - \Delta) \right) && (3.74) \\
&\geq \frac{n^{d'}}{C} K_4 \Delta && (3.75)
\end{aligned}$$

where (3.73) is by lower bounding the volume of the ball-covering of $\tilde{\mathcal{A}}_0$ by the volume of $\tilde{\mathcal{A}}_0$ itself, (3.74) is from the definition of ρ_0 and (3.75) is from Lemma 12.

Continuing from (3.67), we have

$$\begin{aligned}
M(\epsilon) &\geq \frac{n^{d'}}{C} K_4 \Delta \cdot 2^{n\gamma_0' - n\Delta - 2C} && (3.76) \\
&= C n^{d'-1} 2^{n\gamma_0' - 2C - 1} && (3.77)
\end{aligned}$$

where (3.76) is from (3.75), and (3.77) is from $\Delta = \frac{1}{n}$. Replacing γ_0' by (3.62), we obtain

$$\log M(\epsilon) \geq nH + \sigma\sqrt{n} Q^{-1}(\epsilon) + \left(\frac{d'}{2} - 1 \right) \log n + \mathcal{O}(1).$$

□

3.8 Two Stage Code

Using the same steps as in Kosut and Sankar (2014b), we derive the asymptotic third-order coding rate of the Two-Stage code for compression of parametric sources as stated in the following theorem.

Theorem 10. *The overflow coding rate of the Two-Stage code for compression of a d-dimensional parametric exponential family of distributions $\mathcal{P}$, with i.i.d. data generation mechanism is given by*

$$R_n(\Phi^{2^{\mathcal{S}}}, \epsilon, p_{\theta^*}) = H(p_{\theta^*}) + \frac{\sigma(p_{\theta^*})}{\sqrt{n}} Q^{-1}(\epsilon) + \frac{d}{2}\frac{\log n}{n} + \mathcal{O}(\frac{1}{n}) \tag{3.78}$$

Proof. First we derive the number of types in the parametric model setup of section 3.1. Number of type classes is the number of cuboids in the partition of space $\mathcal{T}$. We cover $\mathcal{T}$ by cuboids of sidelength $\frac{h}{n}$. Number of such cuboids (number of type classes) is thus bounded as

$$|\mathcal{P}_n(\mathcal{X})| \leq \frac{(\|\tau_{max}\| + \frac{h}{n})^d - (\|\tau_{min}\| - \frac{h}{n})^d}{\left(\frac{h}{n}\right)^d} = C_{\mathcal{T}} n^d$$

where $C_{\mathcal{T}}$ is a constant independent of n. Hence

$$\log |\mathcal{P}_n(\mathcal{X})| \leq d \log n + \mathcal{O}(1)$$

Thus the number of bits required for the first stage is $d \log n + \mathcal{O}(1)$. The rest of the proof is similar to Kosut and Sankar (2014a), and we omit it due to similarity. □

Chapter 4

VARIABLE TO FIXED LENGTH CODING

4.1 Problem Statement

Let Θ be a compact subset of $\mathbb{R}^d$. Probability distributions in an exponential family can be expressed in the form Merhav and Weinberger (2004)

$$p_\theta(x) = 2^{\langle \theta, \boldsymbol{\tau}(x) \rangle - \psi(\theta)} \tag{4.1}$$

where $\theta \in \Theta$ is the d-dimensional parameter vector, $\boldsymbol{\tau}(x) : \mathcal{X} \to \mathbb{R}^d$ is the vector of sufficient statistics and $\psi(\theta)$ is the normalizing factor. Let the model class $\mathcal{P} = \{p_\theta, \theta \in \Theta\}$, be the exponential family of distributions over the finite alphabet $\mathcal{X} = \{1, \cdots, |\mathcal{X}|\}$, parameterized by $\theta \in \Theta \subset \mathbb{R}^d$, where d is the degrees of freedom in the minimal description of $p_\theta \in \mathcal{P}$, in the sense that no smaller dimensional family can capture the same model class. The degrees of freedom turns out to characterize the richness of the model class in our context. Compactness of Θ, in turn, implies existence of uniform bounds $0 < p_{\min}, p_{\max} < 1$ on the probabilities, i.e.

$$p_{\min} \leq p_\theta(x) \leq p_{\max} \qquad \forall \theta \in \Theta, \forall x \in \mathcal{X}. \tag{4.2}$$

From (4.1), the probability of a sequence $x^\ell = x_1 \cdots x_\ell$ drawn *i.i.d.* from a model $p_\theta \in \mathcal{P}$ in the exponential family takes the form Merhav and Weinberger (2004)

$$\begin{aligned} p_\theta(x^\ell) &= \prod_{i=1}^{\ell} p_\theta(x_i) \\ &= \prod_{i=1}^{\ell} 2^{\langle \theta, \boldsymbol{\tau}(x_i) \rangle - \psi(\theta)} \\ &= 2^{\ell \left[\langle \theta, \boldsymbol{\tau}(x^\ell) \rangle - \psi(\theta) \right]} \end{aligned} \tag{4.3}$$

where

$$\boldsymbol{\tau}(x^\ell) = \frac{\sum_{i=1}^{\ell} \boldsymbol{\tau}(x_i)}{\ell} \in \mathbb{R}^d \tag{4.4}$$

is a minimal sufficient statistic Merhav and Weinberger (2004). Note that $\boldsymbol{\tau}(x)$ and $\boldsymbol{\tau}(x^\ell)$ are differentiated based upon their arguments. We denote the (unknown) true model in force as p_{θ^*}. $\mathbb{P}_\theta$, $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ denote probability, expectation and variance with respect to p_θ, respectively. We denote the set of all finite length strings over $\mathcal{X}$ as $\mathcal{X}^*$. We denote the generic source string of unspecified length as $x^* \in \mathcal{X}^*$. Let $x^\ell x^{\ell'}$ be the concatenation of x^ℓ and $x^{\ell'}$. All logarithms are in base 2. Instead of introducing different indices for every new constant $C_1, C_2, ...$, the same letter C may be used to denote different constants whose precise values are irrelevant.

We consider V-F length codes, where we first parse the source sequence using a parsing dictionary $\mathcal{D}$ of a pre-specified size $|\mathcal{D}| = M$. Dictionary strings (segments) which we denote by $\{x_1^*, \cdots, x_M^*\}$ may have different lengths. Once a segment $x^* \in \mathcal{D}$ is identified as a parsed string, it is then encoded to its lexicographical index within $\mathcal{D}$ using $\log M$ bits. As it does not hurt our analysis, we ignore rounding $\log M$ to the closest integer.

We assume $\mathcal{D}$ is complete, i.e. any infinite length sequence over the alphabet has a prefix in $\mathcal{D}$. In addition, we assume $\mathcal{D}$ is proper, i.e. there are no two segments one being prefix of the other. Completeness along with properness of $\mathcal{D}$, implies that any long enough sequence has a unique prefix in the dictionary. Every complete and proper dictionary can be represented with a rooted complete $\mathcal{X}$-ary tree in which every internal node has $\mathcal{X}$ child nodes. Let us label each of the $|\mathcal{X}|$ edges branching out of a node with different letters from $\mathcal{X}$. Each node corresponds to the string of edge-labels from the root to the node. One can then correspond internal nodes of the tree to the prefixes of the segments, while leaf nodes correspond to the segments themselves.

Let $\mathcal{D}$ be the dictionary of a V-F length code ϕ. Let $X^* \in \mathcal{D}$ be the *random* first parsed segment of the source output X^∞, using the dictionary $\mathcal{D}$. Let $\ell(X^*)$ be the length of X^*. Denote $\ell^\phi(X^\infty) = \ell(X^*)$. We gauge the performance of V-F length code ϕ with a dictionary $\mathcal{D}$ of size M, through the ϵ-coding rate given by

$$\begin{aligned} R_M(\epsilon, \phi, p_{\theta^*}) &:= \inf \left\{ R : \ \mathbb{P}_{\theta^*} \left(\frac{\log M}{\ell^\phi(X^\infty)} \geq R \right) \leq \epsilon \right\} \\ &= \inf \left\{ R : \ \mathbb{P}_{\theta^*} \left(\frac{\log M}{\ell(X^*)} \geq R \right) \leq \epsilon \right\} \end{aligned} \tag{4.5}$$

Our goal is to analyze behavior of $R_M(\epsilon, \phi, p_{\theta^*})$ for large enough dictionary size M.

4.2 Type Complexity Algorithm

In this section, we propose the Type Complexity (TC) algorithm. For a sequence $x^\ell \in \mathcal{X}^*$, define its *quantized type complexity* $S(x^\ell)$ as

$$S(x^\ell) = \log |T_{x^\ell}| + \frac{d}{2} \log \ell \tag{4.6}$$

where T_{x^ℓ} is the quantized type class of x^ℓ defined in (3.5). Our designed dictionary $\mathcal{D}$, consists of sequences in the boundaries of transition from low quantized type complexity to high quantized type complexity. More precisely, for a positive constant γ to be specified in Section 4.5.1, $x^\ell = (x_1, x_2, \cdots, x_\ell)$ is a segment in the dictionary if and only if

$$S(x^\ell) \geq \gamma \text{ and } S(x^{\ell^{-1}}) < \gamma \tag{4.7}$$

where $x^{\ell^{-1}} = (x_1, x_2, \cdots, x_{\ell-1})$.

From construction, it is clear that $\mathcal{D}$ is proper, and furthermore monotonicity of $S(x^\ell)$ implies completeness of $\mathcal{D}$. Intuitively, sequences with high type complexity contain more *information*, implying that the type complexity (TC) code compresses more information into a fixed budget of output bits, which is the promise of the optimal V-F length code.

We note that there is a freedom in defining type classes in (4.6). Different characterization of type classes, leads to different performances. We show that the quantized type is the relevant characterization of type classes for optimal performance.

4.3 Main Result

Let $H(p_\theta) = \mathbb{E}_\theta\left(\log \frac{1}{p_\theta(X)}\right)$ and $\sigma^2(p_\theta) = \mathbb{V}_\theta\left(\log \frac{1}{p_\theta(X)}\right)$ be the entropy and the varentropy of p_θ, repectively. The following theorem exactly characterizes achievable ϵ-rates up to third-order term, as well as asserting that this rate is achievable by the TC code using quantized types.

Theorem 11. *For any stationary memoryless exponential family of distributions parameterized by Θ,*

$$\inf_{\phi} \sup_{\theta\in\Theta} \left[R_M(\epsilon, \phi, p_\theta) - H(p_\theta) - \sigma(p_\theta)\sqrt{\frac{H(p_\theta)}{\log M}}Q^{-1}(\epsilon) - H(p_\theta)\frac{d}{2}\frac{\log\log M}{\log M}\right] = o\left(\frac{\log\log M}{\log M}\right) \tag{4.8}$$

where the infimum is achieved by the TC algorithm using quantized types.

Example 3. *For the class of all binary memoryless sources $d = 1$, and the third-order term in (4.8) matches with the redundancy in Tjalkens and Willems (1992).*

4.4 Preliminary Results

Define

$$\hat{\theta}\left(\boldsymbol{\tau}\right) = \arg\max_{\theta\in\Theta}\left(\langle\theta, \boldsymbol{\tau}\rangle - \psi(\theta)\right). \tag{4.9}$$

Note that since the Hessian matrix of $\psi(\theta)$, $\boldsymbol{\nabla}^2\left(\psi(\theta)\right) = \mathrm{Cov}_\theta\left(\boldsymbol{\tau}(X)\right)$ is positive definite, the log-likelihood function is strictly concave and hence the maximum likelihood $\hat{\theta}(\boldsymbol{\tau})$ is unique.

The next lemma (Iri and Kosut, 2016a, Lemma 1) provides tight upper and lower bounds on the type complexity.

Lemma 13 (Type Complexity)**.** *For large enough ℓ, the quantized type complexity of x^ℓ is bounded as*

$$-\log p_{\hat{\theta}(x^\ell)}(x^\ell) + C_1 \leq S(x^\ell) \leq -\log p_{\hat{\theta}(x^\ell)}(x^\ell) + C_2 \tag{4.10}$$

where C_1, C_2 are constants independent of ℓ.

The type complexity bounds in the previous lemma, are springboards to the following bounds on the lengths of the dictionary segments.

Corollary 3 (Segment Length)**.** *For any long enough $x^\ell \in \mathcal{D}$, we have*

$$\ell < \frac{-\gamma + C_1}{\log p_{\max}} + 1 \tag{4.11}$$

Proof. For any long enough $x^\ell \in \mathcal{D}$, (4.7,4.10) yields

$$-\log p_{\hat{\theta}(x^{\ell^{-1}})}(x^{\ell^{-1}}) + C_1 < S(x^{\ell^{-1}}) < \gamma$$

Hence $p_{\hat{\theta}(x^{\ell^{-1}})}(x^{\ell^{-1}}) > 2^{-\gamma + C_1}$. Since for all $\theta \in \Theta$, $p_\theta(x^{\ell^{-1}}) \leq p_{\max}^{\ell-1}$, upper bound follows. □

Therefore, one can find positive constant $C_3 > 0$, such that for any $x^\ell \in \mathcal{D}$

$$\ell < C_3 \gamma. \tag{4.12}$$

The following lemma shows that, one single observation does not provide much information.

Lemma 14. *Let $x^{\ell+1} = (x_1, \cdots, x_\ell, x_{\ell+1}) = x^\ell x_{\ell+1}$. For a constant $C_4 > 0$*

$$-\log p_{\hat{\theta}(x^{\ell+1})}(x^{\ell+1}) - (-\log p_{\hat{\theta}(x^\ell)}(x^\ell)) \leq C_4 \tag{4.13}$$

Proof. We have

$$
\begin{aligned}
&-\log p_{\hat{\theta}(x^{\ell+1})}(x^{\ell+1}) - (-\log p_{\hat{\theta}(x^{\ell})}(x^{\ell})) = \\
&\qquad \max_{\theta}\left[(\ell+1)(\psi(\theta) - \langle\theta, \boldsymbol{\tau}(x^{\ell+1})\rangle)\right] - \max_{\theta}\left[\ell(\psi(\theta) - \langle\theta, \boldsymbol{\tau}(x^{\ell})\rangle)\right] && (4.14)\\
&\qquad \leq \max_{\theta}\left[(\ell+1)\psi(\theta) - (\ell+1)\langle\theta, \boldsymbol{\tau}(x^{\ell+1})\rangle - \ell\psi(\theta) + \ell\langle\theta, \boldsymbol{\tau}(x^{\ell})\rangle\right] && (4.15)\\
&\qquad \leq C && (4.16)
\end{aligned}
$$

where (4.14) is from the definition (4.9), (4.15) exploits the fact that for any two functions $g_1(\theta), g_2(\theta)$

$$\max_{\theta} g_1(\theta) - \max_{\theta} g_2(\theta) \leq \max_{\theta}\Big((g_1 - g_2)(\theta)\Big),$$

and finally (4.16) follows from $|\boldsymbol{\tau}(x^{\ell}) - \boldsymbol{\tau}(x^{\ell+1})| \leq \frac{C}{\ell}$ for some constant C. □

We appeal to the following normal approximation result from Kontoyiannis and Verdú (2014); Saito *et al.* (2014), in order to bound the percentiles of the type complexity in the achievability proof.

Lemma 15 (Asymptotic Normality of Information)**.** *Kontoyiannis and Verdú (2014); Saito* et al. *(2014) Fix a positive constant α. For a stationary memoryless source, there exists a finite positive constant A, such that for all $n \geq 1$ and z such that $|z| \leq \alpha$,*

$$\left|\mathbb{P}_{\theta^*}\left\{\frac{-\log p_{\theta^*}(X^n) - nH}{\sqrt{n}\sigma} > z\right\} - Q(z)\right| \leq \frac{A}{\sqrt{n}} \tag{4.17}$$

where $H := H(p_{\theta^*})$ and $\sigma^2 := \sigma^2(p_{\theta^*})$, are the entropy and the varentropy of the true model p_{θ^*}, respectively.

4.5 Achievability

4.5.1 Dictionary Size Enforcements

We would like to set γ in (4.7) as high as possible, to compress as much information as we can to a fixed budget of output bits. On the other hand, pre-specified dictionary size enforces upper bounds on γ. In this subsection, we set γ in (4.7), such that the dictionary size does not exceed M.

Let $N_{\ell+1}$ be the number of dictionary segments with length $\ell+1$. For any $x^{\ell+1} \in \mathcal{D}$, it must certainly hold that $S(x^{\ell+1-1}) < \gamma$. Motivated by (Merhav and Neuhoff, 1992, Eq. 3.12), we obtain the following bound

$$\begin{aligned} N_{\ell+1} &\leq |\mathcal{X}| \sum_{\substack{T_{x^\ell} \in \mathcal{T}_\ell: \\ S(x^\ell) \leq \gamma \\ \exists x_{\ell+1} \in \mathcal{X} \text{ such that } S(x^\ell x_{\ell+1}) > \gamma}} |T_{x^\ell}| \\ &\leq |\mathcal{X}| 2^{\gamma - \frac{d}{2}\log \ell} |\mathcal{A}| \end{aligned} \tag{4.18}$$

where (4.18) is from (4.6,4.7,4.10) and

$$\mathcal{A} = \{T \in \mathcal{T}_\ell : \exists x^\ell \in T \text{ with } S(x^\ell) \leq \gamma \text{ and } \exists x_{\ell+1} \text{ such that } S(x^\ell x_{\ell+1}) > \gamma\}. \tag{4.19}$$

We show in the Appendix J that $|\mathcal{A}| = \ell^{d-1}$. Hence

$$\begin{aligned} N_{\ell+1} &\leq |\mathcal{X}| 2^{\gamma - \frac{d}{2}\log \ell} \ell^{d-1} \\ &= |\mathcal{X}| 2^{\gamma} \ell^{\frac{d}{2}-1}. \end{aligned} \tag{4.20}$$

We then upper bound the dictionary size as follows

$$|\mathcal{D}| = \sum_{\ell=0}^{C_3\gamma} N_{\ell+1} \tag{4.21}$$

$$\leq |\mathcal{X}| 2^{\gamma} \sum_{\ell=0}^{C_3\gamma} \ell^{\frac{d}{2}-1} \tag{4.22}$$

$$\leq C 2^{\gamma} \gamma^{\frac{d}{2}} \tag{4.23}$$

where (4.21) is from (4.12), (4.22) follows from (4.20), and (4.23) is from upper bounding the summation with an integral where $C > 0$ is a generic constant whose precise value is irrelevant. Finally, in order to guarantee that the quantized type complexity dictionary (4.7), does not contain more than M segments, it suffices to set γ such that

$$\log C + \gamma + \frac{d}{2}\log\gamma \leq \log M. \tag{4.24}$$

One can show that, there exists a positive constant $C > 0$, such that the following choice of γ, satisfies (4.24)

$$\gamma = \log M - \frac{d}{2}\log\log M - C. \tag{4.25}$$

4.5.2 Coding Rate Analysis

In this subsection, we derive an upper bound for the ϵ-coding rate of the quantized type implementation of the TC algorithm. To this end, we upper bound the overflow probability as follows

$$\begin{aligned}
\mathbb{P}\left(\frac{\log M}{\ell(X^*)} > R\right) &= \mathbb{P}\left(\ell(X^*) < \frac{\log M}{R}\right) \\
&= \mathbb{P}\left(\exists \ell < \frac{\log M}{R} : S(x^\ell) \geq \gamma\right) && (4.26)\\
&\leq \mathbb{P}\left(\exists \ell < \frac{\log M}{R} : -\log p_{\hat{\theta}(x^\ell)}(x^\ell) \geq \gamma - C_2\right) && (4.27)\\
&\leq \mathbb{P}\left(\exists \ell < \frac{\log M}{R} : -\log p_{\theta^*}(x^\ell) \geq \gamma - C_2\right) && (4.28)\\
&= \mathbb{P}\left(-\log p_{\theta^*}(x^{\frac{\log M}{R}}) \geq \gamma - C_2\right) && (4.29)\\
&= \mathbb{P}\left(\frac{-\log p_{\theta^*}(x^{\frac{\log M}{R}}) - \frac{\log M}{R}H}{\sigma\sqrt{\frac{\log M}{R}}} \geq \frac{\gamma - C_2 - \frac{\log M}{R}H}{\sigma\sqrt{\frac{\log M}{R}}}\right) \\
&\leq Q\left(\frac{\gamma - C_2 - \frac{\log M}{R}H}{\sigma\sqrt{\frac{\log M}{R}}}\right) + \frac{A}{\sqrt{\frac{\log M}{R}}} && (4.30)
\end{aligned}$$

where (4.26) is from (4.7), (4.27) is from (4.10), (4.28) is from $p_{\theta^*}(x^\ell) \leq p_{\hat{\theta}(x^\ell)}(x^\ell)$, (4.29) holds since for x^ℓ a prefix of $x^{\ell+1}$, $-\log p_{\theta^*}(x^\ell) \leq -\log p_{\theta^*}(x^{\ell+1})$, and (4.30) is an application of Lemma 15. In the Appendix K, we show that for the rate R specified below, (4.30) and subsequently the overflow probability falls below ϵ

$$R = H + \sigma\sqrt{\frac{H}{\log M}}Q^{-1}(\epsilon) + H\frac{d}{2}\frac{\log\log M}{\log M} + \mathcal{O}(\frac{1}{\log M}). \tag{4.31}$$

Due to (4.5), $R_M(\epsilon, \mathcal{D}, p_{\theta^*}) \leq R$. This completes the achievability proof.

4.6 Converse

We first introduce some notations relevant to F-V length codes. Recall that any F-V length code $\phi^{\texttt{FV}}$ is a one-to-one mapping from a set of words of variable length $\mathcal{W}(\phi^{\texttt{FV}})$ to binary strings. For an infinite length sequence X^∞ emitted from the source, let

$$\ell(\phi^{\texttt{FV}}(X^\infty)) := \ell(\phi^{\texttt{FV}}(X_0^*)) \tag{4.32}$$

where $X_0^* \in \mathcal{W}(\phi^{\texttt{FV}})$ is a word of the code $\phi^{\texttt{FV}}$ that is a prefix of X^∞.

Let $\phi^{\texttt{VF}}$ be an arbitrary V-F length code with M codewords and length function $\ell^{\texttt{VF}}(\cdot)$. Let R be the ϵ-coding rate of $\phi^{\texttt{VF}}$. We show that

$$R \geq H + \sigma\sqrt{\frac{H}{\log M}}Q^{-1}(\epsilon) + H\frac{d}{2}\frac{\log\log M}{\log M} - C\frac{\log\log\log M}{\log M}. \tag{4.33}$$

It is shown in Merhav and Neuhoff (1992) that, for any V-F length code $\phi^{\texttt{VF}}$ with M codewords and length function $\ell^{\texttt{VF}}(\cdot)$, one can construct a F-V length prefix code $\phi^{\texttt{FV}}$ with $2^{\frac{\log M}{R}}$ codewords (i.e. fixed input length is $\frac{\log M}{R}$) and length function $\ell^{\texttt{FV}}(\cdot)$, such that the event $\left\{\ell^{\texttt{VF}}(X^\infty) < \frac{\log M}{R}\right\}$ for $\phi^{\texttt{VF}}$ is equivalent to the event $\{\ell^{\texttt{FV}}(X^\infty) > \log M\}$ for $\phi^{\texttt{FV}}$, where for simplicity $\ell^{\texttt{FV}}(X^\infty) := \ell(\phi^{\texttt{FV}}(X^\infty))$.

Therefore we have

$$\mathbb{P}\left(\frac{\log M}{\ell^{\mathtt{VF}}(X^\infty)} > R\right) = \mathbb{P}\left(\ell^{\mathtt{VF}}(X^\infty) < \frac{\log M}{R}\right) \tag{4.34}$$

$$= \mathbb{P}\left(\ell^{\mathtt{FV}}(X^\infty) > \log M\right) \tag{4.35}$$

$$= \mathbb{P}\left(\frac{\ell^{\mathtt{FV}}(X^\infty)}{\frac{\log M}{R}} > R\right). \tag{4.36}$$

Hence, $\mathbb{P}\left(\frac{\log M}{\ell^{\mathtt{VF}}(X^\infty)} > R\right) \leq \epsilon$ implies

$$\mathbb{P}\left(\frac{\ell^{\mathtt{FV}}(X^\infty)}{\frac{\log M}{R}} > R\right) \leq \epsilon. \tag{4.37}$$

Define the ϵ-coding rate $R(\phi^{\mathtt{FV}}, \epsilon, p)$ of the F-V length code $\phi^{\mathtt{FV}}$ as (Kosut and Sankar, 2014a, Eq. 8)

$$R(\phi^{\mathtt{FV}}, \epsilon, p) = \min\left\{R_0 : \mathbb{P}\left(\frac{\ell^{\mathtt{FV}}(X^\infty)}{\ell(X_0^*)} > R_0\right) \leq \epsilon\right\}.$$

Note that $\ell(X_0^*) = \frac{\log M}{R}$. Hence (4.37) implies

$$R \geq R(\phi^{\mathtt{FV}}, \epsilon, p). \tag{4.38}$$

Converse for fixed-to-variable prefix codes (Kosut and Sankar, 2014a, Theorem 16), in turn implies

$$R(\phi^{\mathtt{FV}}, \epsilon, p) \geq H + \frac{\sigma}{\sqrt{\frac{\log M}{R}}}Q^{-1}(\epsilon) + \frac{d}{2}\frac{\log\frac{\log M}{R}}{\frac{\log M}{R}} - \mathcal{O}\left(\frac{\log\log\frac{\log M}{R}}{\frac{\log M}{R}}\right). \tag{4.39}$$

Combining (4.38,4.39) yields

$$R \geq H + \frac{\sigma}{\sqrt{\frac{\log M}{R}}}Q^{-1}(\epsilon) + \frac{d}{2}\frac{\log\frac{\log M}{R}}{\frac{\log M}{R}} - C\left(\frac{\log\log\frac{\log M}{R}}{\frac{\log M}{R}}\right). \tag{4.40}$$

Through a similar iterative approach as in Appendix K, one can show that (4.40) leads to (4.33).

Chapter 5

CONCLUSION

We derived the fundamental limits of fixed-to-variable (F-V) length as well as variable-to-fixed (V-F) length universal source coding at short blocklengths. We first showed the optimality of the previously introduced Type Size code for F-V compression of Markov sources. We proceeded by considering the universal compression of the d-dimensional exponential family of distributions. We proposed the quantized Type Size code, where type classes are associated with cuboids in the grid partitioning the space of minimal sufficient statistics. We showed that the quantized Type Size code achieves the optimal third-order term $\left(\frac{d}{2}-1\right)\log n$ for compression of d-dimensional exponential family of distributions. Further, the naive point type class approach is considered, where two sequences are in the same type class if and only if they have the same probability under any distribution in the exponential family. In the point type class scenario, each point (rather than a cuboid) in the set of minimal sufficient statistics defines a type class. The third-order term of the point type class approach is shown to be exactly $(\frac{d'}{2}-1)\log n$, where d' is the dimension of the lattice vector representation of the sufficient statistic. Since d' is in general larger than d, our findings reveal that the model class dimension d — rather than the lattice dimension d' — is the relevant dimension for optimal performance. This is a more intuitive result, because it is much easier to understand the role of d as opposed to d'. Moreover, d is a more robust parameter compared to d'; changing the model parameters infinitesimally (i.e. from rational to irrational) can change d', but not d.

For a more general parametric family without any information on the minimal sufficient statistics, one may partition the parameter space into cuboids and define

two sequences to be in the same type class if and only if their maximum likelihood estimates belong to the same cuboid. One interesting future direction of this work is analyzing performance of such approach. As this work does not consider computational complexity of implementing the compression algorithms, an alternative future direction is to consider the blocklength-storage-complexity tradeoff. Finally, the lossy version of this research is also an interesting possible future direction.

Finally, we derived the fundamental limits for universal variable-to-fixed length coding of the d-dimensional exponential family of distributions in the non-vanishing error regime. We proposed a quantized type complexity algorithm that achieves the optimal third-order coding rate. Our algorithm along with prior V-F length codes follow a similar underlying theme; dictionary consists of sequences in the boundaries of transition from low to high complexity. One of the future directions of this work is to probe the diametric opposition to the complexity dogma revealed in this paper, if possible at all. Studying the behavior of the non-prefix codes is also considerably interesting for future work.

REFERENCES

Anderson, T., "Second-order moments of a stationary markov chain and some applications", (1989).

Balasko, Y., *The Equilibrium Manifold* (MIT Press Books, 2009).

Balasubramanian, V., "MDL, Bayesian Inference and the Geometry of the Space of Probability Distributions", in "Advances in Minimum Description Length: Theory and Applications", pp. 81–99 (Cambridge, MA: MIT Press, 2005).

Beirami, A. and F. Fekri, "Fundamental limits of universal lossless one-to-one compression of parametric sources", in "Information Theory Workshop (ITW), 2014 IEEE", pp. 212–216 (IEEE, 2014).

Borovkov, A. A. and A. A. Mogulskii, "Integro-local limit theorems including large deviations for sums of random vectors", Theory Probab. Appl. **43**, 1–12 (1999).

Cover, T. M. and J. A. Thomas, *Elements of Information Theory, 2nd Ed.* (Wiley, New York, 2006).

Cramér, H., *Mathematical methods of statistics*, vol. 9 (Princeton university press, 1999).

Csiszár, I. and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic Press, Orlando, FL, 1982).

Davisson, L., G. Longo and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic markov sources", Information Theory, IEEE Transactions on **27**, 4, 431–438 (1981).

Drmota, M., Y. A. Reznik and W. Szpankowski, "Tunstall code, khodak variations, and random walks", Information Theory, IEEE Transactions on **56**, 6, 2928–2937 (2010).

Harville, D. A., *Matrix algebra from a statistician's perspective* (Springer-Verlag, 1997).

Heydari, J. and A. Tajer, "Quickest search and learning over multiple sequences", in "Information Theory Proceedings (ISIT), 2016 IEEE International Symposium on", pp. 1371–1375 (2017).

Heydari, J., A. Tajer and H. V. Poor, "Quickest detection of markov networks", in "Information Theory Proceedings (ISIT), 2016 IEEE International Symposium on", pp. 1341–1345 (2016a).

Heydari, J., A. Tajer and H. V. Poor, "Quickest linear search over correlated sequences", IEEE Transactions on Information Theory **62**, 10, 5786–5808 (2016b).

Hug, D. and W. Weil, *A Course on Convex Geometry* (2010), URL `http://www.math.kit.edu/iag4/lehre/konvgeo2009w/media/cg.pdf`.

Iri, N. and O. Kosut, "Third-order coding rate for universal compression of Markov sources", in "Information Theory Proceedings (ISIT), 2015 IEEE International Symposium on", pp. 1996–2000 (2015).

Iri, N. and O. Kosut, "Fine asymptotics for universal one-to-one compression of parametric sources", arXiv preprint arXiv:1612.06448 (2016a).

Iri, N. and O. Kosut, "A new type size code for universal one-to-one compression of parametric sources", in "Information Theory Proceedings (ISIT), 2016 IEEE International Symposium on", pp. 1227–1231 (2016b).

Jordan, M. I., "An introduction to probabilistic graphical models", (2003).

Kontoyiannis, I. and S. Verdú, "Optimal lossless data compression: Non-asymptotics and asymptotics", Information Theory, IEEE Transactions on **60**, 2, 777–795 (2014).

Korshunov, D. A., "Limit theorems for general markov chains", Siberian Mathematical Journal **42**, 2, 301–316 (2001).

Kosut, O. and L. Sankar, "Universal fixed-to-variable source coding in the finite blocklength regime", in "Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on", pp. 649–653 (2013).

Kosut, O. and L. Sankar, "Asymptotics and non-asymptotics for universal fixed-to-variable source coding", arXiv preprint arXiv:1412.4444 (2014a).

Kosut, O. and L. Sankar, "New results on third-order coding rate for universal fixed-to-variable source coding", in "Information Theory Proceedings (ISIT), 2014 IEEE International Symposium on", pp. 2689–2693 (2014b).

Krichevsky, R. and V. Trofimov, "The performance of universal encoding", Information Theory, IEEE Transactions on **27**, 199–207 (1981).

Lapinskas, R., "On the rate of convergence in a multidimensional central limit theorem for inhomogeneous markov chains", vol. 14, pp. 46–61 (Springer, 1974).

Lawrence, J., "A new universal coding scheme for the binary memoryless source", Information Theory, IEEE Transactions on **23**, 466–472 (1977).

Merhav, N., "Asymptotically optimal decision rules for joint detection and source coding", Information Theory, IEEE Transactions on **60**, 11, 6787–6795 (2014).

Merhav, N. and M. Feder, "Universal prediction", Information Theory, IEEE Transactions on **44**, 6, 2124–2147 (1998).

Merhav, N. and D. L. Neuhoff, "Variable-to-fixed length codes provide better large deviations performance than fixed-to-variable length codes", IEEE Transactions on Information Theory **38**, 1, 135–140 (1992).

Merhav, N. and M. Weinberger, "On universal simulation of information sources using training data", Information Theory, IEEE Transactions on **50**, 1, 5–20 (2004).

MolavianJazi, E. and J. N. Laneman, "A finite-blocklength perspective on Gaussian multi-access channels", Arxiv.org:1309.2343 (2013).

R.A. Horn, C. J., *Matrix analysis* (Cambridge university press, 2012).

Ren, D., *Topics in integral geometry* (World scientific, 1994).

Rissanen, J., "Universal coding, information, prediction, and estimation", Information Theory, IEEE Transactions on **30**, 4, 629–636 (1984).

Rissanen, J., "Stochastic complexity and modeling", Annals of statistics **14**, 1080–1100 (1986).

Rissanen, J., "Fisher information and stochastic complexity", IEEE Transactions on Information Theory **42**, 1, 40–47 (1996).

Rissanen, J., "Strong optimality of the normalized ML models as universal codes and information in data", IEEE Transactions on Information Theory **47**, 5, 1712–1717 (2001).

Robbin, J. W. and D. A. Salamon., "Introduction to differential geometry", in "ETH, Lecture Notes, preliminary version", (2011), URL `https://people.math.ethz.ch/∼ salamon/PREPRINTS/diffgeo.pdf`.

Ryabko, B., "Prediction of random sequences and universal coding", Problems of information transmission **24**, 2, 87–96 (1988).

Ryabko, B., "Compression-based methods for nonparametric prediction and estimation of some characteristics of time series", Information Theory, IEEE Transactions on **55**, 9, 4309–4315 (2009).

Saito, S., N. Miya and T. Matsushima, "Evaluation of the minimum overflow threshold of Bayes codes for a Markov source", in "Information Theory and its Applications (ISITA), 2014 International Symposium on", pp. 211–215 (IEEE, 2014).

Saito, S., N. Miya and T. Matsushima, "Fundamental limit and pointwise asymptotics of the Bayes code for Markov sources", in "Information Theory Proceedings (ISIT), 2015 IEEE International Symposium on", pp. 1986–1990 (2015).

Seroussi, G. and M. J. Weinberger, "Optimal algorithms for universal random number generation from finite memory sources", IEEE Transactions on Information Theory **61**, 3, 1277–1297 (2015).

Shannon, C. E., ""a mathematical theory of communication.", Bell Syst. Tech. J. **27**, 379–423 (1948).

Stone, C., "On local and ratio limit theorems", in "Fifth Berkeley Symp. Math. Statist. Probab", pp. 217–224 (1967).

Szpankowski, W., "A one-to-one code and its anti-redundancy", IEEE Transactions on Information Theory **54**, 10, 4762–4766 (2008).

Szpankowski, W. and S. Verdu, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints", Information Theory, IEEE Transactions on **57**, 7, 4017–4025 (2011).

Tan, V. Y. F., "Asymptotic estimates in information theory with non-vanishing error probabilities", Found. Trends Commun. Inf. Theory **11**, 1-2, 1–183 (2014).

Tjalkens, T. and F. Willems, "A universal variable-to-fixed length source code based on lawrence's algorithm", Information Theory, IEEE Transactions on **38**, 247–253 (1992).

Tunstall, B. P., *Synthesis of noiseless compression codes* (Ph.D. dissert., Georgia Inst. of Technol., Atlanta, GA, 1967).

van Aardenne-Ehrenfest, T. and N. de Bruijn, "Circuits and trees in oriented linear graphs", in "Classic papers in combinatorics", pp. 149–163 (Springer, 1987).

Visweswariah, K., S. R. Kulkarni and S. Verdu, "Universal variable-to-fixed length source codes", Information Theory, IEEE Transactions on **47**, 1461–1472 (2001).

Wainwright, M. J. and M. I. Jordan, "Graphical models, exponential families, and variational inference", Found. Trends Machine Learning **1**, 1-2, 1–305 (2008).

Weinberger, M., N. Merhav and M. Feder, "Optimal sequential probability assignment for individual sequences", Information Theory, IEEE Transactions on **40**, 2, 384–396 (1994).

Wong, R., *Asymptotic Approximations of Integrals* (Academic Press, Inc., 1989).

Ziv, J., "An efficient universal prediction algorithm for unknown sources with limited training data", Information Theory, IEEE Transactions on **48**, 6, 1690–1693 (2002).

Ziv, J., "On finite memory universal data compression and classification of individual sequences", Information Theory, IEEE Transactions on **54**, 4, 1626–1636 (2008).

APPENDIX A

PROOF OF THE CLAIMS IN PROPOSITION 1

We have used Claims 1 and 3 over the course of the proof of Proposition 1, which we provide their proof here.

Claim 1. *Var*$\Big(\frac{1}{n}\big(\sum_{i=1}^{n} \boldsymbol{U}_i(s)\big)\Big) = \Theta(\frac{1}{n})$ *for all* $s = 0, \ldots, m-1$, *where* $\boldsymbol{U}_i(s)$ *is the* s*-th component of* $\boldsymbol{U}_i$.

Proof. It suffices to show that $\mathrm{Var}\Big(\sum_{i=1}^{n} \mathbf{U}_i(s)\Big) = \Theta(n)$. Note that $\mathrm{Var}\Big(\sum_{i=1}^{n} \mathrm{U}_i(s)\Big)$ for $s = 0, \ldots, m-1$, correspond to the diagonal entries of $\mathrm{Var}\Big(\sum_{i=1}^{n} \mathbf{U}_i\Big)$. We have

$$\mathrm{Var}(\sum_{i=1}^{n} \mathbf{U}_i) = \sum_{i=1}^{n} \mathrm{Var}(\mathbf{U}_i) + 2\sum_{i<j} \mathrm{Cov}(\mathbf{U}_i, \mathbf{U}_j). \tag{A.1}$$

The First term in A.1 is $\Theta(n)$. Since $\mathbf{U}_i$'s are from an irreducible and aperiodic Markov chain, using the same approach as in Lemma (1) one can show that the second term in A.1 vanishes exponentially. □

Let us denote

$$\hat{\mathrm{U}}_i = \sum_{r=1}^{m} j_r \mathbf{U}_i(r)$$

where $\mathbf{U}_i(r)$ is the r-th component of $\mathbf{U}_i$. Throughout this appendix, let $\mathbf{U}_i$ satisfy the assumptions of Proposition 1. Also let

$$\alpha_s = \mathrm{Cov}\Big(\hat{\mathrm{U}}_i, \hat{\mathrm{U}}_{i+s}\Big).$$

Claim 2. *There exists a constant* $\rho < 1$, *such that* $\alpha_s < \rho^s$.

Proof. Let $\mathbf{G}$ be an m-indicator vector, with ith component being $\mathrm{G}_i = j_i$. Observe that $\alpha_s = \mathrm{Cov}\left(\mathbf{G}^T\mathbf{U}_i, \mathbf{G}^T\mathbf{U}_{i+s}\right) = \mathbf{G}^T \mathrm{cov}(\mathbf{U}_i, \mathbf{U}_{i+s})\mathbf{G}$. Using the same approach as in Lemma 1, since $\mathbf{U}_i$'s are irreducible and aperiodic, one can show that $\mathrm{Cov}\left(\mathbf{U}_i, \mathbf{U}_{i+s}\right)$ decreases exponentially and hence α_s decreases exponentially as well. □

Claim 3.

$$\frac{\sigma_\infty - \sigma_n}{\sigma_n} = \mathcal{O}(\frac{1}{\sqrt{n}}) \tag{A.2}$$

Proof. We have

$$\begin{aligned}
\sigma_n^2 &= \mathbf{j}^T \mathbf{V}_n \mathbf{j} = \sum_{r,r'} j_r V_n(r,r') j_{r'} \\
&= \sum_{r,r' \in [m]} j_r \mathrm{Cov}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}_i(r), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}_i(r') \Big) j_{r'} \\
&= \sum_{r,r' \in [m]} j_r \mathbb{E}\left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}_i(r) \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{U}_i(r') \right) \right] j_{r'} \\
&= \mathbb{E}\left[\left(\frac{1}{\sqrt{n}} \sum_{r=1}^{m} \sum_{i=1}^{n} j_r \mathbf{U}_i(r) \right)^2 \right].
\end{aligned}$$

Hence, we have

$$\begin{aligned}
n\sigma_n^2 &= \mathbb{E}\left[\left(\sum_{r=1}^{m} \sum_{i=1}^{n} j_r \mathbf{U}_i(r) \right)^2 \right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} \hat{\mathrm{U}}_i \right)^2 \right] \\
&= \sum_{i,j} \alpha_{j-i} \\
&= \sum_{i=1}^{n} \alpha_0 + \sum_{i<j} 2\alpha_{j-i} \\
&= n\alpha_0 + \sum_{i=1}^{n-1} 2(n-i)\alpha_i.
\end{aligned}$$

Therefore,

$$\sigma_n^2 = \alpha_0 + \sum_{i=1}^{n-1} 2(\frac{n-i}{n})\alpha_i. \tag{A.3}$$

Taking the limits of (A.3), we obtain

$$\sigma_\infty^2 = \alpha_0 + \sum_{i=1}^{\infty} 2\alpha_i. \tag{A.4}$$

Subtracting (A.3) from (A.4) yields

$$\begin{aligned}
\left|\sigma_\infty^2 - \sigma_n^2\right| &= \left|\sum_{i=1}^{n-1} 2(\frac{i}{n})\alpha_i + \sum_{i=n}^{\infty} 2\alpha_i\right| \\
&\leq \sum_{i=1}^{\sqrt{n}} \frac{2}{\sqrt{n}}|\alpha_i| + \sum_{i=\sqrt{n}+1}^{n-1} 2|\alpha_i| + \sum_{i=n}^{\infty} 2|\alpha_i| \\
&\leq \frac{2}{\sqrt{n}} \sum_{i=1}^{\infty} |\alpha_i| + \sum_{i=\sqrt{n}+1}^{\infty} 2|\alpha_i| \\
&\leq \frac{2\rho}{(1-\rho)\sqrt{n}} + \frac{2\rho^{\sqrt{n}+1}}{1-\rho} \\
&= \mathcal{O}(\frac{1}{\sqrt{n}}).
\end{aligned}$$

Claim follows by recalling from Claim 1 that $\sigma_n = \Theta(1)$. □

APPENDIX B

PROOF OF THE CLAIM IN LEMMA 3

Claim 4.

$$\sigma_\infty^2 = \sigma^2. \tag{B.1}$$

Proof. First note that $\bar{\mathbf{U}}((x,y)) = \tilde{q}(x,y) - \tilde{p}(x,y)$. Since $\mathbf{U}_i$'s are indexed with tuples $(x,y) \in \mathcal{X}^2$, so does $\mathbf{j}$ in the statement of the proposition. With the choice of $\mathbf{U}$ and $f(\cdot)$ in the proof of Lemma 3, we have

$$\begin{aligned} j_{(x,y)} &= \frac{\partial f(\bar{\mathbf{u}})}{\partial \bar{\mathbf{u}}(x,y)}\Big|_{\bar{\mathbf{u}}=0} = \frac{\partial f(\bar{\mathbf{u}})}{\partial \tilde{q}(x,y)}\Big|_{\tilde{q}=\tilde{p}} \\ &= -\log \tilde{p}(x,y) - 1 + \log\left(\sum_x \tilde{p}(x,y)\right) + \frac{\tilde{q}(x,y)}{\tilde{q}(x,y)} \\ &= -\log \tilde{p}(x,y) + \log\left(\sum_x \tilde{p}(x,y)\right) \\ &= -\log \tilde{p}(x,y) + \log p(y) = -\log \frac{\tilde{p}(x,y)}{p(y)} = -\log \hat{p}(x|y). \end{aligned} \tag{B.2}$$

Observe that

$$\sigma_n^2 = \mathbf{j}^T \mathbf{V}_n \mathbf{j} = \sum_{(x,y),(x',y') \in \mathcal{X}^2} j_{(x,y)} \mathbf{V}_n\Big((x,y),(x',y')\Big) j_{(x',y')} \tag{B.3}$$

where

$$\begin{aligned} \mathbf{V}_n\Big((x,y),(x',y')\Big) &= \mathbb{E}\Big[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{U}_i^T(x,y)\Big)\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{U}_i^T(x',y')\Big)\Big] \\ &= n\mathbb{E}\Big[\Big(\tilde{q}(x,y) - \tilde{p}(x,y)\Big)\Big(\tilde{q}(x',y') - \tilde{p}(x',y')\Big)\Big]. \end{aligned} \tag{B.4}$$

Plugging (B.2) and (B.4) in (B.3) yields

$$\begin{aligned} \sigma_n^2 = \sum_{(x,y),(x',y') \in \mathcal{X}^2} &(-\log \hat{p}(x|y)) \cdot n\mathbb{E}\left[(\tilde{q}(x,y) - \tilde{p}(x,y))\left(\tilde{q}(x',y') - \tilde{p}(x',y')\right)\right] \\ &\cdot (-\log \hat{p}(x'|y')) \\ = n\mathbb{E}\Big[\Big(\sum_{(x,y)\in\mathcal{X}^2} &\tilde{q}(x,y)(-\log \hat{p}(x|y)) + \sum_{x,y} \tilde{p}(x,y)\log \hat{p}(x|y)\Big)^2\Big] \\ = n\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^n &-\log \hat{p}(x_{i+1}|x_i) - H(X_2|X_1)\Big)^2\Big] \\ = \mathbb{E}\Big[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n &\big(-\log \hat{p}(x_{i+1}|x_i) - H(X_2|X_1)\big)\Big)^2\Big]. \end{aligned}$$

Hence, $\lim_{n\to\infty} \sigma_n^2 = \sigma^2$. □

APPENDIX C

PROOF OF LEMMA 15: ASYMPTOTIC NORMALITY OF INFORMATION

Define

$$e(\boldsymbol{\tau}) = -\max_{\theta} \Big(\langle \theta, \boldsymbol{\tau} \rangle - \psi(\theta) + \langle \theta, \nabla\psi(\theta^*) \rangle \Big). \tag{C.1}$$

Fuethermore, denote $\boldsymbol{U}_i(x^n) = \boldsymbol{\tau}(x_i) - \boldsymbol{\mu}$ for $i = 1, \cdots, n$, where $\boldsymbol{\mu} = \mathbb{E}_{\theta^*}[\boldsymbol{\tau}(X)]$. Therefore, $\boldsymbol{U}_i(X^n)$'s are zero-mean with finite covariance. First, observe that

$$\frac{1}{n} \log p_{\hat{\theta}}(x^n) = -\max_{\theta} \langle \theta, \boldsymbol{\tau}(x^n) - \boldsymbol{\mu} \rangle - \psi(\theta) + \langle \theta, \boldsymbol{\mu} \rangle \tag{C.2}$$

$$= e\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{U}_i(x^n) \right) \tag{C.3}$$

where (C.2) is from (3.9), and since $\boldsymbol{\mu} = \nabla\psi(\theta^*)$ Jordan (2003), (C.3) follows from (C.1,3.3). We then show that $e(\mathbf{0}) = H$. Equating the derivative with respect to θ of the expression inside the parenthesis with zero, we find that θ^* is the maximizing parameter in (C.1). Therefore

$$e(\mathbf{0}) = -\Big(-\psi(\theta^*) + \langle \theta^*, \nabla\psi(\theta^*) \rangle \Big) \tag{C.4}$$

$$= -\Big(-\psi(\theta^*) + \langle \theta^*, \mathbb{E}_{\theta^*}(\boldsymbol{\tau}(X)) \rangle \Big) \tag{C.5}$$

$$= -\mathbb{E}_{\theta^*}\Big(-\psi(\theta^*) + \langle \theta^*, (\boldsymbol{\tau}(X)) \rangle \Big)$$

$$= -\mathbb{E}_{\theta^*}\Big(\log p_{\theta^*}(X) \Big) \tag{C.6}$$

$$= H$$

where (C.4) is from (C.1), (C.5) is an exponential family property Jordan (2003), and (C.6) is from (3.1). Application of the Proposition 1 in Iri and Kosut (2015) completes the proof.

APPENDIX D

PROOF OF LEMMA 6: MAXIMUM LIKELIHOOD APPROXIMATION

We show that $\log p_{\hat{\theta}(x^n)}(x^n)$ is constant away from $\log p_{\hat{\theta}_c(x^n)}(x^n)$. Recall that

$$\log p_{\hat{\theta}(x^n)}(x^n) = n \max_{\theta} \Big[\langle \theta, \boldsymbol{\tau}(x^n) \rangle - \psi(\theta) \Big].$$

For ease of notation, when it is clear from the context, we denote $\boldsymbol{\tau}_c(x^n)$ as $\boldsymbol{\tau}_c$, and similarly we remove the argument in $\hat{\theta}_c(x^n)$ and simply denote it as $\hat{\theta}_c$. Since $\boldsymbol{\tau}(x^n)$ is in a cuboid of side length $\frac{s}{n}$ with center $\boldsymbol{\tau}_c$, we have $\|\boldsymbol{\tau}(x^n) - \boldsymbol{\tau}_c\| \leq \frac{s\sqrt{d}}{2n}$. We hence have

$$\begin{aligned}
\Big| \Big\langle \hat{\theta}_c, \boldsymbol{\tau}(x^n) \Big\rangle - \Big\langle \hat{\theta}_c, \boldsymbol{\tau}_c \Big\rangle \Big| &= \Big| \Big\langle \hat{\theta}_c, \boldsymbol{\tau}(x^n) - \boldsymbol{\tau}_c \Big\rangle \Big| \\
&\leq \|\hat{\theta}_c\| \|\boldsymbol{\tau}(x^n) - \boldsymbol{\tau}_c\| \\
&\leq \wp \frac{s\sqrt{d}}{2n} = \frac{\kappa s}{n} \qquad \text{(D.1)}
\end{aligned}$$

where (D.1) exploits the fact that $\|\theta\| \leq \wp$, for all $\theta \in \Theta$, including $\hat{\theta}_c$. Therefore

$$\begin{aligned}
\log p_{\hat{\theta}_c(x^n)}(x^n) &= n \left[\Big\langle \hat{\theta}_c, \boldsymbol{\tau}(x^n) \Big\rangle - \psi\left(\hat{\theta}_c\right) \right] \\
&\geq n \left[\Big\langle \hat{\theta}_c, \boldsymbol{\tau}_c(x^n) \Big\rangle - \frac{\kappa s}{n} - \psi\left(\hat{\theta}_c\right) \right] \qquad \text{(D.2)} \\
&= n \max_{\theta} \Big[\langle \theta, \boldsymbol{\tau}_c(x^n) \rangle - \frac{\kappa s}{n} - \psi(\theta) \Big] \qquad \text{(D.3)}
\end{aligned}$$

where (D.2) follows from (D.1) and (D.3) is from the definition of $\hat{\theta}_c$. Using the fact that for any two functions $g_1(\theta), g_2(\theta)$

$$\max_{\theta} g_1(\theta) - \max_{\theta} g_2(\theta) \leq \max_{\theta} \Big((g_1 - g_2)(\theta) \Big) \qquad \text{(D.4)}$$

we obtain

$$\begin{aligned}
\log p_{\hat{\theta}(x^n)}(x^n) - \log p_{\hat{\theta}_c}(x^n) &\leq n \max_{\theta} \Big[\langle \theta, \boldsymbol{\tau}(x^n) \rangle - \psi(\theta) \Big] \\
&\qquad - n \max_{\theta} \Big[\langle \theta, \boldsymbol{\tau}_c(x^n) \rangle - \frac{\kappa s}{n} - \psi(\theta) \Big] \qquad \text{(D.5)} \\
&\leq n \max_{\theta} \Big[\langle \theta, \boldsymbol{\tau}(x^n) - \boldsymbol{\tau}_c \rangle + \frac{\kappa s}{n} \Big] \qquad \text{(D.6)}
\end{aligned}$$

where (D.5) exploits (D.3), and (D.6) is from (D.4). Similar to (D.1), one can show that $\langle \theta, \boldsymbol{\tau}(x^n) - \boldsymbol{\tau}_c \rangle \leq \frac{\kappa s}{n}$. Lemma then follows.

APPENDIX E

PROOF OF LEMMA 7: LIPSCHITZNESS OF $F(\cdot)$

Let

$$l(\boldsymbol{\tau}) = \max_{\theta} \left(\langle \theta, \boldsymbol{\tau} \rangle - \psi(\theta) \right). \tag{E.1}$$

Noticing that $\|\nabla f(\boldsymbol{\tau})\| = \|\nabla l(\boldsymbol{\tau})\|$, in order to show the Lipschitzness of $f(\boldsymbol{\tau})$ in (3.15), it suffices to show that $l(\boldsymbol{\tau})$ is a Lipschitz function of $\boldsymbol{\tau}$. We first show that $\|\nabla l(\boldsymbol{\tau})\| = \|\hat{\theta}(\boldsymbol{\tau})\|$. Due to (3.9)

$$l(\boldsymbol{\tau}) = \langle \hat{\theta}(\boldsymbol{\tau}), \boldsymbol{\tau} \rangle - \psi\left(\hat{\theta}(\boldsymbol{\tau})\right).$$

Hence, taking gradient with respect to $\boldsymbol{\tau}$

$$\begin{aligned} \nabla l(\boldsymbol{\tau}) &= \left(\left(\nabla \hat{\theta}(\boldsymbol{\tau}) \right) \boldsymbol{\tau} + \hat{\theta}(\boldsymbol{\tau}) \right) - \nabla \hat{\theta}(\boldsymbol{\tau}) \nabla_{\hat{\theta}} \psi\left(\hat{\theta}(\boldsymbol{\tau})\right) \\ &= \left(\left(\nabla \hat{\theta}(\boldsymbol{\tau}) \right) \boldsymbol{\tau} + \hat{\theta}(\boldsymbol{\tau}) \right) - \nabla \hat{\theta}(\boldsymbol{\tau}) \mathbb{E}_{\hat{\theta}(\boldsymbol{\tau})}(\boldsymbol{\tau}(X)) \qquad \text{(E.2)} \\ &= \hat{\theta}(\boldsymbol{\tau}) \qquad \text{(E.3)} \end{aligned}$$

where $(E.2)$ follows from $\nabla_{\hat{\theta}} \psi\left(\hat{\theta}(\boldsymbol{\tau})\right) = \mathbb{E}_{\hat{\theta}(\boldsymbol{\tau})}\left(\boldsymbol{\tau}(X)\right)$ Jordan (2003), and $(E.3)$ follows from $\mathbb{E}_{\hat{\theta}(\boldsymbol{\tau})}(\boldsymbol{\tau}(X)) = \boldsymbol{\tau}$ (see the proof of Lemma 4). Lemma follows by recalling that $\|\hat{\theta}(\boldsymbol{\tau})\| \leq \wp$.

APPENDIX F

PROOF OF LEMMA 8: LIPSCHITZNESS OF $\rho(\cdot)$

Let $\mathcal{K} = \{\boldsymbol{\tau} \in \mathcal{T} : f(\boldsymbol{\tau}) \leq \lambda\}$ and $\mathcal{K}^c = \mathcal{T} \backslash \mathcal{K}$. We first show that $\mathcal{K}^c$ is a convex body. A sub-level set of $f(\cdot)$ is a sub-level set of $-l(\cdot)$ defined as in (E.1). Therefore, it is enough to show that sub-level sets of $l(\cdot)$ (i.e. $\mathcal{K}^c$) are convex. Maximum of linear functions of $\boldsymbol{\tau}$ is a convex function, therefore $l(\cdot)$ defined in (E.1) is a convex function of $\boldsymbol{\tau}$. Since the sub-level sets of a convex function are convex, $\mathcal{K}^c$ is a convex body.

In order to show that $\rho(\lambda)$ $(= \mathrm{Vol}\,(\mathcal{K}))$ is Lipschitz, we provide an upper bound for the absolute value of its derivative $|\frac{d}{d\lambda}\rho(\lambda)|$. Let us denote the surface area of a convex body $\mathcal{K}^c$ as (Hug and Weil, 2010, Section 3.3)

$$S(\mathcal{K}^c) = \lim_{\epsilon \to 0} \frac{V^{(d)}\Big(\mathcal{K}^c + B(\epsilon)\Big) - V^{(d)}(\mathcal{K}^c)}{\epsilon} \tag{F.1}$$

where $V^{(d)}(\cdot)$ is the d-dimensional volume, $B(\epsilon)$ is the d-dimensional unit ball and addition in $\mathcal{K}^c + B(\epsilon)$ is the Minkowski's sum Hug and Weil (2010). Let us denote $\mathcal{K}^c_\epsilon = \{\boldsymbol{\tau} \in \mathcal{T} : f(\boldsymbol{\tau}) > \lambda - \epsilon\}$. We have

$$\begin{aligned}
\frac{d}{d\lambda}\rho(\lambda) &= \lim_{\epsilon \to 0} \frac{\rho(\lambda) - \rho(\lambda - \epsilon)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{(\mathrm{Vol}(\mathcal{T}) - \rho(\lambda - \epsilon)) - (\mathrm{Vol}(\mathcal{T}) - \rho(\lambda))}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\mathrm{Vol}(\mathcal{K}^c_\epsilon) - \mathrm{Vol}(\mathcal{K}^c)}{\epsilon}.
\end{aligned} \tag{F.2}$$

Let us assume $\epsilon \to 0^+$; the case where $\epsilon \to 0^-$ is handled similarly. Let $\boldsymbol{\tau}_1 \in \mathcal{K}^c_\epsilon$. From the Taylor series expansion of $f(\boldsymbol{\tau}_2)$ in the vicinity of $\boldsymbol{\tau}_1$ with distance at most $\|\boldsymbol{\tau}_2 - \boldsymbol{\tau}_1\| \leq \sqrt{\epsilon}$, we obtain

$$f(\boldsymbol{\tau}_2) = f(\boldsymbol{\tau}_1) + \langle \nabla f(\boldsymbol{\tau}_1), \boldsymbol{\tau}_2 - \boldsymbol{\tau}_1 \rangle + \Delta \tag{F.3}$$

where $|\Delta| \leq C_f \|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\|^2$, for a constant C_f independent of n. Let

$$\boldsymbol{\tau}_2 = \boldsymbol{\tau}_1 + \epsilon \frac{(1 + C_f)\nabla f(\boldsymbol{\tau}_1)}{\|\nabla f(\boldsymbol{\tau}_1)\|^2}. \tag{F.4}$$

With this choice of $\boldsymbol{\tau}_2$, we obtain

$$\begin{aligned}
f(\boldsymbol{\tau}_2) &= f(\boldsymbol{\tau}_1) + \epsilon(1 + C_f) + \Delta \\
&\geq f(\boldsymbol{\tau}_1) + \epsilon + \epsilon C_f - C_f \|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\|^2 \\
&\geq f(\boldsymbol{\tau}_1) + \epsilon && \text{(F.5)} \\
&> \lambda - \epsilon + \epsilon && \text{(F.6)} \\
&= \lambda
\end{aligned}$$

where (F.5) follows form $\|\boldsymbol{\tau}_2 - \boldsymbol{\tau}_1\| \leq \sqrt{\epsilon}$, and (F.6) is a consequence of $\boldsymbol{\tau}_1 \in \mathcal{K}^c_\epsilon$. Hence $\boldsymbol{\tau}_2 \in \mathcal{K}^c$. Since $\boldsymbol{\tau}_1 \in \mathcal{K}^c_\epsilon$ was arbitrary, we have $\mathcal{K}^c_\epsilon \subset \mathcal{K}^c + B\left(\frac{\epsilon(1+C_f)}{\|\nabla f(\boldsymbol{\tau}_1)\|}\right)$.

Therefore, one can upper bound (F.2) in terms of the surface area (F.1) as follows:

$$\left|\frac{d}{d\lambda}\rho(\lambda)\right| \leq \frac{(1+C_f)S(\mathcal{K}^c)}{\|\nabla f(\boldsymbol{\tau}_1)\|} \quad \text{for all } \boldsymbol{\tau}_1 \in \mathcal{K}_\epsilon. \tag{F.7}$$

Since $\mathcal{K}^c, \mathcal{T}$ are convex bodies and $\mathcal{K}^c \subset \mathcal{T}$, consequently $S(\mathcal{K}^c) \leq S(\mathcal{T})$ (Hug and Weil, 2010, Theorem 3.2.2). Since $\mathcal{X}$ is finite, therefore $\mathcal{T}$ is a bounded set, which yields $S(\mathcal{K}^c) \leq S(\mathcal{T}) < \infty$.

From the proof of Lemma 7 in Appendix E, we have $\|\nabla f(\boldsymbol{\tau}_1)\| = \|\hat{\theta}(\boldsymbol{\tau}_1)\|$. That translates (F.7) into

$$\left|\frac{d}{d\lambda}\rho(\lambda)\right| \leq \frac{(1+C_f)S(\mathcal{K}^c)}{\|\hat{\theta}(\boldsymbol{\tau}_1)\|} \quad \text{for all } \boldsymbol{\tau}_1 \in \mathcal{K}_\epsilon. \tag{F.8}$$

We finally show that $\|\hat{\theta}(\boldsymbol{\tau}_1)\|$ is bounded away from zero. Let $\boldsymbol{\tau}_u$ be such that $\hat{\theta}(\boldsymbol{\tau}_u) = (0, \cdots, 0)$ (subscript u stands for the uniform distribution.). Since $\omega = \frac{\log|\mathcal{X}|-H}{5} > 0$ and $f(\boldsymbol{\tau}) = -\frac{1}{n}\log p_{\hat{\theta}(\boldsymbol{\tau})}(x^n) - \Theta\left(\frac{\log n}{n}\right)$, we have that

$$f(\boldsymbol{\tau}_u) \geq \log|\mathcal{X}| - \omega, \text{ for large enough } n. \tag{F.9}$$

From boundedness of $\mathcal{T}$, we have

$$T_{\max} := \max\{\|\boldsymbol{\tau}\| : \boldsymbol{\tau} \in \mathcal{T}\} < \infty.$$

Therefore $\left\|\nabla\psi\left(\hat{\theta}(\boldsymbol{\tau})\right)\right\| = \|\mathbb{E}_{\hat{\theta}}(\boldsymbol{\tau}(X))\| \leq T_{\max}$ is bounded. Hence $\psi\left(\hat{\theta}(\boldsymbol{\tau})\right)$ is a Lipschitz function of $\hat{\theta}(\boldsymbol{\tau})$ with Lipschitz constant $T_{\max}$. Hence if $\left\|\hat{\theta}(\boldsymbol{\tau}) - \hat{\theta}(\boldsymbol{\tau}_u)\right\| \leq \frac{\omega}{T_{\max}}$, then $\left|\psi(\hat{\theta}(\boldsymbol{\tau})) - \psi(\hat{\theta}(\boldsymbol{\tau}_u))\right| \leq \omega$ and furthermore by the Cauchy-Schwarz inequality $\left|\left\langle\hat{\theta}(\boldsymbol{\tau}) - \hat{\theta}(\boldsymbol{\tau}_u), \boldsymbol{\tau}\right\rangle\right| \leq \omega$. Therefore, if $\left\|\hat{\theta}(\boldsymbol{\tau}) - \hat{\theta}(\boldsymbol{\tau}_u)\right\| \leq \frac{\omega}{T_{\max}}$, then

$$\begin{aligned}|f(\boldsymbol{\tau}) - f(\boldsymbol{\tau}_u)| &\leq \left|\left\langle\hat{\theta}(\boldsymbol{\tau}), \boldsymbol{\tau}\right\rangle - \left\langle\hat{\theta}(\boldsymbol{\tau}_u), \boldsymbol{\tau}_u\right\rangle\right| + \left|\psi\left(\hat{\theta}(\boldsymbol{\tau})\right) - \psi\left(\hat{\theta}(\boldsymbol{\tau}_u)\right)\right| \\ &\leq 2\omega \end{aligned} \tag{F.10}$$

where (F.10) follows from $\hat{\theta}(\boldsymbol{\tau}_u) = (0, \cdots, 0)$, $|\langle\hat{\theta}(\boldsymbol{\tau}) - \hat{\theta}(\boldsymbol{\tau}_u), \boldsymbol{\tau}\rangle| \leq \omega$. Finally, for large enough n and for all $\boldsymbol{\tau}_1 \in \mathcal{K}_\epsilon$, it holds that

$$\begin{aligned} f(\boldsymbol{\tau}_1) &\leq \lambda + \epsilon \\ &< (H + \omega) + \omega \qquad \text{(F.11)} \\ &= \log|\mathcal{X}| - 3\omega. \qquad \text{(F.12)} \end{aligned}$$

where (F.11) follows from $\lambda < H + \omega$ and the fact that since $\epsilon \to 0$, $\epsilon < \omega$ and (F.12) is from the definition of ω. From (F.9) and (F.12), we have $|f(\boldsymbol{\tau}_1) - f(\boldsymbol{\tau}_u)| > 2\omega$ for all $\boldsymbol{\tau}_1 \in \mathcal{K}_\epsilon$. Hence by (F.10), we must certainly have $\left\|\hat{\theta}(\boldsymbol{\tau}_1) - \hat{\theta}\left(\boldsymbol{\tau}_u\right)\right\| > \frac{\omega}{T_{\max}}$. On the other hand $\hat{\theta}(\boldsymbol{\tau}_u) = (0, \cdots, 0)$, which entails that $\left\|\hat{\theta}(\boldsymbol{\tau}_1)\right\| > \frac{\omega}{T_{\max}}$. This yields a positive lower bound, independent of n, for the denominator in (F.8).

APPENDIX G

PROOF OF LEMMA 10: POINT TYPE CLASS SIZE

The one-to-one mapping between $\boldsymbol{\tau}(x)$ and $\boldsymbol{L}(x)$, subsequently defines a one-to-one mapping between $\boldsymbol{\tau}(x^n)$ and $\boldsymbol{L}(x^n)$, which consequently defines a one-to-one correspondence between the point type class T_{x^n} and $\boldsymbol{L}(x^n)$. Therefore, for any parameter value $\theta \in \Theta$, it holds that Merhav and Weinberger (2004)

$$|T_{x^n}| = \frac{\mathbb{P}_\theta\{\boldsymbol{L}(X^n) = \boldsymbol{L}(x^n)\}}{p_\theta(x^n)}. \tag{G.1}$$

Since $\boldsymbol{L}(x^n)$ can be written as a sum of integer (lattice) random vectors $\boldsymbol{L}(x_i)$ (Eq. (3.48)), exploiting the local limit theorem of Borovkov and Mogulskii (1999) to bound the numerator in (G.1), yields Merhav and Weinberger (2004)

$$\log|T_{x^n}| = -\log p_{\hat{\theta}(x^n)}(x^n) - \frac{d'}{2}\log 2\pi n - \frac{1}{2}\log\det M\left[\hat{\theta}(x^n)\right] + o(1) \tag{G.2}$$

where $\hat{\theta}(x^n)$ is the maximum likelihood estimate of θ for x^n and $M[\theta]$ denotes the covariance matrix of the random vector $\boldsymbol{L}(X)$ where X is drawn from p_θ.

We show that absolute value of the third term in (G.2), $\left|\frac{1}{2}\log\det M\left[\hat{\theta}(x^n)\right]\right|$, is upper bounded by a constant $C_M > 0$ independent of n. Constant upper bound $C_u > 0$, for $\det M\left[\hat{\theta}(x^n)\right]$ follows from Hadamard's inequality (R.A. Horn, 2012, corollary 7.8.3). For the lower bound, since $\det M[\theta]$ is a continuous function of θ over a compact domain Θ, it attains a minimum at a point in the parameter space, say $\ddot{\theta} \in \Theta$. Let $\ddot{\boldsymbol{P}}$ be a diagonal $(|\mathcal{X}|-1)\times(|\mathcal{X}|-1)$ matrix with diagonal entries $\ddot{\boldsymbol{P}}_{ii} = \mathbb{P}_{\ddot{\theta}}(X = i+1)$ for $i \in \mathcal{X}$, and $\ddot{\boldsymbol{p}}$ be a column vector with $\ddot{\boldsymbol{p}}_i = \mathbb{P}_{\ddot{\theta}}(X = i+1)$ for $i \in \mathcal{X}$. We have

$$\begin{aligned}
M(\ddot{\theta}) &= \mathbb{E}_{\ddot{\theta}}\left([\boldsymbol{L}(X)]\,[\boldsymbol{L}(X)]^T\right) - \mathbb{E}_{\ddot{\theta}}([\boldsymbol{L}(X)])\left(\mathbb{E}_{\ddot{\theta}}([\boldsymbol{L}(X)])\right)^T \\
&= \sum_{x\neq 1} p_{\ddot{\theta}}(x)\boldsymbol{L}(x)\boldsymbol{L}(x)^T - \left(\sum_{x\neq 1} p_{\ddot{\theta}}(x)\boldsymbol{L}(x)\right)\left(\sum_{x\neq 1} p_{\ddot{\theta}}(x)\boldsymbol{L}(x)\right)^T \qquad \text{(G.3)} \\
&= \mathbb{L}\ddot{\boldsymbol{P}}\mathbb{L}^T - (\mathbb{L}\ddot{\boldsymbol{p}})(\mathbb{L}\ddot{\boldsymbol{p}})^T \\
&= \mathbb{L}(\ddot{\boldsymbol{P}} - \ddot{\boldsymbol{p}}\ddot{\boldsymbol{p}}^T)\mathbb{L}^T \qquad \text{(G.4)}
\end{aligned}$$

where (G.3) follows recalling that $\boldsymbol{L}(1) = \boldsymbol{0}$. We then show that $\left(\ddot{\boldsymbol{P}} - \ddot{\boldsymbol{p}}\ddot{\boldsymbol{p}}^T\right)$ is non-singular. Observe that

$$\begin{aligned}
\det(\ddot{\boldsymbol{P}} - \ddot{\boldsymbol{p}}\ddot{\boldsymbol{p}}^T) &= (1 - \boldsymbol{p}^T\ddot{\boldsymbol{P}}^{-1}\ddot{\boldsymbol{p}})\det\ddot{\boldsymbol{P}} \qquad \text{(G.5)} \\
&= \left(1 - (p_{\ddot{\theta}}(2) + \cdots + p_{\ddot{\theta}}(|\mathcal{X}|))\right)\det\ddot{\boldsymbol{P}} \\
&= p_{\ddot{\theta}}(1)\det\ddot{\boldsymbol{P}} \\
&= p_{\ddot{\theta}}(1)p_{\ddot{\theta}}(2)\cdots p_{\ddot{\theta}}(|\mathcal{X}|) \\
&\geq p_{\min}^{|\mathcal{X}|} \qquad \text{(G.6)}
\end{aligned}$$

where (G.5) is from Matrix determinant Lemma Harville (1997), while existence of a constant $p_{\min}$ in (G.6) such that $p_{\ddot{\theta}}(x) \geq p_{\min}\ \forall x \in \mathcal{X}$ follows from compactness of Θ and structure of the exponential family (3.1). Since $\mathbb{L}$ is full rank and rank of a matrix is invariant under multiplication by a non-singular matrix, (G.4) implies $\det M\left[\ddot{\theta}\right] > 0$. Positivity of $\det M\left[\ddot{\theta}\right]$, in turn provides a positive constant lower bound C_l for $\det M\left[\hat{\theta}(x^n)\right]$. Let $C_M = \frac{1}{2}\max\{|\log C_l|, |\log C_u|\}$ and $C = C_M + 1$ be the constant in the lemma. Finally, lemma follows by noticing that

$$\log p_{\hat{\theta}(x^n)}(x^n) = n\left[\left\langle \hat{\theta}(\boldsymbol{\tau}(\boldsymbol{L})), \boldsymbol{\tau}(\boldsymbol{L})\right\rangle - \psi\left(\hat{\theta}(\boldsymbol{\tau}(\boldsymbol{L}))\right)\right]$$

for any x^n with $\boldsymbol{L}(x^n) = \boldsymbol{L}$.

APPENDIX H

PROOF OF LEMMA 11: RATIO OF THE VOLUMES

Similar to the Appendix E, one can show that $f_0(\ell)$ is a Lipschitz function of ℓ. Therefore, for a Lipschitz constant $K_5 > 0$, we have $\|f_0(\ell) - f_0(\boldsymbol{L})\| \leq K_5\|\ell - \boldsymbol{L}\|$. Let $R := \sum_{i=1}^{|\mathcal{X}|} \|\boldsymbol{L}(i)\|$. We first show that

$$\mathcal{A}_R := \left\{\ell \in \mathfrak{L} : \gamma_0' - \Delta + \frac{K_5 R}{n} < f_0(\ell) \leq \gamma_0' - \frac{K_5 R}{n}\right\} \subseteq \bigcup_{\boldsymbol{L}\in\mathcal{A}_0} B_{\frac{R}{n}}(\boldsymbol{L}). \tag{H.1}$$

For an arbitrary $\ell \in \mathcal{A}_R$, since $\mathcal{A}_R$ is a subset of the convex hull of $\mathcal{L}$, one can find real non-negative numbers $a_i, i = 1, ..., |\mathcal{X}|$ such that

$$\sum_{i=1}^{|\mathcal{X}|} a_i = 1 \tag{H.2}$$

and

$$\ell = \sum_{i=1}^{|\mathcal{X}|} a_i \boldsymbol{L}(i). \tag{H.3}$$

For an index j, let $n_i = \lfloor na_i \rfloor$ for $i = 1, ..., j$ and $n_i = \lceil na_i \rceil$ for $i = j+1, ..., |\mathcal{X}|$. We claim one can choose the index $0 \leq j \leq |\mathcal{X}|$ ($j = 0$ corresponds to $n_i = \lceil na_i \rceil$ for all i) such that $\sum_{i=1}^{|\mathcal{X}|} n_i = n$. Observe that for $j = 0$, we have $\sum_{i=1}^{|\mathcal{X}|} n_i \geq n$, while for $j = |\mathcal{X}|$, $\sum_{i=1}^{|\mathcal{X}|} n_i \leq n$. Incrementing j by one, decreases the integer $\sum_{i=1}^{|\mathcal{X}|} n_i$ by at most one. The claim then follows.

It is clear that n_i's satisfy the following condition as well

$$|n_i - na_i| < 1, \ \ \forall i = 1, ..., |\mathcal{X}|. \tag{H.4}$$

Let $x^n \in \mathcal{X}^n$ be any sequence with empirical probability mass function $\left\{\frac{n_i}{n}\right\}$. Observe that

$$\boldsymbol{L}(x^n) = \frac{1}{n}\sum_{i=1}^{|\mathcal{X}|} n_i \boldsymbol{L}(i) \in \mathcal{L}.$$

Therefore one obtains

$$\|\ell - \boldsymbol{L}(x^n)\| \leq \frac{1}{n}\sum_{i=1}^{|\mathcal{X}|} |n_i - na_i| \cdot \|\boldsymbol{L}(i)\| \tag{H.5}$$

$$< \frac{1}{n}\sum_{i=1}^{|\mathcal{X}|} \|\boldsymbol{L}(i)\| \tag{H.6}$$

$$= \frac{R}{n} \tag{H.7}$$

where (H.5) follows from (H.3) and the Cauchy-Schwarz inequality, (H.6) follows from (H.4). Therefore $\ell \in B_{\frac{R}{n}}(\boldsymbol{L}(x^n))$. We then show that $\boldsymbol{L}(x^n) \in \mathcal{A}_0$. From (H.7) and the Lipschitzness of $f_0(\cdot)$ we have

$$f_0(\ell) - \frac{K_5 R}{n} \leq f_0\left(\boldsymbol{L}(x^n)\right) \leq f_0(\ell) + \frac{K_5 R}{n}. \tag{H.8}$$

From (H.8,H.1) and since $\ell \in \mathcal{A}_R$, we obtain

$$\gamma_0' - \Delta < f_0\left(\boldsymbol{L}(x^n)\right) \leq \gamma_0' \tag{H.9}$$

which confirms $\boldsymbol{L}(x^n) \in \mathcal{A}_0$. Since for an arbitrary $\ell \in \mathcal{A}_R$, we are able to find $\boldsymbol{L}(x^n) \in \mathcal{A}_0$ within a distance of $\frac{R}{n}$, (H.1) follows.

We continue by observing the following

$$\mathrm{Vol}\left(\bigcup_{\ell \in \mathcal{A}_R} B_{\frac{1}{2n}}(\ell)\right) \leq \mathrm{Vol}\left(\bigcup_{\boldsymbol{L} \in \mathcal{A}_0} B_{\frac{2R+1}{2n}}(\boldsymbol{L})\right) \tag{H.10}$$

$$\leq (2R+1)^{d'} \mathrm{Vol}\left(\bigcup_{\boldsymbol{L} \in \mathcal{A}_0} B_{\frac{1}{2n}}(\boldsymbol{L})\right) \tag{H.11}$$

where (H.10) is from (H.1) and a geometrical observation (triangle inequality) that if a point is within a distance $\frac{1}{2n}$ of a point in $\mathcal{A}_R$, it is certainly within a distance $\frac{R}{n} + \frac{1}{2n}$ of a point in $\mathcal{A}_0$, (H.11) follows since scaling the radius of an sphere by a constant, changes its volume by a constant multiplicative factor.

Given (H.11), to prove the lemma it is enough to show that for some constant $C > 0$,

$$\frac{\mathrm{Vol}\left(\bigcup_{\ell \in \tilde{\mathcal{A}}_0} B_{\frac{1}{2n}}(\ell)\right)}{\mathrm{Vol}\left(\bigcup_{\ell \in \mathcal{A}_R} B_{\frac{1}{2n}}(\ell)\right)} \leq C. \tag{H.12}$$

Observe the following

$$\mathrm{Vol}\left(\bigcup_{\ell \in \tilde{\mathcal{A}}_0} B_{\frac{1}{2n}}(\ell)\right) - \mathrm{Vol}\left(\bigcup_{\ell \in \mathcal{A}_R} B_{\frac{1}{2n}}(\ell)\right) \tag{H.13}$$

$$\leq \mathrm{Vol}\left(\ell : f_0(\ell) \in \left(\gamma_0' - \Delta - \frac{K_5}{2n}, \gamma_0' + \frac{K_5}{2n}\right]\right) \tag{H.14}$$

$$- \mathrm{Vol}\left(\ell : f_0(\ell) \in \left(\gamma_0' - \Delta + \frac{K_5 R}{2n}, \gamma_0' - \frac{K_5 R}{2n}\right]\right) \tag{H.15}$$

$$= \rho_0\left(\gamma_0' + \frac{K_5}{2n}\right) - \rho_0\left(\gamma_0' - \Delta - \frac{K_5}{2n}\right) + \rho_0\left(\gamma_0' - \Delta + \frac{K_5 R}{2n}\right) - \rho_0\left(\gamma_0' - \frac{K_5 R}{2n}\right) \tag{H.16}$$

$$\leq \frac{C}{n} \tag{H.17}$$

where (H.14) is an upper bound for the first term in (H.13) noticing the definition of $\tilde{\mathcal{A}}_0$ and Lipschitzness of $f_0(\cdot)$, (H.15) is from lower bounding the volume of the ball-covering of $\mathcal{A}_R$ (second term in (H.13)) by the volume of $\mathcal{A}_R$ itself, (H.16) is from the definition of $\rho_0(\cdot)$, and (H.17) is from Lipschitzness of $\rho_0(\cdot)$ and recalling the

choice of $\Delta = \frac{1}{n}$. Therefore

$$\begin{aligned}
\frac{\mathrm{Vol}\left(\bigcup_{\ell\in\tilde{\mathcal{A}}_0} B_{\frac{1}{2n}}(\ell)\right)}{\mathrm{Vol}\left(\bigcup_{\ell\in\mathcal{A}_R} B_{\frac{1}{2n}}(\ell)\right)} &\leq \frac{\mathrm{Vol}\left(\bigcup_{\ell\in\mathcal{A}_R} B_{\frac{1}{2n}}(\ell)\right) + \frac{C}{n}}{\mathrm{Vol}\left(\bigcup_{\ell\in\mathcal{A}_R} B_{\frac{1}{2n}}(\ell)\right)} \\
&= 1 + \frac{C}{n\mathrm{Vol}\left(\bigcup_{\ell\in\mathcal{A}_R} B_{\frac{1}{2n}}(\ell)\right)} \\
&\leq 1 + \frac{C}{n\left(\rho_0\left(\gamma_0' - \frac{K_5 R}{2n}\right) - \rho_0\left(\gamma_0' - \Delta + \frac{K_5 R}{2n}\right)\right)} \qquad \text{(H.18)} \\
&\leq 1 + \frac{C}{K_4(K_5 R + 1)} \qquad \text{(H.19)}
\end{aligned}$$

where (H.18) is by lower bounding the volume of the ball-covering of $\mathcal{A}_R$ by the volume of $\mathcal{A}_R$ itself, along with the definition of $\rho_0(\cdot)$, and (H.19) is an application of Lemma 12 as well as recalling the choice of $\Delta = \frac{1}{n}$. This proves (H.12), and the lemma follows.

APPENDIX I

PROOF OF LEMMA 12: LOWER BOUND ON $|\frac{D}{D\lambda}\rho_0(\lambda)|$

Denote $\mathcal{K}_0 = \{\ell \in \mathfrak{L} : f_0(\ell) \leq \lambda\}$ and $\mathcal{K}_0^c = \mathfrak{L} \backslash \mathcal{K}_0$. Furthermore, let us denote $\mathcal{K}_{0,\epsilon} = \{\ell \in \mathfrak{L} : f_0(\ell) \leq \lambda + \epsilon\}$. We have

$$\begin{aligned}\frac{d}{d\lambda}\rho_0(\lambda) &= \lim_{\epsilon \to 0} \frac{\rho_0(\lambda+\epsilon) - \rho_0(\lambda)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\mathrm{Vol}(\mathcal{K}_{0,\epsilon}) - \mathrm{Vol}(\mathcal{K}_0)}{\epsilon}. \end{aligned} \tag{I.1}$$

Let ℓ_1 be an arbitrary point in $\mathcal{K}_0$. Let

$$\ell_2 = \ell_1 + \frac{\epsilon}{2} \frac{\nabla f_0(\ell_1)}{\|\nabla f_0(\ell_1)\|^2}. \tag{I.2}$$

From Taylor series expansion, we have

$$f_0(\ell_2) = f_0(\ell_1) + \langle \nabla f_0(\ell_1), \ell_2 - \ell_1 \rangle + \Delta_0 \tag{I.3}$$

where $|\Delta_0| \leq C_{f_0} \|\ell_1 - \ell_2\|^2$, for a constant C_{f_0} which is independent of n. First observe from (I.2) that, since $\epsilon \to 0$ is infinitesimal, ℓ_2 resides in the vicinity of ℓ_1 with distance at most

$$\|\ell_2 - \ell_1\| < \sqrt{\frac{\epsilon}{2C_{f_0}}}. \tag{I.4}$$

With the choice of ℓ_2 in (I.2), we have

$$f_0(\ell_2) < f_0(\ell_1) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \tag{I.5}$$

$$\leq \lambda + \epsilon \tag{I.6}$$

where (I.5) follows from (I.2,I.3,I.4), and (I.6) is a consequence of ℓ_1 being a point in $\mathcal{K}_0$. Therefore $\ell_2 \in \mathcal{K}_{0,\epsilon}$. As a conclusion for all $\ell_1 \in \mathcal{K}_0$, $\ell_1 + \frac{\epsilon}{2} \frac{\nabla f_0(\ell_1)}{\|f_0(\ell_1)\|^2} \in \mathcal{K}_{0,\epsilon}$. That translates into the following subset relationship

$$\mathcal{K}_0 + B\left(\frac{\epsilon}{2} \frac{\nabla f_0(\ell_1)}{\|f_0(\ell_1)\|^2}\right) \subset \mathcal{K}_{0,\epsilon}. \tag{I.7}$$

Continuing from (I.1) we have

$$\left|\frac{d}{d\lambda}\rho_0(\lambda)\right| \geq \lim_{\epsilon \to 0} \frac{\mathrm{Vol}\left(\mathcal{K}_0 + B\left(\frac{\epsilon}{2} \frac{\nabla f_0(\ell_1)}{\|\nabla f_0(\ell_1)\|^2}\right)\right) - \mathrm{Vol}(\mathcal{K}_0)}{\epsilon} \tag{I.8}$$

$$\geq \frac{S(\mathcal{K}_0)}{2\|\nabla f_0(\ell_1)\|} \tag{I.9}$$

$$= \frac{S(\mathcal{K}_0)}{2\|\hat{\theta}(\ell_1)\|} \tag{I.10}$$

$$\geq \frac{S(\mathcal{K}_0)}{\wp} \tag{I.11}$$

where (I.8) is a consequence of (I.7), (I.9) is due to the definition of the surface area in (F.1), (I.10) is derived similar to (E.3), and finally (I.11) is from the fact that for all $\theta \in \Theta$, we have $\|\theta\| \leq \wp$.

It remains to provide a positive constant lower bound for $S(\mathcal{K}_0)$ independent of n. We first show that in the range $\gamma_0' - \Delta \leq \lambda \leq \gamma_0'$, there exists a positive constant lower bound for $\mathrm{Vol}(\mathcal{K}_0)$. Since

$$\Upsilon(\ell) := -\frac{1}{n}\left(\left\langle \hat{\theta}(\boldsymbol{\tau}(\ell)), \boldsymbol{\tau}(\ell)\right\rangle - \psi\left(\hat{\theta}\left(\boldsymbol{\tau}(\ell)\right)\right)\right) \tag{I.12}$$

$$= -\frac{1}{n}\left(\left\langle \hat{\theta}(\mathbf{b} + \mathbb{A}\boldsymbol{R}\ell), \mathbf{b} + \mathbb{A}\boldsymbol{R}\ell\right\rangle - \psi\left(\hat{\theta}(\mathbf{b} + \mathbb{A}\boldsymbol{R}\ell)\right)\right) \tag{I.13}$$

is a continuous function of ℓ over a compact domain $\mathfrak{L}$, it attains a minimum at a point, say $\ell^* \in \mathfrak{L}$. This minimum is certainly less than or equal to the minimum of $\Upsilon(\ell)$ over $\mathcal{L}$, which is attained at a point say $\boldsymbol{L}^*$. For any $\theta \in \Theta$, we have

$$\Upsilon(\ell^*) \leq \Upsilon(\boldsymbol{L}^*)$$

$$\leq \sum_{x^n} p_\theta(x^n)\left(-\frac{1}{n}\log p_{\hat{\theta}(x^n)}(x^n)\right) \tag{I.14}$$

$$\leq \sum_{x^n} p_\theta(x^n)\left(-\frac{1}{n}\log p_{\theta}(x^n)\right) \tag{I.15}$$

$$= H(p_\theta) \tag{I.16}$$

where (I.14) follows since $\Upsilon(\boldsymbol{L}^*) = -\frac{1}{n}\log p_{\hat{\theta}(y^n)}(y^n)$ for some $y^n \in \mathcal{X}^n$ with $\boldsymbol{L}(y^n) = \boldsymbol{L}^*$, more precisely

$$\Upsilon(\boldsymbol{L}^*) = \min_{x^n \in \mathcal{X}^n} -\frac{1}{n}\log p_{\hat{\theta}(x^n)}(x^n)$$

and the minimum value of a function is less than or equal to its weighted average with respect to any weighting, (I.15) is from $p_{\hat{\theta}(x^n)}(x^n) \geq p_\theta(x^n)$.

Recall $H := H(p_{\theta^*})$ as the entropy of the underlying model. We provide a positive lower bound, independent of n for δ defined as follows:

$$\delta := H - \Upsilon(\ell^*). \tag{I.17}$$

We assume that the underlying model is not the lowest entropy model in the class, i.e. $H > \min_{\theta\in\Theta} H(p_\theta)$. Since $H(p_\theta)$ is a continuous function of θ over a compact domain, $\min_{\theta\in\Theta} H(p_\theta)$ is achieved for a model in the class, say $\theta_{min} \in \Theta$. We then have

$$\delta \geq H - H(p_{\theta_{min}}) \tag{I.18}$$

$$> 0 \tag{I.19}$$

where (I.18) follows from (I.17) and since (I.16) is true for any $\theta \in \Theta$ including θ_{min}, (I.19) is from the assumption that $H > \min_{\theta\in\Theta} H(p_\theta)$.

Similar to the Appendix E, one can show that $f_0(\ell)$ is a Lipschitz function of ℓ with Lipschitz constant $K_5 > 0$. For any $\ell \in \mathfrak{L}$ with $\|\ell - \ell^*\| \leq \frac{\delta}{2K_5}$, we have

$$f_0(\ell) \leq f_0(\ell^*) + K_5 \cdot \frac{\delta}{2K_5} \tag{I.20}$$
$$= \Upsilon(\ell^*) - \frac{d'}{2n}\log(2\pi n) + \frac{C}{n} + \frac{\delta}{2} \tag{I.21}$$
$$= H - \delta - \frac{d'}{2n}\log(2\pi n) + \frac{C}{n} + \frac{\delta}{2} \tag{I.22}$$
$$< H - \frac{\delta}{3} \tag{I.23}$$
$$< \gamma_0' - \Delta \tag{I.24}$$
$$\leq \lambda \tag{I.25}$$

where (I.20) follows from the Lipschitzness of $f_0(\cdot)$ with Lipschitz constant K_5, (I.21) is from (I.12,3.46), (I.22) is from the definition of δ in (I.17), (I.23) holds for large enough n, and (I.24) holds for large enough n, recalling the choices of γ_0' in (3.62) and $\Delta = \frac{1}{n}$, and (I.25) is due to the range of λ. Therefore, from the definition of $\mathcal{K}_0$, we obtain the following relation

$$\left\{\ell \in \mathfrak{L} : \|\ell - \ell^*\| \leq \frac{\delta}{2K_5}\right\} \subset \mathcal{K}_0.$$

Hence

$$\mathrm{Vol}(\mathcal{K}_0) \geq \mathrm{Vol}\left(\left\{\ell \in \mathfrak{L} : \|\ell - \ell^*\| \leq \frac{\delta}{2K_5}\right\}\right)$$
$$= C\left(\frac{\delta}{2K_5}\right)^{d'} \tag{I.26}$$
$$\geq C\left(\frac{H - H(p_{\theta_{min}})}{2K_5}\right)^{d'} \tag{I.27}$$

where (I.26) is from the fact that the intersection of the sphere $\|\ell - \ell^*\| \leq \frac{\delta}{2K_5}$ and $\mathfrak{L}$ is independent of n and only depends on the constellation of $\mathcal{L}$, and (I.27) is from (I.18).

Finally, since sphere has the smallest surface area among all shapes of a given volume, therefore a positive constant lower bound on $\mathrm{Vol}(\mathcal{K}_0)$, implies a positive constant lower bound on $S(\mathcal{K}_0)$. More precisely, recalling the equations for the volume and the surface area of a d'-dimensional sphere (Ren, 1994, Eq. 1.5.1), we have

$$S(\mathcal{K}_0) \geq C\left(\frac{H - H(p_{\theta_{min}})}{2K_5}\right)^{d'} \frac{2\sqrt{\pi}\Gamma\left(\frac{d'}{2}+1\right)}{\Gamma\left(\frac{d'+1}{2}\right)}.$$

APPENDIX J

PROOF OF $|\mathcal{A}| \leq \ell^{D-1}$

Observe that the type complexity bound (4.10), implies the following subset relationships

$$\begin{aligned}
&\left\{T \in \mathcal{T}_\ell : \exists x^\ell \in T \text{ with } S(x^\ell) < \gamma\right\} \\
&\qquad \subseteq \left\{T \in \mathcal{T}_\ell : \exists x^\ell \in T_{y^\ell} \text{ with } -\log p_{\hat{\theta}(x^\ell)}(x^\ell) + C_1 < \gamma\right\} \qquad \text{(J.1)} \\
&\left\{T \in \mathcal{T}_\ell : \exists x^\ell \in T \text{ such that } \exists x_{\ell+1} \text{ with } S(x^\ell x_{\ell+1}) > \gamma\right\} \\
&\qquad \subseteq \left\{T \in \mathcal{T}_\ell : \exists x^\ell \in T_{y^\ell} \text{ such that } \exists x_{\ell+1} \text{ with } -\log p_{\hat{\theta}(x^\ell x_{\ell+1})}(x^\ell x_{\ell+1}) + C_2 > \gamma\right\}. \qquad \text{(J.2)}
\end{aligned}$$

Hence Lemma 14 along with (J.1,J.2) and the definition of $\mathcal{A}$, imply existence of positive constants $C, C' > 0$ such that

$$\mathcal{A} \subseteq \left\{T \in \mathcal{T}_\ell : \exists x^\ell \in T \text{ with } \gamma - C_2 - C_5 < -\log p_{\hat{\theta}(x^\ell)}(x^\ell) < \gamma - C_1\right\}.$$

On the other hand it is shown in (Iri and Kosut, 2016a, Eq. 32) that

$$\left|\left\{T \in \mathcal{T}_\ell : \exists x^\ell \in T \text{ with } \gamma - C_2 - C_5 < -\log p_{\hat{\theta}(x^\ell)}(x^\ell) < \gamma - C_1\right\}\right| \leq \ell^{d-1}.$$

This completes the proof.

APPENDIX K

SOLVING FOR R IN THE ACHIEVABILITY

In order for (4.30) to be less than or equal to ϵ, it must hold that

$$\gamma - C_2 - \frac{\log M}{R} H \leq \sigma \sqrt{\frac{\log M}{R}} Q^{-1} \left(\epsilon - \frac{A}{\sqrt{\frac{\log M}{R}}} \right).$$

Taylor expansion of $Q^{-1}(\cdot)$ yields

$$R \leq H + \sigma \sqrt{\frac{R}{\log M}} Q^{-1}(\epsilon) + R \frac{d}{2} \frac{\log \log M}{\log M} + \frac{C}{\log M}$$

for some constant C. We then solve for R, iteratively. Assuming M is large enough, one can show that $R \leq H + \delta_1$ with $\delta_1 = o(1)$. Next, one can show that $\delta_1 \leq \sigma \sqrt{\frac{H}{\log M}} Q^{-1}(\epsilon) + \delta_2$ with $\delta_2 = o\left(\frac{1}{\sqrt{\log M}}\right)$, where Taylor expansion of $\sqrt{H + \delta_1}$ is employed. Finally, one can show that $\delta_2 \leq \frac{d}{2} H \frac{\log \log M}{\log M} + \delta_3$, where $\delta_3 = \mathcal{O}\left(\frac{1}{\log M}\right)$.