UNIVERSITY OF CALIFORNIA SAN DIEGO

Universal Probability and Its Applications

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Electrical Engineering (Communication Theory and Systems)

by

Alankrita Bhatt

Committee in charge:

Professor Young-Han Kim, Chair
Professor Ery Arias-Castro
Professor Jelena Bradic
Professor Yoav Freund
Professor Tara Javidi
Professor Piya Pal

2022

The Dissertation of Alankrita Bhatt is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I am tremendously grateful for the guidance and support I have received from several people over the course of my PhD. First of all, I must thank my advisor Young-Han Kim for being an excellent mentor and a role model to me in teaching and research. I have had the good fortune to witness firsthand his breadth of knowledge, deep understanding of information theory and many other subjects, and his excellent sense of humour. I'm also thankful to him for the amount of freedom I had in exploring research problems and directions that interested me.

I am grateful to my committee members Ery Arias-Castro, Jelena Bradic, Yoav Freund, Tara Javidi and Piya Pal for volunteering their precious time to be on my committee, for engaging with my work and providing valuable feedback, and for their encouragement. I have also taken several classes taught by my committee members, and I thank them as well as other Professors at UCSD for the wonderful learning experience of the past 6 years.

I want to thank my groupmates, whom I've had the good fortune to interact with throughout my PhD: Shouvik Ganguly, Nadim Ghaddar, Jiun-Ting Huang, Jongha Ryu, and Pinar Sen. Attending group meetings, TA-ing courses and preparing NSF proposals was always fun thanks to their company. I want to especially thank Jongha for a sequence of joint work, some of which appears in this thesis; Nadim and Jiun-Ting for being my cohort mates; and Shouvik and Pinar for welcoming me into the group when I first joined. I also must thank the group seniors Fatemeh Arbabjolfaei, Yu Xiang and especially Lele Wang for their advice and encouragement. Lele has been a valuable mentor and role model to me from the very beginning of my PhD, and I have truly learnt so much from her—both in terms of mathematical and technical skills as well as general career advice. I am grateful to her for all the research discussions, encouragement and uplifting messages when I needed it.

Apart from groupmates, I am also grateful to collaborators Or Ordentlich, Ankit Pensia, Chi Wang and Ziao Wang for working with me; our interactions taught me so much. I am particularly grateful to Or Ordentlich for his kindness and mentorship, I have learnt a lot from him. I am so thankful for the many big and small conversations I have had with people in UCSD, San Diego and elsewhere that have inspired me. In San Diego, I am particularly grateful to Yeohee Im for her friendship.

My friends from college have been there for me for almost a decade now, to share in the joyful moments and help me through the hard times. I am profoundly grateful to Amrita, Anushka, Jaya, Saloni, Sanjana and Shalini for their continents-spanning friendship, care and love. I also thank Shehzad Hathi for his friendship and shared interests since way back. I am extremely grateful to Supranta for a wonderful friendship and companionship that I hope to share for many years to come.

I want to thank my parents and younger sister for everything they have done for me. I am lucky to have parents with such a firm belief in the importance and power of education, and am so proud of my amazing and most beloved sister. Thank you for everything.

The following co-authored material has been used in the this dissertation.

Chapter 2 includes the material in Alankrita Bhatt, Ziao Wang, Chi Wang, Lele Wang, “Universal graph compression: Stochastic block models”, arXiv 2006.02643 (to be submitted), with a short version in the IEEE International Symposium on Information Theory (ISIT), 2021. The dissertation author was a primary investigator and author of this paper.

Chapter 3, in part, includes the material in Alankrita Bhatt and Young-Han Kim, “Sequential prediction under log-loss with side information,” Algorithmic Learning Theory conference, 2021. The dissertation author was a primary investigator and author of this paper.

Chapter 4 includes the material in Alankrita Bhatt, Jongha Ryu, and Young-Han Kim “On universal portfolios with continuous side information”, arXiv:2207.12382. The dissertation author was a primary investigator and author of this paper.

VITA

2016	B. Tech. in Electrical Engineering, Indian Institute of Technology Kanpur
2020	M. S. in Statistics, University of California San Diego
2022	Ph. D. in Electrical Engineering (Communication Theory and Systems), University of California San Diego

PUBLICATIONS

J. Ryu, A. Bhatt, On Confidence Sequences for Bounded Random Processes via Universal Gambling Strategies, *arXiv:2207.12382*, 2022.

A. Bhatt, J. Ryu, Y.-H. Kim, On Universal Portfolios with Continuous Side Information, *arXiv:2202.02431*, 2022.

J. Ryu, A. Bhatt, Y.-H. Kim, Parameter-free Online Optimization with Side Information via Universal Coin Betting, *AISTATS*, 2022.

A. Bhatt, A. Pensia, Sharp Concentration Inequalities for the Centered Relative Entropy, *Information and Inference: A Journal of the IMA*, 2022.

A. Bhatt, Y.-H. Kim, Sequential prediction under log-loss with side information, in *Algorithmic Learning Theory (ALT)*, 2021.

A. Bhatt, Z. Wang, C. Wang and L. Wang, Universal Graph Compression: Stochastic Block Models. Full version *arXiv 2006.02643 to be submitted.* Short version in *IEEE International Symposium on Information Theory (ISIT),* 2021.

A. Bhatt, B. Nazer, O. Ordentlich and Y. Polyanskiy, Information-distilling quantizers. *IEEE Trans. Inf. Theory*, vol. 67, no. 4, pp. 2472–2487, Apr. 2021.

T.-W Ban, A. Bhatt and Y.-H Kim, An Efficient Method to Monitor Downlink Traffic for 4G and 5G Networks. In *IEEE Global Communications Conference (GLOBECOM)*, 2019.

A. Bhatt, J.-T Huang, Y.-H Kim, J. Ryu and P. Sen, Variations on a theme by Liu, Cuff, and Verdu: The power of posterior sampling. In *IEEE Information Theory Workshop (ITW)*, 2018.

A. Bhatt, J.-T Huang, Y.-H Kim, J. Ryu and P. Sen, Monte carlo methods for randomized likelihood decoding. In *56th Annual Allerton Conference on Communication, Control and Computation* , 2018.

A. Bhatt, N. Ghaddar, and L. Wang, Polar coding for multiple descriptions using monotone chain rules. In *55th Annual Allerton Conference on Communication, Control and Computation*, 2017.

ABSTRACT OF THE DISSERTATION

Universal Probability and Its Applications

by

Alankrita Bhatt

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California San Diego, 2022

Professor Young-Han Kim, Chair

In modern statistical and data science applications, the probability distribution generating the data in question is unknown (or even absent) and decisions must be taken in a purely data-driven manner. Thus motivated, in this dissertation the information-theoretic approach of universal probability is revisited and expanded upon. This approach gives us general principles and guidelines for assigning sequential probabilities to data (based on which a decision can then be made), and has been used successfully over the years to problems in compression and estimation among others. The utility of this approach is then demonstrated through three example problems, motivated by the aforementioned modern statistical applications—-universal compression of graphical data, sequential prediction with side information, and universal portfolio

selection with side information.

Chapter 1

Introduction

Statistical applications in the current era present a host of new challenges that require one to move beyond traditional methods and assumptions. Some of these challenges include unknown data distribution, sequentially available data (as opposed to batch) [1,2], high-dimensionality [3, 4], limited memory, and outliers in data [5]. This dissertation focuses on some techniques and methods that confront the first aforementioned challenge. In particular, the information-theoretic approach of *universal probability* is taken. Presently, we explain this approach and outline some key ideas underpinning it.

1.1 Universal Probability Assignment

Consider the problem of a weatherperson trying to predict whether or not it will rain tomorrow in La Jolla. On day t, she has access to the history of the weather $y^{t-1} \in \{0,1\}^{t-1}$ (where 1 represents rain and 0 represents no rain), and needs to form an estimate $\hat{y}_t \in \{0,1\}$ of what y_t will be (i.e., tell whether or not it will rain tomorrow). After the day is over and the true weather (i.e. y_t) is revealed, she suffers a loss of $\mathbb{1}\{\hat{y}_t \neq y_t\} := |\hat{y}_t - y_t|$, which adds up over time. The final goal is to suffer as little cumulative loss as possible over some time horizon n (say $n = 365$ representing a year).

Now, one may reasonably assume that the rain will fall on any given day with probability $\theta \in [0,1]$ (i.e. that all $y_i \sim \mathrm{Bern}(\theta)$ independently). Unfortunately, the weatherperson doesn't have access to this true probability θ. If indeed she knew θ, she could just output $\hat{y}_t = \mathbb{1}\{\theta \geq 1/2\}$ everyday; and in a certain sense this is the optimal action to take (for example, $\theta \approx 0.1$ for La Jolla, and so the optimal action to take is to declare $\theta = 0$ everyday). However, she must create an estimate $\hat{y}_t \in \{0,1\}$ using just the history y^{t-1}.

This is where the universal probability assignment approach of information theory can be employed. We know that the data $y^n \sim \text{Bern}(\theta)$ independently and identically distributed (i.i.d.) for some unknown $\theta \in [0,1]$ (let us call this distribution p_θ). The key idea is to construct a *universal probability assignment* q for the class of i.i.d. Bernoulli random variables; i.e. construct a measure q such that $q \approx p_\theta$ (in some sense) for *every* $\theta \in [0,1]$. Once this measure is constructed, we can simply pretend that the actual data is being drawn from this distribution can take the optimal action as per this distribution. In the rain prediction example, the weatherperson can simply calculate $q(1|y^{t-1}) = \frac{q(y^{t-1}1)}{q(y^{t-1})}$ and then declare her prediction $\hat{y}_t = \mathbb{1}\{q(1|y^{t-1}) \geq 1/2\}$. Since q is a universal probability assignment for the class of all Bernoulli i.i.d. random variables by assumption, it can well approximate the true underlying distribution p_θ, and so intuitively we don't expect to lose too much and have performance almost as good as if θ was known in advance! The key question, of course, now becomes how does one construct such a q? Does it even exist?

1.1.1 Construction Of Universal Probability

First, we precisely define universality.

Definition 1 (Universality). *A probability assignment q is said to be universal for the class of distributions $\{p_\theta, \theta \in \Theta\}$ if*

$$\frac{1}{n} D(p_\theta(y^n) \| q(y^n)) \longrightarrow 0$$

for all $\theta \in \Theta$, where $D(\cdot\|\cdot)$ represents the Kullback–Leibler (KL) divergence. Thus, $q \approx p_\theta$ in a KL divergence sense for every $\theta \in \Theta$.

Remark 1 (Why KL divergence?)**.** *The reader may wonder why Definition 1 utilizes the KL divergence and not another measure of distance between two distributions such as the total variation (TV) distance or the χ^2 divergence. One reason why the KL divergence is utilized is because universality was initially considered in the context of compression where the log-loss makes a natural appearance; with close operational connections to several other problems as well (such as gambling). Moreover, it is known that universality in KL divergence suffices to deal with several other loss functions as well, a point discussed further in Section 1.1.2 (see also [6]).*

Remark 2 (Random sequence vs. individual sequence)**.** *We have so far considered the data to be random and following a certain distribution, and this is reflected in Definition 1. However, in several cases it is more reasonable to assume that the data is generated by a malicious adversary (for example when dealing with the stock market). In such cases, a more stringent definition of universality must be employed, the so-called* pointwise *universality, which requires that q satisfy*

$$\max_{\theta\in\Theta, y^n} \frac{1}{n} \log \frac{p_\theta(y^n)}{q(y^n)} \longrightarrow 0,$$

clearly a more rigid requirement than that employed for mean *universality in Definition 1. In this thesis, for the most part, we will deal with random data and sources.*

Next, we talk about a few approaches to constructing such a universal probability assignment. The first approach, which forms a major workhorse for obtaining the results presented in this thesis, is the *mixture* approach. The idea is to construct a universal probability assignment q for a class of distributions $p_\theta, \theta \in \Theta$ by taking a weighted mixture of all p_θ with some appropriate choice of prior $w(\theta)$. For example, consider the aforementioned problem of constructing a universal probability assignment for the class of binary i.i.d. probabilities. Using this mixture idea, we could construct, for $y^n \in \{0,1\}^n$

$$q(y^n) = \int_0^1 p_\theta(y^n) w(\theta) d\theta$$

$$= \int_0^1 \theta^{\sum_{i=1}^n y_i} (1-\theta)^{n-\sum_{i=1}^n y_i} w(\theta) d\theta$$

where the second equality follows since for some $y^n \in \{0,1\}^n$ an i.i.d. Bernoulli(θ) probability distribution would assign $p_\theta(y^n) = \prod_{i=1}^n p_\theta(y_i) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}$. One can then choose an appropriate prior $w(\theta)$ and obtain a q accordingly. We consider two important and relevant choices of prior distribution, that yield a universal probability assignment. The omitted proofs of universality may be found in Appendix A.1.

- **Uniform prior** $(w(\theta)=1)$**:** Perhaps the simplest choice would be to take a uniform prior over all $\theta \in [0,1]$. In this case, recalling the definition of the Beta function (as well as its relationship to the Gamma function) given by

  $$B(a,b) := \int_0^1 t^{a-1}(1-t)^{b-1} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

  it follows that for all $y^n \in \{0,1\}^n$

  $$q_{\mathrm{L}}(y^n) := \int_0^1 \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} d\theta = \frac{1}{(n+1)\binom{n}{\sum_{i=1}^n y_i}}$$

  and it can be shown that $q_{\mathrm{L}}(y^n)$ is universal for the class of Bernoulli i.i.d. distributions. Moreover, we can this probability assignment in a very convenient sequential form as

  $$q_{\mathrm{L}}(1|y^t) = \frac{q_{\mathrm{L}}(y^t 1)}{q_{\mathrm{L}}(y^t)} = \frac{\sum_{i=1}^t y_i + 1}{t+2}$$

  and therefore this is often called the *add-1* probability assignment. A simpler and more direct combinatorial proof for the add-1 rule can be found at [7], [8]. It is also known as the *Laplace* probability assignment, as when asked to calculate the probability that the sun will not rise tomorrow given that it rose for t days, Laplace answered with $\frac{1}{t+2}$ (which is the same answer as the above probability assignment would provide, substituting $\sum_{i=1}^t y_i = 0$).

Even though Laplace's answer appears to yield a somewhat alarmingly large probability of doom, we can see that it is nonetheless somewhat justified by the universality of q_{L} and Cromwell's rule.[1]

- **Beta($\frac{1}{2},\frac{1}{2}$) prior** $\left(w(\theta)=\frac{1}{\pi\sqrt{\theta(1-\theta)}}\right)$**:** In this case, we can show that for any $y^n \in \{0,1\}^n$

$$\begin{aligned} q_{\mathsf{KT}}(y^n) &:= \pi^{-1}\int_0^1 \theta^{\sum_{i=1}^n y_i - 1/2}(1-\theta)^{n-\sum_{i=1}^n y_i - 1/2} d\theta \\ &= \frac{\binom{n}{\sum_{i=1}^n y_i}\binom{2n}{n}}{4^n\binom{2n}{2\sum_{i=1}^n y_i}} \end{aligned}$$

  and $q_{\mathsf{KT}}(y^n)$ is universal for the class of Bernoulli i.i.d. distributions. Once again, this probability assignment can be expressed in a very convenient sequential form as

$$q_{\mathsf{KT}}(1|y^t) = \frac{\sum_{i=1}^t y_i + 1/2}{t+1}$$

  and therefore this is often called the *add-1/2* probability assignment. This is also called the Krichevsky–Trofimov probability assignment [9]. We remark that the Beta($\frac{1}{2},\frac{1}{2}$) prior is particularly special as it can be shown to be the *Jeffreys* prior for i.i.d. Bernoulli distributions, which is known to be the optimal choice of prior in a certain sense; see [10].

The two examples above demonstrate the power of the mixture approach for the simple class of Bernoulli i.i.d. distributions. Another illustration of the power of this mixture approach can be seen in the construction of the *context tree weighting* (CTW) probability assignment [11], which is universal for the class of all variable order Markov processes.

We also briefly mention another approach utilized by the seminal Lempel–Ziv probability assignment [12–14], a so-called *dictionary-based* probability assignment which is universal for

[1] Cromwell's rule states that unless a statement is logically true or false (such as "2+2 = 4" or "2+2 = 5") assigning it probability 1 or 0 should be avoided. This rule may be justified by observing that if one assigns 0 probability to an exceedingly rare event, and said event does in fact occur, certain loss functions (such as the log-loss) might blow up to ∞.

the extremely large class of all stationary ergodic processes. Indeed, it is known that for q_{LZ} and the class of stationary ergodic processes $\mathscr{P}$ [15]

$$\max_{y^n, p \in \mathscr{P}} \frac{1}{n} \log \frac{p(y^n)}{q_{\mathrm{LZ}}(y^n)} = \Theta\left(\frac{1}{\log n}\right).$$

A data compressor based on the Lempel–Ziv probability assignment has been enormously successful in practice and is deployed, among others, in the GIF image format.

Going back to the motivating problem of predicting tomorrow's weather, we can show that if the weatherperson predicts $\hat{y}_t = \mathbb{1}\{q_{\mathsf{KT}}(1|y^{t-1}) \geq 1/2\}$, she will on average make not too many more mistakes than she would have if she knew the true probability of rain tomorrow θ in advance—see A.2.1 in the Appendix for the exact statement and proof, as well as a few remarks.

1.1.2 Other Example Applications

There are several other applications where universal probability can be leveraged to accomplish the task at hand. One of the most well-known is universal sequential prediction [6,16]. Consider a sequence $Y^n \sim p_\theta$, where the distribution p_θ is picked from a larger class where $\theta \in \Theta$. This class Θ could be parametric, for example the class of binary i.i.d. processes from earlier, or it could be far richer and nonparametric such as the class of stationary ergodic processes. At time t, based on the history Y^{t-1}, a learner (such as a weatherperson) must take an action $a(Y^{t-1})$ (such as choose $a(Y^{t-1}) = \hat{y}_t \in \{0,1\}$) in order to minimize some loss function $\ell(a(Y^{t-1}), Y_t)$ (such as $\mathbb{1}\{a(Y^{t-1}) \neq Y_t\}$). If the data-generating distribution was known, then the learner could take the *Bayes optimal* action, i.e. choose

$$a_t^* = \arg\inf \mathsf{E}_{p_\theta}[\ell(a(Y^{t-1}), Y_t) | Y^{t-1}].$$

However, if one has a measure q that is universal for the class of measures $\{p_\theta, \theta \in \Theta\}$ then one can take an action

$$\widehat{a}_t = \arg\inf \mathsf{E}_q[\ell(a(Y^{t-1}), Y_t) \,|\, Y^{t-1}]$$

and hope to perform not too poorly compared to how one would have performed had p_θ been known in advance. Apart from the weather prediction example from earlier, this framework also encompasses the classical information theory problem of universal data compression [17], where the learner is required to encode data in an efficient way using as few bits as possible, without knowing the data distribution in advance (had the data distribution been known already, it would make sense to encode more likely data using fewer bits in order to minimize the average number of bits used). It can also be shown [6] that this universal prediction approach works for *any* bounded loss function—a result striking in its generality (see Theorem 6 in Section A.2 of the Appendix for the precise statement). Apart from prediction and compression, this approach has been utilized to great success in domains such as portfolio selection [18, 19], entropy estimation [20], and more recently online linear optimization [21, 22] and obtaining anytime concentration inequalities [23].

1.2 Some Prior Work

Recall that universal compression is simply sequential prediction with log-loss, so universal compression was one of the first universal prediction problems considered with several landmark methods proposed for this and related problems. Some milestone results include (see also [17, Chapter 13] and the references therein for more details): [24] which created a universal probability assignment for the class of stationary ergodic processes, [9] provided the minmax optimal probability assigment for binary i.i.d. processes, [11] proposed a widely used universal probability for the class of variable order Markov processes. A line of work [25–27] quantified the exact minmax redundancy for probability assignment for finite alphabet i.i.d processes, and highlighted operational connections to problems such as gambling. Such a connection was also noted by Cover and Ordentlich, who studied sequential portfolio selection [18, 19] for an adversarial stock market, see [17, Chapters 6 and 16] for a thorough treatment of connections between data compression and gambling.

Parallel and closely related to this, a new line of thinking pioneered by Rissanen and others focused on viewing learning as essentially data compression and culminated in the development of the minimum description length (MDL) principle [28, 29] with the Kolmogorov complexity as a notion of the algorithmic complexity of describing an object, see [17, Chapter 14] for a detailed treatment of this subject.

As mentioned, universal compression is universal prediction under the log-loss—a natural counterpart to study is universal prediction of binary sequences under the Hamming loss. This was considered in the case of individual sequences by [16], see also the detailed survey of Feder and Merhav [6]. Prior to this, Cover [30] studied a similar binary prediction problem and characterized the optimal solution, as well as the achievable performance.

On the applications side, there has been much work on using this theoretical framework to study practical statistical problems arising in modern settings. For example, [31] studied universal compression when the alphabet size is unknown, and could potentially be quite large. The universal approach was applied to denoising problems such as image denoising in [32, 33]. Universal probability principles have also been applied to estimation problems such as estimating directed information [20] and entropy/mutual information [34]. Some application domains for these methods include areas like genomics [35, 36] and finance [18].

1.3 An Outline Of This Thesis

In Chapter 2, universal compression of data that is in the form of a graph is studied. By creating a universal probability assignment for a class of graphical distributions known as the stochastic block model, in turn a universal compressor is provided and its performance analyzed theoretically and empirically.

In Chapter 3, the earlier example of sequential prediction is expanded upon by introducing some additional (sequential) side information that governs the distribution of the data at each time step. Taking the mixture idea further, a universal sequential predictor (with log-loss being the loss function) is constructed and its performance analyzed.

Finally, in Chapter 4, the classical problem of universal portfolio selection is studied, with the additional introduction of continuous-valued side information (i.e. side information taking values from an infinite set). Using results from Chapter 3 the landmark result of Cover and Ordentlich [19] is extended.

Chapter 2

Universal Graph Compression

2.1 Introduction

In many data science applications, data appears in the form of large-scale graphs. For example, in social networks, vertices represent users and an edge between vertices represents friendship; in the World Wide Web, vertices are websites and edges indicate the hyperlinks from one site to the other; in biological systems, vertices can be proteins and edges illustrate protein-to-protein interaction. Such graphs may contain billions of vertices. In addition, edges tend to be correlated with each other since, for example, two people sharing many common friends are likely to be friends as well. How to efficiently compress such large-scale structural information to reduce the I/O and communication costs in storing and transmitting such data is a persisting challenge in the era of big data.

The literature on graph compression is vast. Existing compression schemes follow various different methodologies. Several methods exploited combinatorial properties such as cliques and cuts in the graph [37, 38]. Many works targeted at domain-specific graphs such as web graphs [39], biology networks [40, 41], and social network graphs [42]. Various representations of graphs were proposed, such as the text-based method, where the neighbor list of each vertex is treated as a "word" [43, 44], and the k^2-tree method, where the adjacency matrix is recursively partitioned into k^2 equal-size submatrices [45]. *Succinct* graph representations that enable certain types of fast computation, such as adjacency query or vertex degree query, were also widely studied [46]. While most compression schemes are for labeled graphs, there are also works considering lossless compression of unlabeled graphs [47–49], graphs with marks on its edges and vertices [50–52], or (correlated) data on the graph [53, 54]. We refer the readers to [55] for an exhaustive survey on lossless graph compression and space-efficient graph representations.

In this paper, we take an information theoretic approach to study lossless compression of a graph. We assume the graph is generated by some random graph model and investigate lossless compression schemes that achieve the theoretical limit, i.e., the entropy of the graph, asymptotically as the number of vertices goes to infinity. When the underlying distribution/statistics of the random graph model is known, optimal lossless compression can be achieved by methods like Huffman coding. However, in most real-world applications, the exact distribution is usually hard to obtain and the data we are given is a single realization of this distribution. This motivates us to consider the framework of *universal compression*, in which we assume the underlying distribution belongs to a known family of distributions and require that the encoder and the decoder should not be a function of the underlying distribution. The goal of universal compression is to design a single compression scheme that universally achieves the optimal theoretical limit, for every distribution in the family, without knowing which distribution generates the data. For this paper, we focus on the family of *stochastic block models*, which are widely used random graph models that capture the clustering effect in social networks. Our goal is to develop a universal graph compression scheme for a family of stochastic block models with as wide range of parameters as possible.

How to design computationally efficient universal compression scheme is a fundamental question in information theory. In the past several decades, a large number of universal compressors were proposed for one-dimensional sequences with fixed alphabet size, whose entropy is linear in the number of variables. Prominent results include the Laplace and Krichevsky–Trofimov (KT) compressors for i.i.d. processes [26, 27], Lempel–Ziv compressor [12, 13] and Burrows–Wheeler transform [56] for stationary ergodic processes, and context tree weighting [57] for finite memory processes. Many of these have been adopted in standard data compression applications such as compress, gzip, GIF, TIFF, and bzip2. Despite these exciting developments, existing universal compression techniques fall short of establishing optimality results for graph data due to the following challenges. Firstly, graph data generated from a stochastic block model has non-stationary two-dimensional correlation, so existing techniques do not immediately apply here. Secondly, in many practical applications, where the graph is sparse, the entropy of the graph may be sublinear in the number of entries in the adjacency matrix.

For the first challenge, a natural question arising is: can we convert the two-dimensional adjacency matrix of the graph into a one-dimensional sequence in some order and apply a universal compressor for the sequence? For some simple graph model such as Erdős–Rényi graph, where each edge is generated i.i.d. with probability p, this would indeed work. For more complex graph models including stochastic block models, it is unclear whether there is an ordering of the entries that results in a stationary process. We will show in Section 2.7 several orders including row-by-row, column-by-column, and diagonal-by-diagonal fail to produce a stationary process. We alleviate this challenge by designing a decomposition of the adjacency matrix into blocks. We then show in Theorem 3 that with a carefully chosen parameter, the block decomposition converts two-dimensional correlated entries into a sequence of *almost* i.i.d. blocks with slowly growing alphabet size. To address the second challenge, we adjust the standard definition of universality, which normalizes the compression length by the number of variables. The new definition of universality accommodates data with unknown leading order in its entropy expression.

Lossless compression for stochastic block models was first studied by Abbe [53] (albeit not under the universal compression framework). The focus there is two-fold: 1) compute the entropy of the stochastic block model; 2) explore the relation between community detection and compression. Several interesting questions were presented: Knowing the community assignments will help compression since edges can be grouped into i.i.d. subsets. But is community detection necessary for compression? In the regime when community detection is not possible, how do we compress the graph? We answer these questions in this paper by presenting a universal compressor that does not require knowledge of the edge probabilities, the community assignments, or the number of communities. Our compressor remains universal even in the regime when community detection is information theoretically impossible. As a consequence, universal compression is a fundamentally easier task than community detection for stochastic block models.

Recently, universal compression of graphs with marked edges and vertices is studied by Delgosha and Anantharam [52, 58]. They focus on the *sparse* graph regime, where the number of edges is in the same order as the number of vertices n. They employ the framework of local weak convergence, which provides a technique to view a sequence of graphs as a sequence of distributions on neighbourhood structures. Built on this framework, they propose an algorithm that compresses graphs by describing the local neighbourhood structures. Moreover, they introduce a universality/optimality criterion through a notion of entropy for graph sequences under the local weak convergence framework, known as the *BC entropy* [59]. This universality criterion is stronger than the one used in this paper. It requires the asymptotic length of the compressor to match the constants in both first and second order terms in Shannon entropy, whereas the universality criterion we use only requires to match the first order term. As a consequence of the stronger criterion, the compressor in [58] is universal over a smaller random graph family. In comparison, we expand the range of edge numbers from $\Theta(n)$ in the sparse regime to $\Theta(n^{\alpha})$ for every $0 < \alpha \leq 2$ and propose a single universal compressor for the whole family under the weaker universality criterion. In Section 2.6, we evaluate the proposed compressor under the criterion in [58] for the family of *symmetric* SBMs. The proposed compressor achieves a similar performance in terms of BC entropy in the sparse regime.

The rest of the paper is organized as follows. In Section 2.1.1, we define universality over a family of graph distributions and the stochastic block models. We present our main result in Section 2.1.2, which is a graph compressor that is universal for a family containing most non-trivial stochastic block models. We describe the proposed graph compressor in Section 2.2. We illustrate key steps in establishing universality in Section 2.3 and elaborate the proof of each step in Section 2.4. In Section 2.6, we provide the second order analysis of the expected length of our compressor and compare it to the one in [58] In Section 2.7, we explain why existing universal compressors developed for stationary processes may not be immediately applicable for some one-dimensional ordering of entries in the adjacency matrix. In Section 2.8, we implement our compressor in four benchmark graph datasets and compare its empirical performance to four competing algorithms.

Notation. For an integer n, let $[n] = \{1,2,\dots,n\}$. Let $\log(\cdot) = \log_2(\cdot)$. We follow the standard order notation: $f(n) = O(g(n))$ if $\lim_{n\to\infty} \frac{|f(n)|}{g(n)} < \infty$; $f(n) = \Omega(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} > 0$; $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$; $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$; $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} = \infty$; and $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$.

2.1.1 Problem Setup

For simplicity, we focus on simple (undirected, unweighted, no self-loop) graphs with labeled vertices in this paper. But our compression scheme and the corresponding analysis can be extended to more general graphs. Let $\mathscr{A}_n$ be the set of all labeled simple graphs on n vertices. Let $\{0,1\}^i$ be the set of binary sequences of length i, and set $\{0,1\}^* = \cup_{i=0}^{\infty}\{0,1\}^i$. A lossless graph compressor $C\colon \mathscr{A}_n \to \{0,1\}^*$ is a one-to-one function that maps a graph to a binary sequence. Let $\ell(C(A_n))$ denote the length of the output sequence. When A_n is generated from a distribution, it is known that the entropy $H(A_n)$ is a fundamental lower bound on the expected length of any lossless compressor [60, Theorem 8.3]

$$H(A_n) - \log(e(H(A_n)+1)) \le \mathsf{E}[\ell(C(A_n))], \tag{2.1}$$

and therefore

$$\liminf_{n\to\infty} \frac{\mathsf{E}[\ell(C(A_n))]}{H(A_n)} \ge 1.$$

Thus, a graph compressor is said to be *universal* for the family of distributions $\mathscr{P}$ if for all distribution $\mathsf{P} \in \mathscr{P}$ and $A_n \sim \mathsf{P}$, we have

$$\limsup_{n\to\infty} \frac{\mathsf{E}[\ell(C(A_n))]}{H(A_n)} = 1. \tag{2.2}$$

A stochastic block model $\mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W})$ defines a probability distribution over $\mathscr{A}_n$. Here n is the number of vertices, L is the number of communities. Each vertex $i \in [n]$ is associated with a community assignment $X_i \in [L]$. The length-L column vector $\mathbf{p} = (p_1, p_2, \ldots, p_L)^T$ is a probability distribution over $[L]$, where p_i indicates the probability that any vertex is assigned community i. $\mathbf{W}$ is an $L \times L$ symmetric matrix, where W_{ij} represents the probability of having an edge between a vertex with community assignment i and a vertex with community assignment j. We say $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W})$ if the community assignments $X_1, X_2, \ldots, X_n$ are generated i.i.d. according to $\mathbf{p}$ and for every pair $1 \le i < j \le n$, an edge is generated between vertex i and vertex j with probability W_{X_i, X_j}. In other words, in the adjacency matrix A_n of the graph, $A_{ij} \sim \mathrm{Bern}(W_{X_i, X_j})$ for $i < j$; the diagonal entries $A_{ii} = 0$ for all $i \in [n]$; and $A_{ij} = A_{ji}$ for $i > j$. We assume all the entries in $\mathbf{W}$ are in the same regime $f(n)$ and write $\mathbf{W} = f(n)\mathbf{Q}$, where $\mathbf{Q}$ is an $L \times L$ symmetric matrix with constant entries $Q_{ij} = \Theta(1)$ for all $i, j \in [L]$. We assume all entries in $\mathbf{p}$ are $\Theta(1)$. We will consider two families of stochastic block models: For $0 < \varepsilon < 1$,

$$\mathscr{P}_1(\varepsilon)\colon \text{SBM with } L = \Theta(1), f(n) = O(1), f(n) = \Omega\left(\tfrac{1}{n^{2-\varepsilon}}\right), \tag{2.3}$$

$$\mathscr{P}_2(\varepsilon)\colon \text{SBM with } L = \Theta(1), f(n) = o(1), f(n) = \Omega\left(\tfrac{1}{n^{2-\varepsilon}}\right). \tag{2.4}$$

Note that the edge probability $\frac{1}{n^2}$ is the threshold for a random graph to contain an edge with high probability [61]. Thus, the family $\mathscr{P}_1(\varepsilon)$ covers most non-trivial SBM graphs. Clearly, $\mathscr{P}_2(\varepsilon)$ is a strict subset of $\mathscr{P}_1(\varepsilon)$, as it does not contain the constant regime $f(n) = 1$.

2.1.2 Main Results

The main contribution of this paper is providing two compressors universal over the classes $\mathscr{P}_1(\varepsilon)$ and $\mathscr{P}_2(\varepsilon)$ respectively for $0 < \varepsilon < 1$. Note that a compressor universal over the class $\mathscr{P}_1(\varepsilon)$ is also universal over the class $\mathscr{P}_2(\varepsilon)$, but our compressor designed specifically for the class $\mathscr{P}_2(\varepsilon)$ has a lower computational complexity. We will formally state the results in the next two theorems.

Theorem 1 (Universality over $\mathscr{P}_1$)**.** *For every* $0 < \varepsilon < 1$*, the graph compressor* C_k *defined in Section 2.2 is universal over the family* $\mathscr{P}_1(\varepsilon)$ *provided that*

$$0 < \delta < \varepsilon, \quad k \leq \sqrt{\delta \log n}, \quad \textit{and} \quad k = \omega(1).$$

Theorem 2 (Universality over $\mathscr{P}_2$)**.** *For every* $0 < \varepsilon < 1$*, the graph compressor* C_1 *defined in Section 2.2 is universal over the family* $\mathscr{P}_2(\varepsilon)$*.*

For now, one can think of k as a parameter that defines a compression scheme C_k—the exact definition will become clear in the next section when we precisely define the compressors.

2.2 Algorithm: Universal Graph Compressor

In this section, we describe our universal graph compression scheme. For each k that divides n, the graph compressor $C_k \colon \mathscr{A}_n \to \{0,1\}^*$ is defined as follows.

- **Block decomposition.** Let $n' = \frac{n}{k}$. For $1 \leq i, j \leq n'$, let $\mathbf{B}_{ij}$ be the submatrix of A_n formed by the rows $(i-1)k+1, (i-1)k+2, \ldots, ik$ and the columns $(j-1)k+1, (j-$

$1)k+2,\ldots,jk$. For example, we have

$$\mathbf{B}_{12} = \begin{bmatrix} A_{1,k+1} & A_{1,k+2} & \cdots & A_{1,2k} \\ A_{2,k+1} & A_{2,k+2} & \cdots & A_{2,2k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k,k+1} & A_{k,k+2} & \cdots & A_{k,2k} \end{bmatrix}. \tag{2.5}$$

We then write A_n in the block-matrix form as

$$A_n = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1,n'} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2,n'} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{n',1} & \mathbf{B}_{n',2} & \cdots & \mathbf{B}_{n',n'} \end{bmatrix}. \tag{2.6}$$

Denote

$$\mathbf{B}_{\text{ut}} := \mathbf{B}_{12}, \mathbf{B}_{13}, \mathbf{B}_{23}, \mathbf{B}_{14}, \mathbf{B}_{24}, \mathbf{B}_{34}, \ldots, \mathbf{B}_{1,n'}, \cdots, \mathbf{B}_{n'-1,n'} \tag{2.7}$$

as the sequence of off-diagonal blocks in the upper triangle and

$$\mathbf{B}_{\text{d}} := \mathbf{B}_{11}, \mathbf{B}_{22}, \ldots, \mathbf{B}_{n',n'} \tag{2.8}$$

as the sequence of diagonal blocks.

- **Binary to m-ary conversion.** Let $m := 2^{k^2}$. Each $k \times k$ block with binary entries in the two block sequences $\mathbf{B}_{\text{ut}}$ and $\mathbf{B}_{\text{d}}$ is converted into a symbol in $[m]$.

- **KT probability assignment.** Apply KT sequential probability assignment for the two m-ary sequences $\mathbf{B}_{\text{ut}}$ and $\mathbf{B}_{\text{d}}$ respectively. Given an m-ary sequence $x_1, x_2, \ldots, x_N$, *KT sequential probability assignment* defines N conditional probability distributions over $[m]$

as follows. For $j = 0, 1, 2, \ldots, N-1$, assign conditional probability

$$q_{\mathrm{KT}}(i|x^j) \coloneqq q_{\mathrm{KT}}(X_{j+1} = i | X^j = x^j) = \frac{N_i(x^j) + 1/2}{j + m/2} \quad \text{for each } i \in [m], \tag{2.9}$$

where $X^j \coloneqq (X_1, \ldots, X_j), x^j \coloneqq (x_1, x_2, \ldots, x_j)$, and $N_i(x^j) \coloneqq \sum_{k=1}^{j} \mathbb{1}\{x_k = i\}$ counts the number of symbol i in x^j.

- **Adaptive arithmetic coding.** With the KT sequential probability assignments, compress the two sequences $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ separately using adaptive arithmetic coding [62] (see description in Algorithm 1). In case $k = 1$, the diagonal sequence $\mathbf{B}_{\mathrm{d}}$ becomes an all-zero sequence since we assume the graph is simple. So we will only compress the off-diagonal sequence $\mathbf{B}_{\mathrm{ut}}$.

Algorithm 1: *m*-ary adaptive arithmetic encoding with KT probability assignment

```
Input  : Data sequence $x^N$, alphabet size $m$
Initialize lower $= 0$, upper $= 1$, logprob $= 0, N_1 = N_2 = \cdots = N_m = 0$;
for $j = 0, 1, \ldots, N-1$ do
    range $\leftarrow$ upper $-$ lower;
    for $i = 1, 2, \ldots, x_{j+1}$ do
        Compute $q_{\mathrm{KT}}(i|x^j) = \frac{N_i + 1/2}{j + m/2}$;
    upper $\leftarrow$ lower $+$ range $\cdot \sum_{i=1}^{x_{j+1}} q_{\mathrm{KT}}(i|x^j)$;
    lower $\leftarrow$ upper $-$ range $\cdot q_{\mathrm{KT}}(x_{j+1}|x^j)$;
    $N_{x_{j+1}} \leftarrow N_{x_{j+1}} + 1$;
    logprob $\leftarrow$ logprob $+ \log(q_{\mathrm{KT}}(x_{j+1}|x^j))$;
Output : the binary representation of $\frac{1}{2}$(lower $+$ upper) with $\lceil -$logprob$\rceil + 1$ bits
```

Given the compressed graph sequence y^L, the number of vertices n and the block size k, the graph decompressor $D_k \colon \{0,1\}^* \to \mathscr{A}_n$ is defined as follows.

- **Adaptive arithmetic decoding.** With the KT sequential probability assignments defined in (2.9), decompress the two code sequences for $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ separately using adaptive arithmetic decoding (see Algorithm 2). The length of data sequence $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ are $\frac{n}{k}(\frac{n}{k} - 1)/2$ and $\frac{n}{k}$ respectively.

Algorithm 2: *m*-ary adaptive arithmetic decoding with KT probability assignment

```
Input  : Binary sequence $y^L$, alphabet size $m = 2^{k^2}$, length of data sequence $N$
Add '0.' before sequence $y^L$ and convert it into a decimal real number $Y$. Initialize
  $\texttt{lower} = 0, \texttt{upper} = 1, N_1 = N_2 = \cdots = N_m = 0$;
for $j = 0, 1, \ldots, N-1$ do
    $\texttt{range} \leftarrow \texttt{upper} - \texttt{lower}$;
    for $i = 1, 2, \ldots, m$ do
        Compute $q_{\mathrm{KT}}(i|x^j) = \frac{N_i + 1/2}{j + m/2}$;
    Find minimum $z \in [m]$ such that $\texttt{lower} + \texttt{range} \cdot \sum_{i=1}^{z} q_{\mathrm{KT}}(i|x^j) > Y$;
    $\texttt{upper} \leftarrow \texttt{lower} + \texttt{range} \cdot \sum_{i=1}^{z} q_{\mathrm{KT}}(i|x^j)$;
    $\texttt{lower} \leftarrow \texttt{upper} - \texttt{range} \cdot q_{\mathrm{KT}}(z|x^j)$;
    $N_z \leftarrow N_z + 1$;
    $x_{j+1} \leftarrow z$;
Output : the $m$-ary data sequence $x_1, x_2, \cdots, x_N$
```

- ***m*-ary to binary conversion.** Each *m*-ary symbol in the sequence is converted to a k^2-bit binary number and further converted into a $k \times k$ block with binary entries.

- **Adjacency matrix recovery.** With the blocks in $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$, recover the adjacency matrix of A_n in the order described in (2.6), (2.7), and (2.8).

One can check that C_k is well-defined. The block decomposition and the binary to *m*-ary conversion are clearly one-to-one. It is also known that for any valid probability assignment, arithmetic coding produces a prefix code, which as also one-to-one.

The computational complexity of the proposed algorithm is $O(2^{k^2} n^2)$. For the choice of k that achieves universality over $\mathscr{P}_1(\varepsilon)$ family in Theorem 1, $O(2^{k^2} n^2) = O(n^{2+\delta})$ for $\delta < \varepsilon$. For the choice of k that achieves universality over $\mathscr{P}_2(\varepsilon)$ family in Theorem 2, $O(2^{k^2} n^2) = O(n^2)$.

The orders in $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ do not matter in terms of establishing universality. The current orders in (2.7) and (2.8) together with arithmetic coding enable a *horizon free* implementation. That is, the encoder does not need to know the *horizon n* to start processing the data and can output partial coded bits *on the fly* before receiving all the data. This leads to short encoding and decoding delay. For some real-world applications, for example, when the number of users increases in a large social network, this compressor has the advantage of not requiring to re-process existing data and re-compress the whole graph from scratch.

Remark 1 (Laplace probability assignment). As an alternative to the KT sequential probability assignment, one can also use the Laplace sequential probability assignment. Given an m-ary sequence $x_1, x_2, \ldots, x_N$, *Laplace sequential probability assignment* defines N conditional probability distributions over $[m]$ as follows. For $j = 0, 1, 2, \ldots, N-1$, we assign conditional probability

$$q_{\mathrm{L}}(X_{j+1} = i | X^j = x^j) = \frac{N_i(x^j) + 1}{j + m} \quad \text{for each } i \in [m]. \tag{2.10}$$

Both methods can be shown to be universal, while Laplace probability assignment has a much cleaner derivation. However, KT probability assignment produces a better empirical performance. For this reason, we keep both in the paper.

2.3 Main Ideas in Establishing Universality

In this section, we establish the universality of the graph compressor in Section 2.2.

Graph Entropy

We first calculate the entropy of the (random) graph A_n, which, recall, is the fundamental lower bound on the expected compression length for any compression scheme. Since to establish optimality we need to show that $\limsup_{n\to\infty} \frac{\mathsf{E}[\ell(C(A_n))]}{H(A_n)} \leq 1$, we will only be concerned with the first order term in $H(A_n)$.

Lemma 1 (Graph entropy)**.** *Let $A_n \sim SBM(n, L, \mathbf{p}, f(n)\mathbf{Q})$ with $f(n) = O(1), f(n) = \Omega\left(\frac{1}{n^2}\right)$, and $L = \Theta(1)$. For $0 \leq p \leq 1$, let $h(p) \triangleq -p\log(p) - (1-p)\log(1-p)$ denote the binary entropy function. For a matrix W with entries in $[0,1]$, let $h(W)$ be a matrix of the same dimension whose (i,j) entry is $h(W_{ij})$. Then*

$$H(A_n) = \binom{n}{2} H(A_{12}|X_1, X_2)(1 + o(1)) \tag{2.11}$$

$$= \binom{n}{2} \mathbf{p}^T h\big(f(n)\mathbf{Q}\big)\mathbf{p} + o\left(n^2 h\big(f(n)\big)\right). \tag{2.12}$$

In particular, when $f(n) = \Omega\left(\frac{1}{n^2}\right)$ *and* $f(n) = o(1)$*, expression* (2.12) *can be further simplified as*

$$H(A_n) = \binom{n}{2} f(n) \log\left(\frac{1}{f(n)}\right) (\mathbf{p}^T\mathbf{Q}\mathbf{p} + o(1)). \tag{2.13}$$

Remark 2. In the regime $f(n) = \Omega\left(\frac{1}{n}\right)$ and $f(n) = O(1)$, the above result has been established in [53]. We extend the analysis to the regime $f(n) = o\left(\frac{1}{n}\right)$ and $f(n) = \Omega(\frac{1}{n^2})$.

Remark 3. Lemma 1 can be used to calculate the entropy of the graph for certain important regimes of $f(n)$, in which the SBM displays characteristic behavior. For $f(n) = 1$, we have $H(A_n) = \binom{n}{2} h\left(\mathbf{p}^T\mathbf{Q}\mathbf{p}\right)(1+o(1))$; for $f(n) = \frac{\log n}{n}$ (the regime where the phase transition for exact recovery of the community assignments occurs [63,64]), we have $H(A_n) = \frac{n\log^2 n}{2}(\mathbf{p}^T\mathbf{Q}\mathbf{p}+o(1))$; when $f(n) = \frac{1}{n}$ (the regime where the phase transition for detection between SBM and the Erdős–Rényi model occurs [65]), we have $H(A_n) = \frac{n\log n}{2}(\mathbf{p}^T\mathbf{Q}\mathbf{p}+o(1))$; when $f(n) = \frac{1}{n^2}$ (the regime where the phase transition for the existence of an edge occurs), we have $H(A_n) = \log n(\mathbf{p}^T\mathbf{Q}\mathbf{p}+o(1))$.

Asymptotic i.i.d. via Block Decomposition

To compress the matrix A_n, we wish to decompose it into a large number of components that have little correlation between them. This leads to the idea of block decomposition described previously. Since the sequence of blocks are used to compress A_n, the next theorem claims these blocks are identically distributed and asymptotically independent in a precise sense described as follows.

Theorem 3 (Block decomposition)**.** *Let* $A_n \sim SBM(n, L, \mathbf{p}, f(n)\mathbf{Q})$ *with* $f(n) = \Omega\left(\frac{1}{n^{2-\varepsilon}}\right)$ *for some* $0 < \varepsilon < 1$*,* $f(n) = O(1)$*, and* $L = \Theta(1)$*. Let k be an integer that divides n and* $n' = n/k$*. Consider the* $k \times k$ *block decomposition in* (2.6)*. We have all the off-diagonal blocks share the same joint distribution; all the diagonal blocks share the same joint distribution. In other words,* *for any* $1 \leq i_1, i_2, j_1, j_2 \leq n'$ *with* $i_1 \neq j_1, i_2 \neq j_2$ *and* $1 \leq l_1, l_2 \leq n'$*, we have*

$$\mathbf{B}_{i_1,j_1} \overset{d}{=} \mathbf{B}_{i_2,j_2},$$
$$\mathbf{B}_{l_1,l_1} \overset{d}{=} \mathbf{B}_{l_2,l_2}.$$

In addition, if $k = \omega(1)$ *and* $k = o(n)$*, we have*

$$\lim_{n\to\infty} \frac{H(\mathbf{B}_{\mathrm{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} = 1. \tag{2.14}$$

Length Analysis for Correlated Sequences

Thanks to this property of the block decomposition, we hope to compress these blocks as if they are independent using a Laplace probability assignment (which, recall, is universal for the class of all m-ary iid processes). However, since these blocks are still correlated (albeit weakly), we will need a result on the performance of Laplace probability assignment on correlated sequences with identical marginals, which we give next.

Theorem 4 (Laplace probability assignment for correlated sequence)**.** *Consider arbitrarily correlated* $Z_1, Z_2, \ldots, Z_N$*, where the marginal distribution of each* Z_i *is identically distributed over an alphabet of size* $m \geq 2$*. Let* $\ell_{\mathrm{L}}(z^N) = \log \frac{1}{q_{\mathrm{L}}(z^N)}$ *where* $q_{\mathrm{L}}(\cdot)$ *is the marginal distribution induced by Laplace probability assignment in* (2.10)

$$q_{\mathrm{L}}(z^N) := \frac{N_1! N_2! \cdots N_m!}{N!} \cdot \frac{1}{\binom{N+m-1}{m-1}}. \tag{2.15}$$

We then have

$$\mathsf{E}[\ell_{\mathrm{L}}(Z^N)] \leq m \log(2eN) + NH(Z_1). \tag{2.16}$$

We provide a similar result for the KT probability assignment.

Theorem 5 (KT probability assignment for correlated sequence)**.** *Consider arbitrarily correlated* $Z_1, Z_2, \ldots, Z_N$*, where the marginal distribution of each* Z_i *is identically distributed over an* *alphabet of size $m \geq 2$. Let $\ell_{\mathrm{KT}}(z^N) = \log \frac{1}{q_{\mathrm{KT}}(z^N)}$ where $q_{\mathrm{KT}}(\cdot)$ is the marginal distribution induced by KT probability assignment in* (2.9)

$$q_{\mathrm{KT}}(z^N) = \frac{(2N_1-1)!!(2N_2-1)!!\cdots(2N_m-1)!!}{m(m+2)\cdots(m+2N-2)} \tag{2.17}$$

with $(-1)!! \triangleq 1$. *We then have*

$$\mathsf{E}[\ell_{\mathrm{KT}}(Z^N)] \leq \tfrac{m}{2}\log\left(e\left(1+\tfrac{2N}{m}\right)\right) + \tfrac{1}{2}\log(\pi N) + NH(Z_1). \tag{2.18}$$

We are now ready to prove Theorem 1.

Proof of Theorem 1. We will prove the universality of C_k for both KT probability assignment and Laplace probability assignment. Note that the upper bound on the expected length of KT in (2.18) is upper bounded by the upper bound on the length of Laplace in (2.16). So it suffices to show Laplace probability assignment is universal.

We use the bound in Theorem 4 to establish the upper bound on the length of the code. Recall that here we compress the diagonal blocks $\mathbf{B}_{\mathrm{d}}$ ($m = 2^{k^2}$-sized alphabet, $N = n'$ blocks) and the off-diagonal blocks $\mathbf{B}_{\mathrm{ut}}$ ($m = 2^{k^2}$-sized alphabet, $N = \binom{n'}{2}$ blocks) separately. We have,

$$\begin{aligned}
\frac{\mathsf{E}(\ell(C_k(A_n)))}{H(A_n)} &= \frac{\mathsf{E}(\ell_{\mathrm{L}}(\mathbf{B}_{\mathrm{ut}})) + \mathsf{E}(\ell_{\mathrm{L}}(\mathbf{B}_{\mathrm{d}}))}{H(A_n)} \\
&\leq \frac{\binom{n'}{2}H(\mathbf{B}_{12}) + 2^{k^2}\log\left(2e\binom{n'}{2}\right) + n'H(\mathbf{B}_{11}) + 2^{k^2}\log(2en')}{H(A_n)} \\
&\overset{(a)}{\leq} \frac{\binom{n'}{2}H(\mathbf{B}_{12}) + 2^{k^2}\log\left(en^2\right) + nH(\mathbf{B}_{11}) + 2^{k^2}\log(2en)}{H(A_n)} \\
&\overset{(b)}{\leq} \frac{\binom{n'}{2}H(\mathbf{B}_{12}) + 2^{k^2}\log\left(2e^2n^3\right) + nk^2H(A_{12})}{H(A_n)} \\
&= \frac{\binom{n'}{2}H(\mathbf{B}_{12})}{H(A_n)} + \frac{2^{k^2}\log\left(2e^2n^3\right)}{H(A_n)} + \frac{nk^2H(A_{12})}{H(A_n)},
\end{aligned} \tag{2.19}$$

where in (a) we bound $\binom{n'}{2} \leq n^2$ and $n' \leq n$, and in (b) we note that $H(\mathbf{B}_{11}) \leq k^2H(A_{12})$ since there are k^2-k elements of the matrix (all apart from the diagonal elements) are distributed identically as A_{12}. We will now analyze each of these three terms separately. Firstly, using Theorem 3 yields that $\frac{\binom{n'}{2}H(\mathbf{B}_{12})}{H(A_n)} \to 1$. Next, since $f(n) = \Omega\left(\frac{1}{n^{2-\varepsilon}}\right)$, we have $H(A_n) = \Omega(n^{\varepsilon}\log n)$ and subsequently substituting $k \le \sqrt{\delta \log n}$, we have

$$\frac{2^{k^2}\log(2en^3)}{H(A_n)} = O\left(\frac{n^{\delta}\log n}{n^{\varepsilon}\log n}\right) = O\left(n^{\delta-\varepsilon}\right) = o(1)$$

since $\delta < \varepsilon$. Moreover, we have

$$\frac{nk^2H(A_{12})}{H(A_n)} \le \frac{nk^2H(A_{12})}{H(A_n|X^n)} = \frac{nk^2H(A_{12})}{\binom{n}{2}H(A_{12}|X_1,X_2)} = O\left(\frac{k^2}{n}\right) = o(1),$$

where the penultimate equality used the fact that $H(A_{12}) \sim H(A_{12}|X_1,X_2)$ (since $h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}) \sim \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p}$). We have then established that

$$\begin{aligned}\frac{\mathsf{E}(\ell(C_k(A_n)))}{H(A_n)} &\le \frac{\binom{n'}{2}H(\mathbf{B}_{12})}{H(A_n)} + \frac{2^{k^2}\log\left(2en^3\right)}{H(A_n)} + \frac{nk^2H(A_{12})}{H(A_n)} \\ &= 1 + o(1),\end{aligned}$$

which finishes the proof. □

The proof of Theorem 2 follows similar arguments as in Theorem 1 and is deferred to Section 2.4.5.

2.4 Proof of Universality

2.4.1 Graph Entropy

Proof of Lemma 1. Note that

$$H(A_n) = H(A_n|X^n) + I(X^n;A_n)$$

$$= \binom{n}{2} H(A_{12}|X_1, X_2) + I(X^n; A_n) \tag{2.20}$$

$$= \binom{n}{2} \mathbf{p}^T h\big(f(n)\mathbf{Q}\big)\mathbf{p} + I(X^n; A_n), \tag{2.21}$$

where (2.21) follows since all the $\binom{n}{2}$ edges are identically distributed and also independent given X^n and consequently

$$\begin{aligned} H(A_n|X^n) &= \binom{n}{2} H(A_{12}|X_1, X_2) \\ &= \binom{n}{2} \sum_{i,j} H(A_{12}|X_1 = i, X_2 = j) p_i p_j \\ &= \binom{n}{2} \mathbf{p}^T h\left(f(n)\mathbf{Q}\right)\mathbf{p}. \end{aligned}$$

When $f(n) = \Theta(1)$, we see that since

$$0 \leq I(X^n; A_n) \leq H(X^n) = nH(X_1) \leq n \log L,$$

we have that $H(A_n) = \binom{n}{2}\mathbf{p}^T h\big(f(n)\mathbf{Q}\big)\mathbf{p} + o\left(n^2 h(f(n))\right)$.

Next, consider the case when $f(n) = o(1)$ and $f(n) = \Omega\left(\frac{1}{n^2}\right)$. By properties of the entropy, we have

$$H(A_n|X^n) \leq H(A_n) \leq \binom{n}{2} H(A_{12}). \tag{2.22}$$

Note that

$$\mathsf{P}(A_{12} = 1) = \sum_{i,j} \mathsf{P}(A_{12} = 1|X_1 = i, X_2 = j) p_i p_j = \mathbf{p}^T f(n)\mathbf{Q}\mathbf{p},$$

which yields that $H(A_{12}) = h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right)$. Substituting this in (2.22) gives

$$\binom{n}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} \leq H(A_n) \leq \binom{n}{2} h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right). \tag{2.23}$$

Note now for any $g(n) = o(1)$, we have

$$\begin{aligned} h(g(n)) &= -g(n)\log g(n) - (1-g(n))\log(1-g(n)) \\ &= -g(n)\log g(n)\left(1 + \frac{(1-g(n))\log(1-g(n))}{g(n)\log g(n)}\right). \end{aligned}$$

By noting that $\frac{\log(1-g(n))}{g(n)} \to -1$ and $\frac{1}{\log(g(n))} \to 0$ as $g(n) \to 0$ we see that

$$h(g(n)) = g(n)\log\frac{1}{g(n)}(1+o(1)).$$

Using this, we note that

$$\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} = \mathbf{p}^T\mathbf{Q}\mathbf{p}f(n)\log\frac{1}{f(n)}(1+o(1))$$

and

$$h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}) = \mathbf{p}^T\mathbf{Q}\mathbf{p}f(n)\log\frac{1}{f(n)}(1+o(1)).$$

Finally, substituting this into (2.23) yields

$$H(A_n) = \binom{n}{2}\mathbf{p}^T\mathbf{Q}\mathbf{p}f(n)\log\frac{1}{f(n)}(1+o(1))$$

as required. □

2.4.2 Asymptotic i.i.d. via Block Decomposition

We first invoke a known property of stochastic block models (see, for example, [66]). We include the proof here for completeness.

Lemma 2 (Exchangeability of SBM). *Let $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W})$. For a permutation $\pi : [n] \to [n]$, let $\pi(A_n)$ be an $n \times n$ matrix whose (i, j) entry is given by $A_{\pi(i),\pi(j)}$. Then, for any permutation*

$\pi : [n] \to [n]$, the joint distribution of A_n is the same as the joint distribution of $\pi(A_n)$, i.e.,

$$A_n \overset{d}{=} \pi(A_n). \tag{2.24}$$

Proof. Let a_n be a realization of the random matrix A_n and $\pi(X^n)$ be the permuted vector $(X_{\pi(1)}, \dots, X_{\pi(n)})$. For any symmetric binary matrix a_n with zero diagonal entries, we have

$$\begin{aligned}
\mathsf{P}(A_n = a_n) &= \sum_{x^n \in [L]^n} \mathsf{P}(A_n = a_n, X^n = x^n) \\
&= \sum_{x^n \in [L]^n} \mathsf{P}(A_n = a_n | X^n = x^n) \prod_{i=1}^{n} \mathsf{P}(X_i = x_i) \\
&\overset{(a)}{=} \sum_{x^n \in [L]^n} \prod_{\substack{i,j \\ 1 \le i < j \le n}} \mathsf{P}(A_{ij} = a_{ij} | X_i = x_i, X_j = x_j) \prod_{i=1}^{n} \mathsf{P}(X_{\pi(i)} = x_i) \\
&\overset{(b)}{=} \sum_{x^n \in [L]^n} \prod_{\substack{i,j \\ 1 \le i < j \le n}} (W_{x_i,x_j})^{a_{ij}} (1 - W_{x_i,x_j})^{1-a_{ij}} \prod_{i=1}^{n} \mathsf{P}(X_{\pi(i)} = x_i) \\
&\overset{(c)}{=} \sum_{x^n \in [L]^n} \prod_{\substack{i,j \\ 1 \le i < j \le n}} \mathsf{P}(A_{\pi(i),\pi(j)} = a_{ij} | X_{\pi(i)} = x_i, X_{\pi(j)} = x_j) \prod_{i=1}^{n} \mathsf{P}(X_{\pi(i)} = x_i) \\
&= \sum_{x^n \in [L]^n} \mathsf{P}(\pi(A_n) = a_n, \pi(X^n) = x^n) \\
&= \mathsf{P}(\pi(A_n) = a_n),
\end{aligned}$$

where (a) follows since X^n are i.i.d. and thus $\mathsf{P}(X_i = x_i) = \mathsf{P}(X_{\pi(i)} = x_i)$ and (b) follows since $A_{ij} \sim \text{Bern}(W_{X_i,X_j})$, and thus

$$\begin{aligned}
\mathsf{P}(A_{ij} = a_{ij} | X_i = x_i, X_j = x_j) &= \begin{cases} W_{x_i,x_j} & \text{if } a_{ij} = 1 \\ 1 - W_{x_i,x_j} & \text{if } a_{ij} = 0 \end{cases} && (2.25) \\
&= (W_{x_i,x_j})^{a_{ij}} (1 - W_{x_i,x_j})^{1-a_{ij}}. && (2.26)
\end{aligned}$$

The step in (c) follows since $A_{\pi(i),\pi(j)} \sim \text{Bern}(W_{X_{\pi(i)},X_{\pi(j)}})$ and the conditional probability has the same expression as in (2.26). □

Now we are ready to establish Theorem 3.

Proof of Theorem 3. For any $i_1 \neq j_1$ and $i_2 \neq j_2$, consider a permutation $\pi_1 : [n] \to [n]$ that has

$$\pi_1(x) = \begin{cases} x + (i_2 - i_1)k & \text{for } (i_1 - 1)k + 1 \leq x \leq i_1 k \\ x + (j_2 - j_1)k & \text{for } (j_1 - 1)k + 1 \leq x \leq j_1 k \end{cases}$$

and the remaining $n - 2k$ arguments are mapped to the $n - 2k$ values in $[n] \setminus \{(i_2 - 1)k + 1, \ldots, i_2 k, (j_2 - 1)k, \ldots, j_2 k\}$ in any order. Lemma 2 implies that $\mathbf{B}_{i_1, j_1}$, which is the submatrix formed by the rows $(i_1 - 1)k + 1, \ldots, i_1 k$ and the columns $(j_1 - 1)k + 1, \ldots, j_1 k$ has the same distribution as the submatrix formed by the rows $\pi_1((i_1 - 1)k + 1), \ldots, \pi_1(i_1 k)$ and the columns $\pi_1((j_1 - 1)k + 1), \ldots, \pi_1(j_1 k)$. From the definition of π_1, we see that the latter submatrix is $\mathbf{B}_{i_2, j_2}$ and we establish that $\mathbf{B}_{i_1, j_1} \overset{d}{=} \mathbf{B}_{i_2, j_2}$. Similarly, defining a permutation $\pi_2 : [n] \to [n]$ which has

$$\pi_2(x) = \; x + (l_2 - l_1)k \quad \text{for } (l_1 - 1)k + 1 \leq x \leq l_1 k$$

and invoking Lemma 2 establishes $\mathbf{B}_{l_1, l_1} \overset{d}{=} \mathbf{B}_{l_2, l_2}$.

Now, clearly $H(\mathbf{B}_{\text{ut}}) \leq \binom{n'}{2} H(\mathbf{B}_{12})$, and therefore we have

$$\limsup_{n \to \infty} \frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} \leq 1. \tag{2.27}$$

Moreover we have $H(A_n) = H(\mathbf{B}_{\text{ut}}, \mathbf{B}_{\text{d}}) \leq H(\mathbf{B}_{\text{ut}}) + H(\mathbf{B}_{\text{d}}) \leq H(\mathbf{B}_{\text{ut}}) + n' H(\mathbf{B}_{11}) \leq H(\mathbf{B}_{\text{ut}}) + n' k^2 h(A_{12})$ where the last inequality follows by noting that except for the diagonal elements of $\mathbf{B}_{\text{d}}$ (which are zero and thus have zero entropy), all other elements have the same distribution as A_{12}. We therefore obtain $H(\mathbf{B}_{\text{ut}}) \geq H(A_n) - n' k^2 h(A_{12}) = H(A_n) - nkh(A_{12}) \geq H(A_n | X_1^n) -$

$nkh(A_{12}) = \binom{n}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} - nkh(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})$. Consequently,

$$\frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2}H(\mathbf{B}_{12})} \geq \frac{\binom{n}{2}\left(\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} - \frac{2kh(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})}{n-1}\right)}{\binom{n'}{2}H(\mathbf{B}_{12})}. \tag{2.28}$$

We will now analyze the right hand side of (2.28) in two parameter regimes.

- $\underline{f(n) = 1:}$ We have

$$\begin{aligned} H(\mathbf{B}_{12}) &\overset{(a)}{\leq} H(\mathbf{B}_{12}|X_1^{2k}) + H(X_1^{2k}) \\ &\leq H(\mathbf{B}_{12}|X_1^{2k}) + 2kH(\mathbf{p}) \\ &\overset{(b)}{=} k^2 H(A_{1,k}|X_1, X_k) + 2kH(\mathbf{p}) \\ &\leq k^2\left(\mathbf{p}^T h(\mathbf{Q})\mathbf{p} + 2\frac{\log L}{k}\right), \end{aligned} \tag{2.29}$$

where (a) follows from the chain rule and (b) follows since all elements of the matrix $\mathbf{B}_{12}$ are independent given $X_1, \cdots, X_{2k}$. Plugging this into the right hand side of (2.28) we obtain

$$\frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2}H(\mathbf{B}_{12})} \geq \frac{\binom{n}{2}\left(\mathbf{p}^T h(\mathbf{Q})\mathbf{p} - \frac{2kh(\mathbf{p}^T\mathbf{Q}\mathbf{p})}{n-1}\right)}{\binom{n'}{2}k^2\left(\mathbf{p}^T h(\mathbf{Q})\mathbf{p} + 2\frac{\log L}{k}\right)}. \tag{2.30}$$

Since $k = o(n), k = \omega(1)$ and $\binom{n'}{2}k^2 \sim \binom{n}{2}$, we have from (2.30)

$$\liminf_{n\to\infty} \frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2}H(\mathbf{B}_{12})} \geq 1, \tag{2.31}$$

which together with (2.27) yields the required result.

- $\underline{f(n) = \Omega\left(\frac{1}{n^2}\right), f(n) = o(1):}$ Since $\mathbf{B}_{12}$ is a matrix of k^2 identically distributed Bernoulli random variables, we have

$$H(\mathbf{B}_{12}) \leq k^2 h(A_{1,k}) = k^2 h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right). \tag{2.32}$$

Plugging this into the RHS of (2.28) then yields

$$\frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2}H(\mathbf{B}_{12})} \geq \frac{\binom{n}{2}\left(\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} - \frac{2kh(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})}{n-1}\right)}{\binom{n'}{2}k^2 h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right)}. \tag{2.33}$$

We first observe that in this parameter range, since $f(n) = o(1)$, we have by Lemma 1

$$\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} \sim h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right). \tag{2.34}$$

Finally using that $k = o(n)$ and $\binom{n'}{2}k^2 \sim \binom{n}{2}$ establishes

$$\liminf_{n\to\infty} \frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2}H(\mathbf{B}_{12})} \geq 1, \tag{2.35}$$

which together with (2.27) yields the required result.

□

2.4.3 Length of the Laplace Probability Assignment

Proof of Theorem 4. Let us first elaborate the relation between probability assignment and compression length. In Algorithm 1, the terms $\log(q(x_{j+1}|x^j))$ are added up, which lead to the marginal probability implied by the sequential probability assignment

$$\sum_{j=0}^{N-1}\log(q(x_{j+1}|x^j)) = \log\left(\prod_{j=0}^{N-1} q(x_{j+1}|x^j)\right) = \log(q(x^N)). \tag{2.36}$$

The compression output length of Algorithm 1 is $\left\lceil \log\frac{1}{q(x^N)} \right\rceil + 1$.

Now we analyze the compression length of Laplace compressor for the sequence $Z_1, Z_2, \dots, Z_N$. Define $\theta_i \coloneqq \mathsf{P}(Z_1 = i), N_i \coloneqq \sum_{k=1}^{N} \mathbb{1}\{Z_k = i\}, i \in [m]$. We have

$$\begin{aligned}
\ell_{\mathrm{L}}(z^n) &= \log \frac{1}{q_{\mathrm{L}}(z^N)} \\
&= \log \frac{\theta_1^{N_1} \theta_2^{N_2} \cdots \theta_m^{N_m}}{q_{\mathrm{L}}(z^N)} + \log \frac{1}{\theta_1^{N_1} \theta_2^{N_2} \cdots \theta_m^{N_m}} \\
&= \log \binom{N+m-1}{m-1} + \log \left(\frac{N!}{N_1! N_2! \cdots N_m!} \theta_1^{N_1} \theta_2^{N_2} \cdots \theta_m^{N_m} \right) \\
&\qquad\qquad\qquad\qquad + \log \frac{1}{\theta_1^{N_1} \theta_2^{N_2} \cdots \theta_m^{N_m}} \\
&\overset{(a)}{\leq} \log \binom{N+m-1}{m-1} + \log \frac{1}{\theta_1^{N_1} \theta_2^{N_2} \cdots \theta_m^{N_m}} \\
&\overset{(b)}{\leq} (m-1) \log \left(e \left(\frac{N}{m-1} + 1 \right) \right) + \log \frac{1}{\theta_1^{N_1} \theta_2^{N_2} \cdots \theta_m^{N_m}} \\
&\leq m \log(2eN) + \sum_{i=1}^{m} N_i \log \frac{1}{\theta_i},
\end{aligned} \tag{2.37}$$

where (a) follows since $\frac{N!}{N_1! N_2! \cdots N_m!} \theta_1^{N_1} \theta_2^{N_2} \cdots \theta_m^{N_m}$ is a multinomial probability which is always upper bounded by 1, and (b) follows since $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$. Taking expectation on both sides of (2.37), we obtain

$$\begin{aligned}
\mathsf{E}[\ell_{\mathrm{L}}(Z^N)] &\leq m \log(2eN) + \sum_{i=1}^{m} \mathsf{E}[N_i] \log \frac{1}{\theta_i} \\
&\overset{(a)}{=} m \log(2eN) + \sum_{i=1}^{m} N \theta_i \log \frac{1}{\theta_i} \\
&= m \log(2eN) + N H(Z_1),
\end{aligned}$$

where (a) follows since $\mathsf{E}[N_i] = \sum_{k=1}^{N} \mathsf{E}[\mathbb{1}\{Z_k = i\}] = N \mathsf{P}(Z_1 = i)$ since the Z_i are identically distributed. □

2.4.4 Length of the KT probability assignment

Lemma 3. *For any integer $m > 0$, $N_1, N_2, \cdots N_m \in \mathbb{N}$ and probability distribution $(\theta_1, \cdots \theta_m)$,*

$$\frac{\binom{N}{N_1,N_2\cdots N_m}\theta_1^{N_1}\cdots\theta_m^{N_m}}{\binom{2N}{2N_1,2N_2\cdots 2N_m}\theta_1^{2N_1}\cdots\theta_m^{2N_m}} \geq 1,$$

where $N = \sum_{i=1}^m N_i$.

Remark 4. Equivalently, consider an urn containing known number of balls with m different colours. The lemma claims that the probability of getting N_1 balls of colour 1, N_2 of balls of colour 2, $\cdots$ N_m balls of colour m out of N draws with replacement is always greater than the probability of getting $2N_1$ balls of colour 1, $2N_2$ of balls of colour 2, $\cdots$ $2N_m$ balls of colour m out of $2N$ draws with replacement.

Proof. Let $p_1 = N_1/N, p_2 = N_2/N, \cdots, p_m = N_m/N$. Notice that $\sum_{i=1}^m p_i = 1$, so $(p_1, \cdots p_m)$ can be viewed as a probability distribution. And the entropy of this distribution is $H(p_1, \cdots p_m) = \sum_{i=1}^m -p_i \log p_i$. Firstly we consider the case when $N_1, N_2 \cdots N_m$ are all positive and none of them equal to N. By Stirling's approximation for factorial $\sqrt{2\pi n}(\frac{n}{e})^n e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}(\frac{n}{e})^n e^{1/12n}$, we can bound

$$\begin{aligned}\binom{N}{N_1, N_2 \cdots N_m} &\geq \frac{\sqrt{2\pi N}N^N \exp\left(\frac{1}{12N+1} - \frac{1}{12N_1} - \frac{1}{12N_2} - \cdots - \frac{1}{12N_m}\right)}{(2\pi)^{m/2}(N_1 N_2 \cdots N_m)^{1/2} N_1^{N_1} N_2^{N_2} \cdots N_m^{N_m}} \\ &= \frac{\exp\left(\frac{1}{12N+1} - \frac{1}{12N_1} - \frac{1}{12N_2} - \cdots - \frac{1}{12N_m}\right)}{(2\pi)^{\frac{m-1}{2}}(p_1 p_2 \cdots p_m)^{1/2} N^{\frac{m-1}{2}} 2^{-NH(p_1, p_2, \cdots, p_m)}}.\end{aligned}$$

Similarly, we have

$$\binom{2N}{2N_1, 2N_2 \cdots 2N_m} \leq \frac{\exp\left(\frac{1}{24N} - \frac{1}{24N_1+1} - \frac{1}{24N_2+1} - \cdots - \frac{1}{24N_m+1}\right)}{(2\pi)^{\frac{m-1}{2}} 2^{\frac{m-1}{2}} (p_1 p_2 \cdots p_m)^{1/2} N^{\frac{m-1}{2}} 2^{-2NH(p_1 \cdots p_m)}}.$$

Consider the function

$$\begin{aligned}
&f(N_1,N_2,\cdots,N_m)\\
&\quad = \tfrac{1}{12N+1} - \tfrac{1}{24N} + (\tfrac{1}{24N_1+1} - \tfrac{1}{12N_1}) + (\tfrac{1}{24N_2+1} - \tfrac{1}{12N_2}) + \cdots + (\tfrac{1}{24N_m+1} - \tfrac{1}{12N_m})
\end{aligned}$$

and the function

$$g(n) = \frac{1}{24n+1} - \frac{1}{12n},$$

where n is a positive integer. Function $g(n)$ is minimized with $n = 1$ and $\min g(n) = 1/25 - 1/12$ and we can bound function $f(N_1,N_2,\cdots,N_m) \geq \frac{1}{12N+1} - \frac{1}{24N} + (1/25-1/12)m$. Finally we are ready to prove the lemma.

$$\begin{aligned}
\frac{\binom{N}{N_1,N_2\cdots N_m}\theta_1^{N_1}\cdots\theta_m^{N_m}}{\binom{2N}{2N_1,2N_2\cdots 2N_m}\theta_1^{2N_1}\cdots\theta_m^{2N_m}} &\geq \frac{2^{\frac{m-1}{2}}\exp(f(N_1,N_2,\cdots,N_m))}{2^{NH(p_1\cdots p_m)}\theta_1^{N_1}\cdots\theta_m^{N_m}}\\
&\geq \frac{2^{\frac{m-1}{2}}\exp\left(\frac{1}{12N+1} - \frac{1}{24N} + (1/25-1/12)m\right)}{2^{-ND_{\mathrm{KL}}(p||\theta)}}\\
&= 2^{\frac{m-1}{2}}2^{ND_{\mathrm{KL}}(p||\theta)}2^{\log e(\frac{1}{12N+1}-\frac{1}{24N}+(1/25-1/12)m)}.
\end{aligned}$$

Notice that $\frac{1}{12N+1} - \frac{1}{24N}$ goes to zero when $N \to \infty$, $\frac{m-1}{2} > (1/25-1/12)m$ and $D_{\mathrm{KL}}(P||\theta) \geq 0$. Therefore in this case,

$$\frac{\binom{N}{N_1,N_2\cdots N_m}\theta_1^{N_1}\cdots\theta_m^{N_m}}{\binom{2N}{2N_1,2N_2\cdots 2N_m}\theta_1^{2N_1}\cdots\theta_m^{2N_m}} \geq 1.$$

When one of $\{N_i\}_{i=1}^{N}$ equals to N, without loss of generality, we assume that $N_1 = N$. We have

$$\frac{\binom{N}{N_1,N_2\cdots N_m}\theta_1^{N_1}\cdots\theta_m^{N_m}}{\binom{2N}{2N_1,2N_2\cdots 2N_m}\theta_1^{2N_1}\cdots\theta_m^{2N_m}} = \frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} > 1.$$

When there are k numbers out of $N_1,N_2,\cdots,N_m$ that equal to zero, we can simply remove these values and consider the case with alphabet size $m-k$. And this will yield the same result. □

Proof of Theorem 5. In this proof, we define a generalized form of factorial function. Let x be a positive integer, $(x+\frac{1}{2})! = \frac{1}{2}\frac{3}{2}\cdots(x+\frac{1}{2})$. Since $(2N_1-1)!! = \frac{(2N_1)!}{2^{N_1}(N_1)!}$, we have

$$\begin{aligned}
&m(m+2)\cdots(m+2N-2)\\
&\quad = 2^N\left(\frac{m}{2}\right)\left(\frac{m+2}{2}\right)\cdots\left(\frac{m+2N-2}{2}\right)\\
&\quad = 2^N\frac{\left(\frac{m}{2}+N-1\right)!}{\left(\frac{m}{2}-1\right)!}.
\end{aligned}$$

Therefore we can rewrite the KT probability assignment in (2.17) as

$$\begin{aligned}
q_{\mathrm{KT}}(z^N) &= \frac{(\frac{m}{2}-1)!}{2^N(\frac{m}{2}+N-1)!}\frac{\binom{2N}{N}}{\binom{2N}{N}}\prod_{i=1}^{m}\frac{(2N_i)!}{N_i!2^{N_i}}\\
&= \frac{(\frac{m}{2}-1)!}{2^N(\frac{m}{2}+N-1)!}\binom{2N}{N}N!\frac{N!}{(2N)!}\prod_{i=1}^{m}\frac{(2N_i)!}{N_i!2^{N_i}}\\
&\overset{(a)}{\geq} \frac{\left(\frac{m}{2}-1\right)!\binom{2N}{N}}{4^N(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}\frac{N!}{(2N)!}\prod_{i=1}^{m}\frac{(2N_i)!}{N_i!}\\
&\overset{(b)}{=} \frac{\theta_1^{N_1}\cdots\theta_m^{N_m}(\frac{m}{2}-1)!\binom{2N}{N}}{4^N(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}\frac{\binom{N}{N_1,N_2\cdots N_m}\theta_1^{N_1}\cdots\theta_m^{N_m}}{\binom{2N}{2N_1,2N_2\cdots 2N_m}\theta_1^{2N_1}\cdots\theta_m^{2N_m}},
\end{aligned}$$

where (a) follows that when m is even, $\frac{N!}{(\frac{m}{2}+N-1)!} = \frac{1}{(N+1)\cdots(\frac{m}{2}+N-1)} \geq \frac{1}{(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}$ and when m is odd, $\frac{N!}{(\frac{m}{2}+N-1)!} \geq \frac{N!}{(\frac{m}{2}+N-\frac{1}{2})!} = \frac{1}{(N+1)\cdots(\frac{m}{2}+N-\frac{1}{2})} \geq \frac{1}{(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}$, (b) follows that $\binom{N}{N_1,N_2\cdots N_m} = \frac{N!}{\prod_{i=1}^{m}N_i!}$ and $\theta_i \triangleq \mathbb{P}(Z_1=i)$. By lemma 3, we have $q_{\mathrm{KT}}(z^N) \geq \frac{\theta_1^{N_1}\cdots\theta_m^{N_m}(\frac{m}{2}-1)!\binom{2N}{N}}{4^N(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}$. Thus,

$$\begin{aligned}
&\ell_{\mathrm{KT}}(z^N)\\
&= \log\frac{1}{q_{\mathrm{KT}}(z^N)}\\
&\leq \log\frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} + \log\frac{4^N(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}{(\frac{m}{2}-1)!\binom{2N}{N}}\\
&= \log\frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} + \frac{m-1}{2}\log\left(N+\frac{m-1}{2}\right) + \log\frac{4^N}{\binom{2N}{N}} - \log\left(\frac{m}{2}-1\right)!\\
&\overset{(a)}{\leq} \log\frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} + \frac{m-1}{2}\log\left(N+\frac{m-1}{2}\right) + \log\frac{4^N}{\binom{2N}{N}} - \left(\frac{m}{2}-1\right)\log\left(\frac{\frac{m}{2}-1}{e}\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\sim} \log \frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} + \frac{m-1}{2}\log\left(N+\frac{m-1}{2}\right) + \log\sqrt{\pi N} - \left(\frac{m}{2}-1\right)\log\left(\frac{\frac{m}{2}-1}{e}\right)\\
&\sim \frac{m}{2}\log\frac{e(\frac{m}{2}+N)}{m/2} + \log\sqrt{\pi N} + \log\frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}}\\
&= \frac{m}{2}\log\left(e\left(1+\tfrac{2N}{m}\right)\right) + \frac{1}{2}\log(\pi N) + \sum_{i=1}^{m} N_i \log\frac{1}{\theta_i},
\end{aligned}
$$

where (a) follows Stirling's approximation $k! \geq \sqrt{2\pi k}(\frac{k}{e})^k e^{\frac{1}{12k+1}}$ and (b) follows Stirling's approximation for binomial coefficient, i.e., $\binom{2N}{N} \sim \frac{4^N}{\sqrt{\pi N}}$. Therefore, we have

$$
\mathsf{E}[\ell_{\mathrm{KT}}(Z^N)] \leq \frac{1}{2}m\log\left(e\left(1+\tfrac{2N}{m}\right)\right) + \frac{1}{2}\log(\pi N) + NH(Z_1).
$$

□

2.4.5 Proof of Theorem 2

Proof. Once again, we establish universality for both KT and Laplace probability assignment. Following a similar argument as in the proof of Theorem 1, it suffices to show the universality of Laplace. Since we are compressing $N = \binom{n}{2}$ identically distributed bits using a Laplace probability assignment, Theorem 4 yields

$$
\begin{aligned}
\frac{\mathsf{E}(\ell(C_1(A_n)))}{H(A_n)} &\leq \frac{\log(2eN) + NH\left(A_{12}\right)}{H(A_n)}\\
&\leq \frac{\log(2eN) + NH\left(A_{12}\right)}{H(A_n|X_1^n)}\\
&= \left(\frac{\log(2eN) + NH\left(A_{12}\right)}{NH(A_{12})}\right)\frac{H(A_{12})}{H(A_{12}|X_1,X_2)}\\
&= \left(1+\frac{\log(2eN)}{Nh(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})}\right)\frac{h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})}{\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p}}\\
&\overset{(a)}{=} 1+o(1).
\end{aligned}
$$

Here, (a) is justified by noting that $\frac{\log(2eN)}{Nh(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})} \leq \frac{\log(2en^2)}{\binom{n}{2}h(n^{-(2-\varepsilon)}\mathbf{p}^T\mathbf{Q}\mathbf{p})}\frac{h(n^{-(2-\varepsilon)}\mathbf{p}^T\mathbf{Q}\mathbf{p})}{h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})}$, and then noting that $\frac{\log(2en^2)}{\binom{n}{2}h(n^{-(2-\varepsilon)}\mathbf{p}^T\mathbf{Q}\mathbf{p})} = o(1)$ and $\frac{h(n^{-(2-\varepsilon)}\mathbf{p}^T\mathbf{Q}\mathbf{p})}{h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})} = O(1)$ when $f(n) = \Omega\left(\frac{1}{n^{2-\varepsilon}}\right)$ and that

$H(h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}) \sim \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p}.$ □

Remark 5. When $f(n) = 1$, the compressor C_1 is strictly suboptimal. This is because the length achieved by C_1 is $\binom{n}{2} h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right)(1+o(1))$, whereas the first order term in the entropy is $\binom{n}{2}\mathbf{p}^T h\left(f(n)\mathbf{Q}\right)\mathbf{p}^T$. When $f(n)$ is $o(1)$, these two have the same first order term. However, when $f(n)$ is constant, $\mathbf{p}^T h\left(f(n)\mathbf{Q}\right)\mathbf{p}^T$ is strictly smaller than $h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right)$ by concavity of entropy.

2.5 Redundancy analysis

Let A_n be a random graph generated from certain graph generation model and let C be a graph compressor. We define the *redundancy* of compressor C for random graph A_n as

$$R(C_k, A_n) \triangleq \mathsf{E}[\ell(C_k(A_n))] - H(A_n).$$

Theorem 6 (Redundancy in the vanishing probability regime)**.** *Let*

$$A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})$$

with $f(n) = o(1)$*,* $f(n) = \Omega(\frac{1}{n^{2-\varepsilon}})$ *for any* $0 < \varepsilon < 1$*, and* $L = \Theta(1)$*. If* $k = \omega(1)$ *and* $k \le \sqrt{\delta \log n}$ *for some* $0 < \delta < \varepsilon$*, then the redundancy of compressor* C_k *defined in Section 2.2 is upper bounded as*

$$\begin{aligned} R(C_k, A_n) &\triangleq \mathsf{E}[\ell(C_k(A_n))] - H(A_n) \\ &\le \binom{n}{2} f(n) \left(\mathbf{p}^T\mathbf{Q}\mathbf{p} \log\left(\frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}}\right) - \mathbf{p}^T\mathbf{Q}^*\mathbf{p}\right) + o(n^2 f(n)), \end{aligned}$$

where $\mathbf{Q}^*$ *denotes an* $L \times L$ *matrix whose* (i,j) *entry is* $Q_{ij}\log(\frac{1}{Q_{ij}})$*.*

Proof. First we lower bound $H(A_n)$ using the conditional entropy $H(A_n|X^n)$:

$$
\begin{aligned}
&H(A_n)\\
&\geq H(A_n|X^n)\\
&= \binom{n}{2}\mathbf{p^T}h(f(n)\mathbf{Q})\mathbf{p}\\
&= \binom{n}{2}\sum_{i,j\in[n]} p_ip_jh(f(n)Q_{ij})\\
&\overset{(a)}{=} \binom{n}{2}\sum_{i,j\in[n]} p_ip_j\left(f(n)Q_{ij}\log\left(\frac{1}{f(n)Q_{ij}}\right)+f(n)Q_{ij}\log e+o(f(n))\right)\\
&= \binom{n}{2}\sum_{i,j\in[n]} p_ip_j\Bigg(f(n)Q_{ij}\log\left(\frac{1}{f(n)}\right)\\
&\qquad\qquad\qquad\qquad +f(n)Q_{ij}\log e+f(n)Q_{ij}\log\left(\frac{1}{Q_{ij}}\right)+o(f(n))\Bigg)\\
&= \binom{n}{2}f(n)\left(\log\left(\frac{1}{f(n)}\right)\mathbf{p}^T\mathbf{Q}\mathbf{p}+\mathbf{p}^T\mathbf{Q}\mathbf{p}\log e+\mathbf{p^T}\mathbf{Q}^*\mathbf{p}+o(1)\right), \qquad (2.38)
\end{aligned}
$$

where (a) follows since $h(g(n)) = g(n)\log\frac{1}{g(n)}+g(n)\log e+o(g(n))$ (see, for example, [67]). From (2.19) in the proof of Theorem 1, we have

$$
\mathsf{E}(\ell(C_k(A_n))) \leq \binom{n'}{2}H(\mathbf{B}_{12})+2^{k^2}\log(2e^2n^3)+nk^2H(A_{12}).
$$

Now, we upper bound the three terms separately. We have

$$
\begin{aligned}
&\binom{n'}{2}H(\mathbf{B}_{12})\\
&\quad\leq \binom{n'}{2}k^2H(A_{12})\\
&\quad= \binom{n'}{2}k^2h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})\\
&\quad\overset{(b)}{=} \binom{n}{2}f(n)\left(\log\frac{1}{f(n)}\mathbf{p}^T\mathbf{Q}\mathbf{p}+\mathbf{p}^T\mathbf{Q}\mathbf{p}\log e+\mathbf{p}^T\mathbf{Q}\mathbf{p}\log\frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}}+o(1)\right), \qquad (2.39)
\end{aligned}
$$

where (b) follows for the same reason as (a). Moreover, we have

$$2^{k^2}\log(2e^2n^3) \le 2^{\delta\log n}\log(2e^2n^3) = n^{\delta}\log(2e^2n^3) = o(n^2 f(n)), \tag{2.40}$$

and

$$nk^2 H(A_{12}) \le (n\delta\log n)\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\log\frac{e}{f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}} + o(f(n))\right) = o(n^2 f(n)). \tag{2.41}$$

Combining bounds (2.38), (2.39), (2.40) and (2.41), we have

$$R(C_k, A_n) \le \binom{n}{2} f(n)\left(\mathbf{p}^T\mathbf{Q}\mathbf{p}\log\frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} - \mathbf{p}^T\mathbf{Q}^*\mathbf{p}\right) + o(n^2 f(n)).$$

□

Theorem 7 (Redundancy in constant probability regime)**.** *Let*

$$A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})$$

with $f(n) = \Theta(1)$ and $L = \Theta(1)$. If $k = \omega(1)$ and $k \le \sqrt{\delta\log n}$ for some $0 < \delta < 1$, then the redundancy of compressor C_k defined in Section 2.2 is upper bounded as

$$R(C_k, A_n) \triangleq \mathsf{E}(\ell(C_k(A_n))) - H(A_n) \le H(\mathbf{p})\frac{n^2}{k} + o\left(\frac{n^2}{k}\right),$$

where $H(\mathbf{p}) = \sum_{i=1}^{L} p_i \log\left(\frac{1}{p_i}\right)$.

Proof. Firstly, we lower bound $H(A_n)$ by the conditional entropy $H(A_n|X^n)$

$$H(A_n) \ge H(A_n|X^n) = \binom{n}{2}\mathbf{p}^{\mathbf{T}} h(f(n)\mathbf{Q})\mathbf{p}. \tag{2.42}$$

Still we can upper bound the expected length of compressor C_k:

$$\mathsf{E}(\ell(C_k(A_n))) \le \binom{n'}{2} H(\mathbf{B}_{12}) + 2^{k^2}\log(2e^2n^3) + nk^2H(A_{12}).$$

Now, we bound the three terms separately. We have

$$\begin{aligned}
\binom{n'}{2} H(\mathbf{B}_{12}) &= \binom{n'}{2}\left(H(\mathbf{B}_{12}|X_1^{2k}) + I(X_1^{2k};\mathbf{B}_{12})\right)\\
&= \binom{n'}{2}\left(k^2H(A_{1,k+1}|X_1,X_{k+1}) + I(X_1^{2k};\mathbf{B}_{12})\right)\\
&\le \binom{n'}{2}\left(k^2\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + H(X_1^{2k})\right)\\
&\le \binom{n'}{2}\left(k^2\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + 2kH(\mathbf{p})\right)\\
&= \frac{n(n-k)}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + n(n'-1)H(\mathbf{p}),
\end{aligned} \tag{2.43}$$

$$2^{k^2}\log(2e^2n^3) \le n^{\delta}\log(2e^2n^3) = o\left(\frac{n^2}{k}\right), \tag{2.44}$$

and

$$nk^2H(A_{12}) \le n\delta\log nH(A_{12}) = o\left(\frac{n^2}{k}\right). \tag{2.45}$$

Combining bounds (2.42), (2.43), (2.44) and (2.45) gives

$$\begin{aligned}
&R(C_k,A_n)\\
&\le \frac{n(n-k)}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + n(n'-1)H(\mathbf{p}) - \frac{n(n-1)}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + o\left(\frac{n^2}{k}\right)\\
&= n(n'-1)H(\mathbf{p}) + \frac{n(1-k)}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + o\left(\frac{n^2}{k}\right)\\
&\le H(\mathbf{p})\frac{n^2}{k} + o\left(\frac{n^2}{k}\right).
\end{aligned}$$

□

2.6 Second order analysis in the sparse regime

So far, we have shown that our algorithm always matches the first order term in the Shannon entropy. Now, we proceed to analyze the second order term of the expected length of our proposed compressor. We focus on the family of *symmetric* SBM with edge probability $f(n) = 1/n$ and evaluate the performance of our compressor using the framework of local weak convergence, as introduced in [59]. This would allow us to compare the performance of our compressor to the compressor proposed in [50]. We first introduce some basic definitions on rooted graphs in Subsection 2.6.1. Then, we define the local weak convergence of graphs and derive the local weak convergence limit of the symmetric stochastic block model in Subsection 2.6.2. Finally, we review the definition of BC entropy in Subsection 2.6.3 and state the performance guarantee of our compression algorithm in Subsection 2.6.4.

2.6.1 Basic definitions on rooted graphs

Let $G = (V,E)$ be a simple graph (undirected, unweighted, no self-loop), with V a countable set of vertices and E a countable set of edges. Let $u \overset{G}{\sim} v$ denote the connectivity of vertices u and v in G. G is said to be *locally finite* if, for all $v \in V$, the degree of v in G is finite. A rooted graph (G,o) is a locally finite and connected graph $G = (V,E,o)$ with a distinguished vertex $o \in V$, called the root. Two rooted graphs $(G_1,o_1) = (V_1,E_1,o_1)$ and $(G_2,o_2) = (V_2,E_2,o_2)$ are *isomorphic*, denoted as $(G_1,o_1) \simeq (G_2,o_2)$, if there exists a bijection $\pi : V_1 \to V_2$ such that $\pi(o_1) = o_2$ and $u \overset{G_1}{\sim} v$ if and only if $\pi(u) \overset{G_2}{\sim} \pi(v)$ for all $u,v \in V_1$. One can verify that this notion of isomorphism defines an equivalence relation on rooted graphs. Let $[G,o]$ denote the equivalence class corresponding to (G,o). Let $\mathscr{G}^*$ denote the set of all locally finite and connected rooted graphs. For $(G,o) \in \mathscr{G}^*$ and $h \in \mathbb{N}$, we write $(G,o)_h$ for the truncated graph at depth h of the graph (G,o), in other words, the induced subgraph on the vertices such that their distance from the root is less than or equal to h. The equivalence classes $[G,o]_h$ follows the similar definition. Let $\mathscr{G}_h^*$ denote the set of all $[G,o]_h$. Now, we define the metric d^* on $\mathscr{G}^*$.

For any $[G_1,o_1]$ and $[G_2,o_2]$, let $\hat{h} :=$

$$\sup\{h \in \mathbb{Z}^+ : (G_1,o_1)_h \simeq (G_2,o_2)_h \text{ for some } (G_1,o_1) \in [G_1,o_1], (G_2,o_2) \in [G_2,o_2]\}$$

and define the metric d^* as

$$d^*([G_1,o_1],[G_2,o_2]) := \frac{1}{1+\hat{h}}.$$

As shown in [50], equipped with the metric defined above, $\mathscr{G}^*$ is a Polish space, i.e, a complete separable metric space. For this Polish space, let $\mathscr{P}(\mathscr{G}^*)$ denote the Borel probability measures on it. We say that a sequence of measures $\mu_n \in \mathscr{P}(\mathscr{G}^*)$ *converges weakly* to $\mu \in \mathscr{P}(\mathscr{G}^*)$, written as $\mu_n \rightsquigarrow \mu$, if for any bounded continuous function f on $\mathscr{G}^*$, we have $\int f d\mu_n \to \int f d\mu$. It was shown in [68] that $\mu_n \rightsquigarrow \mu$ if for any uniformly continuous and bounded functions f, we have $\int f d\mu_n \to \int f d\mu$. For $\mu \in \mathscr{P}(\mathscr{G}^*)$, $h \in \{0,1,2,\ldots\}$, and $[G,o] \in \mathscr{G}^*$, let μ_h denote the h-neighborhood marginal of μ

$$\mu_h([G,o]) = \sum_{[G',o] \in \mathscr{G}^* : [G',o]_h = [G,o]} \mu([G',o]).$$

For a locally finite graph $G = (V,E)$ and a vertex $v \in V$, let $G(v)$ denote the graph component in G that is connected to v. By our previous definitions, $(G(v),v)$ denotes the rooted graph of the connected component of v and the root is located at v and $[G(v),v]$ denotes the equivalence class corresponding to $(G(v),v)$. Now, the *rooted neighbourhood distribution* of G is defined as the distribution of the rooted graph when the root is chosen uniformly at random over V

$$U(G) := \frac{1}{|V|} \sum_{v \in V} \delta_{[G(v),v]}, \tag{2.46}$$

where δ is the Dirac delta function.

2.6.2 Local weak convergence

For our study of stochastic block model, which is a sequence of *random* graphs $\{A_n\}_{n=1}^{\infty}$, $U(A_n)$ as defined in (2.46) becomes a random distribution. In the section, we establish the asymptotic behavior of the average neighbourhood distribution $\mathsf{E}U(A_n)$ averaged over the randomness of the graph A_n.

To state the limiting distribution, we define the *Galton–Watson tree* probability distribution on rooted trees $\mathrm{GWT}(\mathrm{P}_\lambda)$ as follows. Let P_λ denote the Poisson distribution with mean λ. We take a vertex as the root and generate $Z^{(1)} \sim \mathrm{P}_\lambda$ as the number of children of the first generation. For the first generation, independent of $Z^{(1)}$, we generate $\xi_1^{(1)}, \ldots, \xi_{Z^{(1)}}^{(1)}$ i.i.d. according to P_λ as the number of children of each vertex in the first generation. Let $Z^{(2)} = \sum_{i=1}^{Z^{(1)}} \xi_i^{(1)}$ denote the total number of vertices in the first generation. In general, for the jth generation, $j = 1, 2, \ldots$, generate the number of children for each vertex in the jth generation $\xi_1^{(j)}, \ldots, \xi_{Z^{(j)}}^{(j)}$ i.i.d. according to P_λ, independent of all previous variables $\{\xi_1^{(i-1)}, \ldots, \xi_{Z^{(i-1)}}^{(i-1)}, Z^{(i)}, \text{ for all } i \leq j\}$. Let $Z^{(j+1)} = \sum_{k=1}^{Z^{(j)}} \xi_k^{(j)}$ denote the total number of vertices in the jth generation. In this way, we iteratively defined a measure on rooted trees. With the definitions above, we are ready to establish the local weak convergence of the symmetric stochastic block model.

Lemma 4 (Local weak convergence of sparse symmetric SBMs)**.** *Let A_n denote a graph generated from a symmetric stochastic block model $\mathrm{SBM}(n, L, \mathbf{p}, \frac{1}{n}\mathbf{Q})$ with $\mathbf{p} = (\frac{1}{L}, \ldots, \frac{1}{L})$, $\mathbf{Q}_{ii} = a, \forall i \in [n]$ and $\mathbf{Q}_{ij} = b, \forall i, j \in [n], i \neq j$. Let $U(A_n)$, defined as in* (2.46)*, be the random rooted neighbourhood distribution of A_n. Then, the average neighbourhood distribution $\mathsf{E}U(A_n)$ converges weakly to a Poisson Galton–Walson tree*

$$\mathsf{E}U(A_n) \rightsquigarrow \mathrm{GWT}(\mathrm{P}_\lambda),$$

where $\lambda = \frac{a+(L-1)b}{L}$.

Remark 6. When $a = b$, the symmetric stochastic block model recovers the well-known local weak convergence result on Erdős–Rényi model (see, e.g., [69, Theorem 3.12]).

Proof of Lemma 4. We want to show that for any uniformly continuous and bounded function f,

$$\left|\int f d\mathsf{E}U(A_n) - \int f d\mathrm{GWT}(\mathrm{P}_\lambda)\right| \to 0$$

as $n \to \infty$. Since f is a uniformly continuous function on $\mathscr{G}^*$, for every $\varepsilon > 0$ there exists $\delta > 0$ such that, for any pair of rooted graphs $[G_1, o_1]$ and $[G_2, o_2] \in \mathscr{G}^*$ with $d^*([G_1, o_1], [G_2, o_2]) < \delta$ we have $|f(G_1, o_1) - f(G_2, o_2)| < \varepsilon$. Recall that $d^*([G_1, o_1], [G_2, o_2]) := \frac{1}{1+\hat{h}}$, where $\hat{h}$ denotes the maximum layers of matching between $[G_1, o_1]$ and $[G_2, o_2]$. Therefore, as long as $h > \frac{1}{\delta} - 1$, we have $|f((G,o)_h) - f(G,o)| < \varepsilon$. It follows that $|f([i,o]) - f([g,o])| < \varepsilon$, if $[i,o]_h = [g,o]$. Let $\mu \in \mathscr{P}(\mathscr{G}^*)$ and assume $h > \frac{1}{\delta} - 1$. We have

$$\left|\int f d\mu_h - \int f d\mu\right| = \left|\sum_{[g,o]\in\mathscr{G}_h^*} f([g,o])\mu_h([g,o]) - \sum_{[i,o]\in\mathscr{G}^*} f([i,o])\mu([i,o])\right| \tag{2.47}$$

$$\leq \sum_{[g,o]\in\mathscr{G}_h^*} \left| f([g,o])\mu_h([g,o]) - \sum_{[i,o]\in\mathscr{G}^*:[i,o]_h=[g,o]} f([i,o])\mu([i,o])\right| \tag{2.48}$$

$$= \sum_{[g,o]\in\mathscr{G}_h^*} \left| \sum_{[i,o]\in\mathscr{G}^*:[i,o]_h=[g,o]} (f([g,o]) - f([i,o]))\mu([i,o])\right| \tag{2.49}$$

$$\leq \sum_{[g,o]\in\mathscr{G}_h^*} \sum_{[i,o]\in\mathscr{G}^*:[i,o]_h=[g,o]} |f([g,o]) - f([i,o])|\,\mu([i,o]) \tag{2.50}$$

$$\leq \sum_{[g,o]\in\mathscr{G}_h^*} \sum_{[i,o]\in\mathscr{G}^*:[i,o]_h=[g,o]} \varepsilon\mu([i,o]) = \varepsilon, \tag{2.51}$$

where (3) follows since $\mu_h([g,o]) = \sum_{[i,o]\in\mathscr{G}^*:[i,o]_h=[g,o]} \mu([i,o])$. Thus, we have $|\int f d\mathsf{E}U(A_n)_h - \int f d\mathsf{E}U(A_n)| < \varepsilon$ and $|\int f d\mathrm{GWT}(\mathrm{P}_\lambda)_h - \int f d\mathrm{GWT}(\mathrm{P}_\lambda)| < \varepsilon$. Let $B \subseteq \mathscr{G}^*$ be a measurable event in $\mathscr{G}^*$. By exchangability property of the SBM, we have $\mathsf{E}U(A_n)(B) = \frac{1}{n}\sum_{i=1}^n \mathsf{P}([A_n(i), i] \in B) = \mathsf{P}([A_n(1), 1] \in B)$. In other words, $\mathsf{E}U(A_n)$ is simply the neighbourhood distribution at vertex 1. By the analogous argument as in proposition 2 of [65], for any $\varepsilon > 0$, there exists n_0 such that if $n \geq n_0$ and $\frac{\ln n}{10\ln(2(a+(L-1)b))} \geq R$, we have $d_{TV}(\mathrm{GWT}(\mathrm{P}_\lambda)_R, \mathsf{E}U(A_n)_R) < \varepsilon$, where $d_{TV}(\cdot,\cdot)$

denotes the total variation distance between two measures. Remember here the total variation distance is $d_{TV}(\mu_1,\mu_2) := \sup_{g:\mathscr{G}^* \to [-1,1]}(\int g d\mu_1 - \int g d\mu_2)$. Since f is a bounded function, we have $|\int f d\mathrm{GWT}(\mathrm{P}_\lambda)_R - \int f d\mathsf{E}U(A_n)_R| < \varepsilon$, as long as n is large enough. Therefore, if we take n large enough such that $\frac{\ln n}{10\ln(2(a+(L-1)b))} > \frac{1}{\delta} - 1$ and $|\int f d\mathrm{GWT}(\mathrm{P}_\lambda)_h - \int f d\mathsf{E}U(A_n)_h| < \varepsilon$, we have

$$\begin{aligned}\left|\int f d\mathsf{E}U(A_n) - \int f d\mathrm{GWT}(\mathrm{P}_\lambda)\right| &\le \left|\int f d\mathsf{E}U(A_n)_h - \int f d\mathsf{E}U(A_n)\right| \\ &+ \left|\int f d\mathrm{GWT}(\mathrm{P}_\lambda)_h - \int f d\mathrm{GWT}(\mathrm{P}_\lambda)\right| \\ &+ \left|\int f d\mathrm{GWT}(\mathrm{P}_\lambda)_h - \int f d\mathsf{E}U(A_n)_h\right| \\ &< 3\varepsilon,\end{aligned}$$

which completes the proof. □

2.6.3 BC entropy

In this section, we review the notion of BC entropy introduced in [59], which is shown to be the fundamental limit of universal lossless compression for certain graph family [50].

For a Polish space Ω, let $\mathscr{P}(\Omega)$ denote the set of all Borel probability measures on Ω. Let A be a Borel set in Ω, we define the *ε-extension of A*, denoted A^ε, as the union of the open balls with radius ε centered around the points in A. For two probability measures μ and ν in $\mathscr{P}(\Omega)$, we define the *Lévy–Prokhorov distance* $d_{\mathrm{LP}}(\mu,\nu) := \inf\{\varepsilon > 0 : \mu(A) \le \nu(A^\varepsilon) + \varepsilon$ and $\nu(A) \le \mu(A^\varepsilon) + \varepsilon, \forall A \in \mathscr{B}(\Omega)\}$, where $\mathscr{B}(\Omega)$ denotes the Borel sigma algebra of Ω. Let $\rho \in \mathscr{P}(\mathscr{G}^*)$. Let d be the expected number of neighbours of root under the law ρ and let a sequence $m = m(n)$ such that $m/n \to d/2$, as $n \to \infty$. Define $\mathscr{G}_{n,m}$ to be the set of graphs with n vertices and m edges. For $\varepsilon > 0$, define

$$\mathscr{G}_{n,m}(\rho,\varepsilon) = \{G \in \mathscr{G}_{n,m} : U(G) \in B(\rho,\varepsilon)\},$$

where $B(\rho,\varepsilon)$ denotes the open ball with radius ε around ρ with respect to Lévy–Prokhorov metric. Now, we define the *ε-upper BC entropy* of ρ as

$$\overline{\Sigma}(\rho,\varepsilon) = \limsup_{n\to\infty} \frac{\log|\mathscr{G}_{n,m}(\rho,\varepsilon)| - m\log n}{n}$$

and define the *upper BC entropy* of ρ as

$$\overline{\Sigma}(\rho) = \lim_{\varepsilon\to 0} \overline{\Sigma}(\rho,\varepsilon).$$

Similarly we define the *ε-lower BC entropy* $\underline{\Sigma}(\rho,\varepsilon)$ and *lower BC entropy* $\underline{\Sigma}(\rho)$ with $\limsup$ replaced by $\liminf$ in above definitions. If ρ is such that $\overline{\Sigma}(\rho) = \underline{\Sigma}(\rho)$, then this common limit is called the *BC entropy* of ρ

$$\Sigma(\rho) := \overline{\Sigma}(\rho) = \underline{\Sigma}(\rho).$$

The following lemma states the BC entropy of the Galton–Waston tree distribution.

Lemma 5 (Corollary 1.4 of [59])**.** *The BC entropy of the Galton–Watson tree distribution* $\mathrm{GWT}(\mathrm{P}_\lambda)$ *is given by*

$$\Sigma\left(\mathrm{GWT}(\mathrm{P}_\lambda)\right) = \frac{\lambda}{2}\log\frac{e}{\lambda} \quad \text{bits.}$$

2.6.4 Achieving BC entropy in the sparse regime

With the Lemma above, we can give a performance guarantee of our algorithm corresponding to the BC entropy. It is a Theorem analog to Proposition 1 in [50].

Theorem 8. *Let* $A_n \sim \mathrm{SBM}\left(n, L, \mathbf{p}, \frac{1}{n}\mathbf{Q}\right)$ *with* $\mathbf{p} = \left(\frac{1}{L},\ldots,\frac{1}{L}\right)$*,* $\mathbf{Q}_{ii} = a, \forall i \in [n]$ *and* $\mathbf{Q}_{ij} = b, \forall i,j \in [n], i \neq j$*. Let* $\lambda = \mathbf{p}^T\mathbf{Q}\mathbf{p} = \frac{a+(L-1)b}{L}$ *and* $m = \binom{n}{2}\frac{\lambda}{n}$ *be the expected number of edges in the model. Then, our compression algorithm achieves the BC entropy of the local weak limit of stochastic block models in the sense that*

$$\limsup_{n\to\infty} \frac{\mathsf{E}[\ell(C_k(A_n))] - m\log n}{n} \leq \Sigma\left(\mathrm{GWT}(\mathrm{P}_\lambda)\right).$$

Proof. By our proof of theorem (need to fill in ref), we have

$$\mathsf{E}[\ell(C_k(A_n))] \leq \binom{n'}{2} H(\mathbf{B}_{12}) + 2^{k^2}\log(2en^3) + nk_n^2 H(A_{12}).$$

Notice that

$$\begin{aligned}
\binom{n'}{2} H(\mathbf{B}_{12}) &\leq \binom{n'}{2} k^2 H(A_{12}) \\
&= \binom{n'}{2} k^2 h(\lambda/n) \\
&\overset{(1)}{=} \binom{n'}{2} k^2 \left(\frac{1}{n}\lambda \log \frac{ne}{\lambda} + o\left(\frac{1}{n}\right)\right) \\
&\overset{(2)}{\sim} \binom{n}{2} \left(\frac{1}{n}\lambda \log n + \frac{1}{n}\lambda \log \frac{e}{\lambda} + o\left(\frac{1}{n}\right)\right) \\
&= \binom{n}{2} \frac{1}{n}\lambda \log n + \frac{\lambda \log e - \lambda \log \lambda}{2} n + o(n) \\
&\overset{(3)}{=} m \log n + n\Sigma\left(\mathrm{GWT}(\mathrm{P}_\lambda)\right) + o(n)
\end{aligned}$$

where (1) follows since $h(p) = p\log\frac{e}{p} - \frac{\log e}{2}p^2 + o(p^2)$, (2) follows since $n'k = n$ and (3) follows from Lemma 5. Then it suffices to that the remaining terms in the upper bound of $\mathsf{E}[\ell(C_k(A_n))]$ are all $o(n)$. Indeed we have

$$2^{k^2}\log(2en^3) \leq 2^{\delta \log n}\log(2en^3) = n^{\delta}\log(2en^3) = o(n)$$

since $\delta < 1$ and

$$\begin{aligned}
nk_n^2 H(A_{12}) &= nk_n^2 h(\lambda/n) \\
&= nk_n^2 \left(\frac{1}{n}\lambda \log \frac{ne}{\lambda} + o\left(\frac{1}{n}\right)\right) \\
&\leq n\delta \log n \left(\frac{1}{n}\lambda \log \frac{ne}{\lambda} + o\left(\frac{1}{n}\right)\right) \\
&= \delta \log n \left(\lambda \log \frac{ne}{\lambda}\right) + o\left(\log n\right)
\end{aligned}$$

$$= o(n).$$

□

Remark 7. For sparse symmetric SBMs, Theorem 8 shows that our compressor achieves the BC entropy of the Galton–Watson tree that is the local weak convergence limit of the underlying sequence of graphs. We note, however, that for the family of sparse symmetric SBMs, it is unclear if this BC entropy is the fundamental limit of lossless compression. This is because the family of sparse symmetric SBMs does not belong to the family of random graphs considered in [58], where a converse statement can be established.

2.7 Stationarity in the stochastic block model

In this section, we take a closer look at the correlation among entries in the adjacency matrix and explain why existing universal compressors developed for stationary processes may not be immediately applicable for certain orderings of the entries.

Compressing A_n entails compressing

$$A_{12},\dots,A_{1,n},A_{23},\dots,A_{n-1,n},$$

i.e. the bits in the upper triangle of A_n. Clearly, these are not independent (because of the dependency through X_1^n) so one cannot use any of the compressors universal for the class of iid processes to compress A_n. So, one hopes that it is possible to list the $\binom{n}{2}$ random variables $A_{12},\dots,A_{1,n},A_{23},\dots,A_{n-1,n}$ in an order that makes the resulting sequence stationary, so that the Lempel–Ziv compressor (which, recall, is universal for the class of stationary processes) may be used. However, we show now that some of the most natural orders of listing these $\binom{n}{2}$ bits result in a sequence that is nonstationary.

1. **Horizontally:** Listing the bits in the upper triangle row-wise (i.e. first listing the bits in the first row, followed by the bits in the second and so on, ending with $A_{n-1,n}$) we get the

following sequence

$$A_{12},\ldots,A_{1,n},A_{23},\ldots,A_{2,n},\ldots,A_{n-1,n},$$

which can be seen to be nonstationary. Consider the case when $n=4, L=2, Q_{11}=Q_{12}=1, Q_{12}=0$. In this case the horizontal ordering is

$$A_{12},A_{13},A_{14},A_{23},A_{24},A_{34}$$

and this is seen to be nonstationary by observing $P(A_{12}=1,A_{13}=0,A_{14}=1)>0$ but $P(A_{23}=1,A_{24}=0,A_{34}=1)=0$.

2. **Vertically:** Listing the bits in the upper triangle column-wise (i.e. first listing the bits in the first column, followed by the bits in the second and so on, ending with $A_{n-1,n}$) we get the following sequence

$$A_{12},A_{13},A_{23},\ldots,A_{1,n},\ldots,A_{n-1,n},$$

which can be seen to be nonstationary. Consider the case when $n=4, L=2, Q_{11}=Q_{12}=1, Q_{12}=0$. In this case the vertical ordering is

$$A_{12},A_{13},A_{23},A_{14},A_{24},A_{34}$$

and this is seen to be nonstationary by observing $P(A_{12}=1,A_{13}=0,A_{23}=1)=0$ but $P(A_{14}=1,A_{24}=0,A_{34}=1)>0$.

3. **Diagonally:** Consider $\lfloor \frac{n}{2} \rfloor$ sequences defined as

$$S_1 := A_{12},A_{23},A_{34},\ldots,A_{n-1,n},A_{n,1}$$

$$S_2 := A_{13},A_{24},A_{35},\ldots,A_{n-2,n},A_{n-1,1},A_{n,2}$$

$$\vdots$$

$$S_{\lfloor \frac{n}{2} \rfloor - 1} := A_{1,1+\lfloor \frac{n}{2} \rfloor - 1}, A_{2,2+\lfloor \frac{n}{2} \rfloor - 1}, \cdots, A_{n,\lfloor \frac{n}{2} \rfloor - 1}$$

and

$$S_{\lfloor \frac{n}{2} \rfloor} = \begin{cases} A_{1,1+n/2}, A_{2,2+n/2}, \ldots, A_{n/2,n}, & \text{when } n \text{ is even,} \\ A_{1,1+\lfloor \frac{n}{2} \rfloor}, A_{2,2+\lfloor \frac{n}{2} \rfloor}, \ldots, A_{n,n+\lfloor \frac{n}{2} \rfloor}, & \text{when } n \text{ is odd.} \end{cases}$$

Concatenating $S_1, \ldots, S_{\lfloor \frac{n}{2} \rfloor}$ yields a sequence of length $\binom{n}{2}$. This corresponds to listing the bits diagonal-wise. However, even this does not yield a sequence that is stationary which can be illustrated by considering the case when $n = 4, L = 2, Q_{11} = Q_{12} = 1, Q_{12} = 0$. In this case the diagonal ordering is

$$A_{12}, A_{23}, A_{34}, A_{41}, A_{13}, A_{24}$$

and this is seen to be nonstationary by observing $P(A_{12} = 0, A_{23} = 1, A_{34} = 1) > 0$ but $P(A_{34} = 0, A_{41} = 1, A_{13} = 1) = 0$.

2.8 Experiments

We implement the proposed universal graph compressor (UGC) in four widely used benchmark graph datasets: protein-to-protein interaction network (PPI) [70], LiveJournal friendship network (Blogcatalog) [71], Flickr user network (Flickr) [71], and YouTube user network (YouTube) [72]. The block decomposition size k is chosen to be $1, 2, 3, 4$ and we present in Table 2.1 the compression ratios (the ratio between output length and input length of the encoder) of UGC for different choices of k. We present in Table 2.2 the compression ratios of four competing algorithms.

- CSR: Compressed sparse row is a widely used sparse matrix representation format. In the experiment, we further optimize its default compressor exploiting the fact that the graph is simple and its adjacency matrix is symmetric with binary entries.

- Ligra+: This is another powerful sparse matrix representation format [73, 74], which improves upon CSR using byte codes with run-length coding.
- LZ: This is an implementation of the algorithm proposed in [75], which first transforms the two-dimensional adjacency matrix into a one-dimensional sequence using the Peano–Hilbert space filling curve and then compresses the sequence using Lempel–Ziv 78 algorithm [13].
- PNG: The adjacency matrix of the graph is treated as a gray-scaled image and the PNG lossless image compressor is applied.

The compression ratios of the five algorithms implemented on four datasets are given as follows. The proposed UGC outperforms all competing algorithms in all datasets. The compression ratios from competing algorithms are 2.4 to 27 times that of the universal graph compressor.

Table 2.1. Compression ratio of UGC under different k values.

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
PPI	0.0228	**0.0226**	0.0227	0.034
Blogcatalog	0.0275	0.0270	**0.0267**	0.0288
Flickr	0.00960	0.00935	0.00915	**0.00907**
YouTube	4.51×10^{-5}	4.11×10^{-5}	$\mathbf{3.98 \times 10^{-5}}$	4.00×10^{-5}

Table 2.2. Compression ratios of competing algorithms.

	CSR	Ligra+	LZ	PNG
PPI	0.166	0.0605	0.06	0.089
Blogcatalog	0.203	0.0682	0.080	0.096
Flickr	0.0584	0.0217	0.0307	0.0262
YouTube	3.23×10^{-4}	9.90×10^{-5}	1.09×10^{-4}	1.10×10^{-3}

Note, however, that CSR and Ligra+ are designed to enable fast computation, such as adjacency query or vertex degree query, in addition to compressing the matrix. Our proposed compressor does not possess such a functionality and is designed solely for the purpose of compression.

2.9 Acknowledgement

This chapter includes the material in Alankrita Bhatt, Ziao Wang, Chi Wang, Lele Wang, “Universal graph compression: Stochastic block models”, arXiv 2006.02643 (to be submitted), with a short version in the IEEE International Symposium on Information Theory (ISIT), 2021. The dissertation author was a primary investigator and author of this paper.

Chapter 3

Sequential Prediction With Log-loss and Side Information

3.1 Introduction

We consider a variant of the problem of sequential prediction under log-loss with side information. The particular variant under consideration was first studied in [76]. Let $X \in \mathscr{X}$ and $Y \in \{0,1\}$ denote two jointly distributed random variables. Let the marginal distribution of X be denoted by $P_X(x)$. A hypothesis f in the *hypothesis class* $\mathscr{F}$ determines the conditional distribution $P_f(y|x)$, or equivalently, the conditional probability mass function (pmf) $p_f(y|x)$, for $y \in \{0,1\}$ and $x \in \mathscr{X}$. Each hypothesis is characterized by a tuple $f = (g, \theta_0, \theta_1)$ where

1. $\theta_0, \theta_1 \in [0,1]$
2. $g \in \mathscr{G} \subset \{\mathscr{X} \to \{0,1\}\}$.

In other words, g belongs to a class $\mathscr{G}$ of binary functions. We assume that $\mathscr{G}$ has finite VC dimension, denoted by VCdim$(\mathscr{G})$.

Given a chosen hypothesis $f = (g, \theta_0, \theta_1)$ we then have

$$Y \,|\, \{X = x\} \sim \text{Bernoulli}(\theta_{g(x)}).$$

Thus, given the *side information* X, the random variable Y is distributed as either Bernoulli(θ_0)

or Bernoulli(θ_1). Picking a hypothesis $f \in \mathscr{F}$, let $(X_i,Y_i)_{i=1}^n$ be drawn i.i.d. from the joint distribution of X and Y characterized by the hypothesis f, so

$$P(x^n,y^n) = \prod_{i=1}^{n} P_X(x_i)P_f(y_i|x_i). \tag{3.1}$$

The problem of sequential prediction under log-loss, also known as the sequential probability assignment problem, can be thought of as a game between the player and nature. First, nature picks a hypothesis $f \in \mathscr{F}$ unbeknownst to the player, and X^n, Y^n are then generated according to the law (3.1). At each time step $i \in [n]$, X_i is revealed to the player, who then assigns a probability mass function (pmf) $q(\cdot|X^i,Y^{i-1})$ to Y_i. Next, Y_i is revealed and the player incurs loss $-\log q(\cdot|X^i,Y^{i-1})$. Nature assigns the pmf $p_f(\cdot|X_i)$ at each time step i and incurs loss $-\log p_f(Y_i|X_i)$. The goal of the game is to minimize the expected value of cumulative loss relative to nature (known as the regret), without knowledge of f. Importantly, we also wish to do this without knowing P_X either.

To make this notion precise, define the regret incurred by the probability assignment q when nature picked f and the distribution of X is P_X as

$$R_{n,P_X}(q,f) := \mathbb{E}\left[\sum_{i=1}^{n} \log\frac{1}{q(Y_i|X^i,Y^{i-1})} - \sum_{i=1}^{n} \log\frac{1}{p_f(Y_i|X_i)}\right]. \tag{3.2}$$

Then, the worst-case regret for the probability assignment q is

$$R_n(q) := \max_{P_X,f} R_{n,P_X}(q,f). \tag{3.3}$$

In this paper, we aim to calculate the min-max regret

$$R_n := \min_{q} R_n(q). \tag{3.4}$$

and discover a probability assignment q that is optimal or near-optimal in the sense of achieving

$R_n(q)$ close to the optimal value (3.4).

The log-loss is of central importance in information theory as it connects two canonical problems in data science—compression and prediction; see the survey [77]. To motivate the use of the log-loss in the current problem, we view it as an extension of the problem of universal compression. Indeed, if there is no side information X present, then the problem is equivalent to universal compression of an i.i.d. Bernoulli source which has been well studied [25, 78–81]. The minimax regret R_n then is significant operationally, representing the number of extra bits above the entropy one must pay as the price for compressing the source without knowing its distribution. Remarkably, one can show that $R_n = \frac{1}{2}\log n + o(\log n)$ in this setting. In a similar vein, [82] studies a closely related problem where a compressed version of the sequence Y^n is available as side information noncausally (i.e. not sequentially) and demonstrate its equivalence to lossy compression.

In the current setting, if the function g is known, then simple extensions of the techniques developed to tackle the problem of universal compression of an i.i.d. Bernoulli source can be used to show that $R_n \leq \log n + o(\log n)$, and we will elaborate on this important special case in detail in Section 3.2.1. The problem becomes nontrivial when the function g is not known, and new techniques need to be developed to characterize R_n in this case.

In the standard study of classification in statistical learning theory, the loss function employed is the 0-1 loss or the indicator loss, and the notion of VC dimension plays a crucial role in characterizing the fundamental limits of binary classification [83]. In particular, $\text{VCdim}(\mathscr{G}) < \infty$ implies the PAC-learnability of the hypothesis class $\mathscr{G}$. Viewing the current setting as a log-loss variant of the standard classification problem studied in statistical learning (which uses the indicator loss) motivates the choice of constraint $\text{VCdim}(\mathscr{G}) < \infty$. A variant of the current problem with indicator loss instead of log-loss was studied in [84]. We have considered a specific class of conditional distributions to compete against (recall that under hypothesis f we have $p_f(Y = 0|X = x) = \text{Bern}(\theta_{g(x)})$). As mentioned in the preceding paragraphs, our motivation stems from universal compression with side information, and to consider a log-loss variant of the standard binary classification problem. In both these cases, the choice of the considered class seems natural. However, in general, one could view this problem as an online conditional density estimation problem and correspondingly consider an arbitrary class $\mathscr{F}$ where any $f \in \mathscr{F}$ may characterize the conditional distribution $p_f(y|x)$ in a far more complex manner. It then makes sense to expect R_n in this case to depend on a measure of complexity of $\mathscr{F}$ akin to the VC dimension. Indeed, in [85] the authors develop a remarkable theory parallel to statistical learning theory when the data is non-i.i.d. They develop analogues of several combinatorial dimensions and the Rademacher complexity in the non-i.i.d. case. They then leverage this theory in [86] to study the minmax regret in several online learning problems (with adversarial data). This approach is employed to study sequential prediction with the log-loss in [87] and [88]. However, it is important to note that the proofs in these works are nonconstructive—they proceed via using minmax duality and analyzing the dual game, which does not provide a strategy (i.e. a probability assignment) achieving the regret upper bound that is proven. Our method on the other hand involves construction of a sequential probability assignment. In the next subsection, we will mention and compare our results with the aforementioned two papers studying the log-loss.

3.1.1 Main Results

Our first main result is a probability assignment that yields an upper bound on R_n.

Theorem 1. *If $\mathscr{G}$ is such that VCdim$(\mathscr{G}) = d < \infty$, we have for an absolute constant $C \leq 250$, for a probability assignment q^* (which is specified in detail further on)*

$$R_n(q^*) \leq 125C\sqrt{dn}\log(2n) + d(\log n)^2 + 2. \tag{3.5}$$

Moreover, for any $P_X, f, \delta \in (0,1)$, with probability greater than $1-\delta$,

$$\sum_{i=1}^{n} \log \frac{1}{q^*(Y_i|X^i,Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)}$$

$$\leq 25C\sqrt{dn}\log(2n)\left(C\sqrt{d}+\sqrt{2\log\frac{2\log n}{\delta}}\right)+d(\log n)^2+2 \tag{3.6}$$

The proof is deferred to Section 3.4, where we construct and analyze the probability assignment q^*. In [76], the authors showed $R_n = O(d\sqrt{n}\log n)$, and $R_n \leq \left(2d+1+\log\frac{1}{\delta}\right)\sqrt{n}\log n$ with probability $\geq 1-\delta$. Our proof (and probability assignment) is different and achieves the same dependence on n, and a better dependence on δ in the high-probability version of the result.

We also establish a lower bound on R_n.

Theorem 2. *We have*

$$R_n \geq d+\log(n+1)-2\sqrt{e}d^2e^{-3n/100d}-\log(\pi e).$$

The proof is deferred to Section 3.5.

The non-constructive approaches of the papers [87] and [88] mentioned earlier establish an $O(d\log n)$ upper bound for the $\mathscr{F}$ under consideration. In conjunction with Theorem 2 we see that the dependence of R_n on n is indeed $\Theta(\log n)$. This implies that the q^* employed to prove Theorem 1 is suboptimal and a better probability assignment could be constructed. As a starting step, we considered a few special cases of the function class $\mathscr{G}$ in the hypothesis class and provide a sequential probability assignment achieving $O(d\log n)$ upper bound. These upper bounds constitute our third main result.

3.1.2 Organization and Notation

In Section 3.2 we provide basic notation and results that will be used in the proofs of our main results. In Section 3.3, we provide logarithmic upper bounds on R_n for a few special cases. Section 3.4 is devoted to the proof of Theorem 1, and Section 3.5 is devoted to the proof of Theorem 2. Finally, Section 3.6 concludes. All the proofs throughout the paper are relegated to the Appendix.

Notation: Throughout the paper, $\log(\cdot)$ refers to the logarithm to base 2, and $\ln(\cdot)$ refers to logarithm to base e. The Hamming distance between two binary vectors x and y is denoted by $d_{\mathrm{H}}(x,y)$. The fact that two random variables Z_1 and Z_2 have the same distribution is denoted by $Z_1 \overset{(d)}{=} Z_2$.

3.2 Mathematical Preliminaries

This section introduces some basic notation and results that form the building blocks of the proofs of our main results.

To prove an upper bound on R_n, it suffices to show a probability assignment $q(Y_i|X^i,Y^{i-1})$ that achieves regret $R_n(q)$ that is less than the given upper bound. To this end, we will use a *mixture probability assignment*

$$q_{\mathrm{mix}}(y_i|x^i,y^{i-1}) := \frac{\mathbb{E}_F\left[p_F\left(y^i|x^i\right)\right]}{\mathbb{E}_F\left[p_F\left(y^{i-1}|x^{i-1}\right)\right]} \tag{3.7}$$

where $F = (\Theta_0,\Theta_1,G) \in \mathscr{F}$ is a random variable with some distribution over the hypothesis class $\mathscr{F}$. The usage of such a mixture probability assignment is inspired by previous work in universal prediction and universal compression, and we discuss this choice in further detail in Section 3.2.1. It can be verified that $q_{\mathrm{mix}}(y_i|x^i,y^{i-1})$ is indeed a probability assignment.

Proposition 1. *For any x^i, y^{i-1}, we have $\sum_{y_i\in\mathscr{Y}} q_{\mathrm{mix}}(y_i|x^i,y^{i-1}) = 1$.*

For the probability assignment q_{mix} in (3.7) we can establish the following.

Proposition 2. *We have*

$$\sum_{i=1}^{n} \log \frac{1}{q_{\mathrm{mix}}(Y_i|X^i,Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)} = \log \frac{p_f(Y^n|X^n)}{\mathbb{E}[p_F(Y^n|X^n)]}. \tag{3.8}$$

The choice of the distribution of F is important and greatly affects $R_n(q_{\mathrm{mix}})$. Almost all throughout this paper, for $F = (\Theta_0,\Theta_1,G)$, we will choose Θ_0, Θ_1 and G to be mutually independent, with $\Theta_0,\Theta_1 \sim \mathrm{Beta}\left(\frac{1}{2},\frac{1}{2}\right)$ each. The choice of the distribution of G (which, recall, is over the class of functions $\mathscr{G}$) will be varied across different problems. The Beta $\left(\frac{1}{2},\frac{1}{2}\right)$ density is denoted by $w(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}}$. The choice $w(\theta)$ is elaborated upon in the next subsection.

3.2.1 When $|\mathscr{G}| = 1$

In this subsection, we consider the rather simple case when the class of functions $|\mathscr{G}|$ contains only one function g^* (or, equivalently, the function g^* picked by nature is known). Thus, in this case, the hypothesis f picked is of the form $(\theta_0, \theta_1, g^*)$. Considering g^* to be the function for which $g^*(x) = 0\, \forall\, x \in \mathscr{X}$, as mentioned previously in the introduction, we recover the setting of universal compression over the class of binary i.i.d processes. In this case, the minmax regret R_n in (3.4) reduces to

$$R_n = \min_{q} \max_{\theta \in [0,1]} \mathbb{E}\left[\log \frac{p_\theta(Y^n)}{q(Y^n)}\right] = \min_{q} \max_{\theta \in [0,1]} D_{\mathrm{KL}}(p_\theta(Y^n)||q(Y^n)) \tag{3.9}$$

where $Y_i \sim \text{Bernoulli}(\theta)$ i.i.d. and $p_\theta(\cdot)$ is the probability law for this process. As mentioned in the introduction, it is well known that in this case

$$R_n = \frac{1}{2}\log n + o(\log n), \tag{3.10}$$

and that this is asymptotically achieved by an instance of the mixture probability assignment (3.7) given by

$$q_{\mathrm{KT}}(y_i|y^{i-1}) = \frac{\int_0^1 p_\theta(y^i)w(\theta)d\theta}{\int_0^1 p_\theta(y^{i-1})w(\theta)d\theta} = \frac{\mathbb{E}_\Theta[p_\Theta(y^i)]}{\mathbb{E}_\Theta[p_\Theta(y^{i-1})]}$$

with $\Theta \sim \text{Beta}(1/2, 1/2)$. This probability assignment is known as the Krichevsky–Trofimov (KT) probability assignment [89] and thus motivates the utilization of the Beta(1/2,1/2) prior for Θ_0 and Θ_1. For a sequence y^n, the sequential probability assignment $q_{\mathrm{KT}}(y_{i+1}|y^i)$ turns out to be the so-called “add-1/2" estimator which sets $q_{\mathrm{KT}}(0|y^i) = \frac{\sum_{t=1}^{i} \mathbb{1}\{y_t=0\}+1/2}{i+1}$. Moreover, it can be shown that if $k = \sum_{i=1}^{n} y_i$,

$$q_{\mathrm{KT}}(y^n) = \int_0^1 p_\theta(y^n)w(\theta)d\theta = \frac{1}{4^n}\frac{\binom{n}{k}\binom{2n}{n}}{\binom{2n}{2k}} \tag{3.11}$$

When the range of g^* includes both 0 and 1, a modification of the KT probability assignment can achieve regret $\log n + o(\log n)$.

Consider the sequential probability assignment

$$q_{\mathrm{KT}}(0|x^i, y^{i-1}) = \frac{\sum_{t=1}^{i-1} \mathbb{1}\{y_t = 0, g^*(x_t) = g^*(x_i)\} + 1/2}{\sum_{t=1}^{i-1} \mathbb{1}\{g^*(x_t) = g^*(x_i)\} + 1}$$

Without the x_i, this can be seen to be the standard Krichevsky–Trofimov (KT) probability assignment for binary i.i.d. processes. With the side information x_i, q_{KT} is seen to be a “block-wise" or “symbol-wise" KT probability assignment. This can be seen to be a probability assignment of the form in (3.7) with

$$q_{\mathrm{KT}}(y_i|x^i, y^{i-1}) = \frac{\int_0^1 \int_0^1 p_{g^*,\theta_0,\theta_1}(y^i|x^i)w(\theta_0)w(\theta_1)d\theta_0 d\theta_1}{\int_0^1 \int_0^1 p_{g^*,\theta_0,\theta_1}(y^{i-1}|x^{i-1})w(\theta_0)w(\theta_1)d\theta_0 d\theta_1}$$

We can then bound the regret achieved by the probability assignment q_{KT}.

Lemma 1. *When the function class $\mathscr{G}$ is such that $|\mathscr{G}| = 1$, we have*

$$R_n(q_{\mathrm{KT}}) \leq \log\left(\frac{n}{2} + 1\right) + \log\frac{\pi^2}{8}. \tag{3.12}$$

Remark 3 (Laplace probability assignment)**.** *Instead of using the* $\mathrm{Beta}(1/2, 1/2)$ *prior, one can use the* $\mathrm{Uniform}[0,1]$ *prior and choose the sequential probability assignment*

$$q_{\mathrm{L}}(y_{i+1}|y^i) = \frac{\int_0^1 p_\theta(y^{i+1})d\theta}{\int_0^1 p_\theta(y^i)d\theta},$$

which yields the so-called Laplace or the add-1 probability assignment. It can be shown that for *the problem* (3.9)*, $q_{\mathrm{L}}(\cdot)$ can achieve $R_n(q_{\mathrm{L}}) \leq \log n + o(\log n)$. Thus, the Laplace probability assignment achieves the optimal regret in order but with a slightly larger constant, a result that even holds for very rich expert classes [90]. It can be shown that if $k = \sum_{i=1}^{n} y_i$, we have $q_{\mathrm{L}}(y^n) = \int_0^1 p_\theta(y^n)d\theta = \frac{1}{(n+1)\binom{n}{k}}$. For mathematical convenience, we will use the Laplace probability assignment later in the paper, specifically in Sections 3.2.4 and 3.4.*

3.2.2 When $|\mathscr{G}| < \infty$

When $|\mathscr{G}| < \infty$, we can use a probability assignment (3.7) with $G \sim \mathrm{Uniform}(\mathscr{G})$. Then, for this choice of mixture, we have

$$\mathbb{E}_F\left[p_F\left(y^i|x^i\right)\right] = \frac{1}{|\mathscr{G}|}\sum_{g\in\mathscr{G}}\int_0^1\int_0^1 p_{g,\theta_0,\theta_1}(y^i|x^i)w(\theta_0)w(\theta_1)d\theta_0 d\theta_1 \tag{3.13}$$

where $w(x) = \frac{1}{\sqrt{x(1-x)}}$ is the $\mathrm{Beta}(1/2,1/2)$ prior as before.

We can then establish the following upper bound on the regret for the probability assignment q_{mix} characterized by the mixture (3.13).

Lemma 2. *For the probability assignment q_{mix} with characterized by the mixture* (3.13)*, we have*

$$R_n(q_{\mathrm{mix}}) \leq \log|\mathscr{G}| + \log\left(\frac{n}{2}+1\right) + \log\frac{\pi^2}{8}. \tag{3.14}$$

3.2.3 Side Information X^n Available Noncausally

In this subsection we consider the special case when the side information X^n is available *noncausally* instead of sequentially. The results and intuition developed in this section will be used in proofs further ahead.

When the side information X^n is available noncausally, the probability assignment for Y_i

is of the form $q(Y_i|X^n, Y^{i-1})$ and the regret for a probability assignment q can be seen to be

$$R_{n,\text{nc}}(q) = \max_{P_X, f} \mathbb{E}_{X^n,Y^n}\left[\sum_{i=1}^{n} \log \frac{1}{q(Y_i|X^n, Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)}\right]. \tag{3.15}$$

Since the side information X^n is available in advance, we can choose our mixture over the hypothesis class $\mathscr{F}$ to be dependent on X^n. As done so far, for $H = (\Theta_0, \Theta_1, G)$ we will choose Θ_0, Θ_1 and G to be mutually independent with $\Theta_0, \Theta_1 \sim \text{Beta}(1/2, 1/2)$. We will now define a distribution over $\mathscr{G}$ that is dependent on the side information X^n.

Given X^n, define the set $\mathscr{P}_n(X^n) = \{(g(X_1), \ldots, g(X_n)), g \in \mathscr{G}\} \subseteq \{0,1\}^n$. For the remainder of this subsection, for brevity we will refer to $\mathscr{P}_n(X^n)$ by just $\mathscr{P}_n$. Enumerating the elements of $\mathscr{P}_n$ by $1, 2, \ldots, |\mathscr{P}_n|$, we now define the set

$$I_j = \{g \in \mathscr{G}, (g(X_1), \ldots, g(X_n)) = \mathscr{P}_n(j)\} \tag{3.16}$$

where $\mathscr{P}_n(j)$ represents the j-th element in $\mathscr{P}_n$. Clearly, the sets $I_1, \ldots, I_{|\mathscr{P}_n|}$ are nonempty and partition $\mathscr{G}$. So, X^n can be thought of as partitioning $\mathscr{G}$ into sets where any two functions g_1, g_2 in the same partition have $g_1(X_j) = g_2(X_j) \forall j \in [n]$.

Pick an arbitrary $g_i \in I_i$ for $i \in [|\mathscr{P}_n|]$. Choosing $G \sim \text{Uniform}\left\{g_1, \ldots, g_{|\mathscr{P}_n|}\right\}$, we have

$$\mathbb{E}_F\left[p_F\left(y^i|x^i\right)\right] = \frac{1}{|\mathscr{P}_n|}\sum_{i=1}^{|\mathscr{P}_n|} \int_0^1 \int_0^1 p_{g_i,\theta_0,\theta_1}(y^i|x^i) w(\theta_0) w(\theta_1) d\theta_0 d\theta_1. \tag{3.17}$$

Lemma 3. *For the probability assignment q_{mix} characterized by the mixture* (3.17)*, we have*

$$R_{n,\text{nc}}(q_{\text{mix}}) \leq d \log(en/d) + \log\left(\frac{n}{2} + 1\right) + \log \frac{\pi^2}{8}. \tag{3.18}$$

3.2.4 When P_X is Known

Consider the case when the distribution P_X is known. In this case, the main idea is to choose the distribution of G to be uniform over a *finite* set of functions in $\mathscr{G}$ that form a fine-enough covering of $\mathscr{G}$. We make this idea precise next. First we will need the following Lemma.

Lemma 4 (Lemma 13.6 of [91])**.** *For $f,g \in \mathscr{G}$, define the metric $L^2(P_X)$ as*

$$\|f-g\|_{L^2(P_X)} = \left(\mathbb{E}[f(X)-g(X)]^2\right)^{1/2}.$$

Let $\mathscr{N}(\mathscr{G},L^2(P_X),\varepsilon)$ denote the covering number of $\mathscr{G}$ in the metric $L^2(P_X)$. Then, we have

$$\mathscr{N}(\mathscr{G},L^2(P_X),\varepsilon) \leq \left(\frac{e^2}{\varepsilon}\right)^{2d}. \tag{3.19}$$

Consider the metric $d(f,g) = \mathbb{P}(g(X) \neq f(X))$ for $f,g \in \mathscr{G}$. Since

$$\|f-g\|_{L^2(P_X)} = \sqrt{\mathbb{P}(g(X) \neq f(X))},$$

any $\sqrt{\varepsilon}$ covering of $\mathscr{G}$ in the $L^2(P_X)$ metric is a ε covering of $\mathscr{G}$ in the metric d. Therefore,

$$\mathscr{N}(\mathscr{G},d,\varepsilon) \leq \left(\frac{e^4}{\varepsilon}\right)^{d}. \tag{3.20}$$

We will now construct a mixture probability assignment of the form in (3.7). To do this, we must specify a distribution over the hypothesis class $\mathscr{F}$. Consider $g_1,g_2,\ldots,g_{\lfloor(e^4n)^d\rfloor}$ that form a $1/n$ covering of $\mathscr{G}$ in the metric d. By (3.20), $\lfloor(e^4n)^d\rfloor$ such functions exist. Take G,Θ_0 and Θ_1 to be independent, with $\Theta_0,\Theta_1 \sim \text{Uniform}[0,1]$ [1] and $G \sim \text{Uniform}\{g_1,\ldots,g_{\lfloor(e^4n)^d\rfloor}\}$.

[1] As mentioned in Remark 3, this corresponds to the Laplace probability assignment and we do this because it considerably simplifies the proof at just the cost of a slightly larger constant.

We then have

$$\mathbb{E}_F[p_F(y^i|x^i)] = \frac{1}{\lfloor (e^4n)^d \rfloor} \sum_{i=1}^{\lfloor (e^4n)^d \rfloor} \int_0^1 \int_0^1 p_{g_i,\theta_0,\theta_1} d\theta_0 d\theta_1 \tag{3.21}$$

which we substitute into (3.7) to construct q_{mix}. We can then prove the following.

Lemma 5. *For q_{mix} characterized by the mixture* (3.21)*, we have*

$$\max_{f^*} R_{n,P_X}(q_{\text{mix}}, f^*) \leq (d+8)\log\left(e^4 n\right) + 6. \tag{3.22}$$

3.3 Logarithmic upper bounds

In this section, we consider some special instances of the function class $\mathscr{G}$ and distributions P_X for which we can establish that the probability assignment q_{mix} in (3.7) achieves $O(d\log n)$ regret for an appropriate choice of the mixture distribution (i.e. the distribution over the class $\mathscr{F}$).

3.3.1 Finite Function Class

When $|\mathscr{G}| < \infty$, we have already shown in Lemma 2 in Section 3.2.2 that the regret is logarithmic for any distribution P_X.

3.3.2 Function Class of Halfspaces

In this subsection, we will consider the case when $\mathscr{G}$ is the class of *halfspaces*, defined precisely as follows. Let $X \in \mathscr{X} = \mathbb{S}^{d-1}$. Recall that $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Define the function $g_a(x) : \mathscr{X} \to \{-1, 1\}$ as $g_a(x) = \text{sign}\left(a^T x\right)$. The class of functions $\text{HS}_d := \{g_a, a \in \mathbb{S}^{d-1}\}$ is known as the class of $d-$dimensional (homogenous) halfspaces, and is known to have $\text{VCdim}(\mathscr{G}) = d$ [83]. Consider $X_1^n \sim \text{Uniform}(\mathbb{S}^{d-1})$ i.i.d. We will now evaluate the regret of q_{mix} in (3.7).

As in the previous section, characterizing q_{mix} requires specifying a distribution over the hypothesis class $\mathscr{F}$, which in turn requires specifying a distribution over the function class HS_d (recall that Θ_0 and Θ_1 are chosen to be $\text{Beta}(1/2, 1/2)$ independently of each other and of G). We will choose $A \sim \text{Uniform}[\mathbb{S}^{d-1}]$. We then have

$$\mathbb{E}_F[p_F(y^i|x^i)] = \mathbb{E}_A[\mathbb{E}_{\Theta_0,\Theta_1}[p_{A,\Theta_0,\Theta_1}(y^i|x^i)]] \tag{3.23}$$

Now, using the notation

$$q_{a,\text{mix}}(y^i|x^i) := \mathbb{E}_{\Theta_0,\Theta_1}[p_{a,\Theta_0,\Theta_1}(y^i|x^i)] = \int_0^1 \int_0^1 p_{a,\theta_0,\theta_1}(y^i|x^i) w(\theta_0) w(\theta_1) d\theta_0 d\theta_1 \tag{3.24}$$

for an $a \in \mathbb{S}^{d-1}$, we see that

$$\mathbb{E}_F[p_F(y^i|x^i)] = \mathbb{E}_A[q_{A,\text{mix}}(y^i|x^i)]. \tag{3.25}$$

where $A \sim \text{Uniform}[\mathbb{S}^{d-1}]$ as mentioned previously. We can make the following assertion.

Proposition 3. *If $P_X = \text{Uniform}[\mathbb{S}^{d-1}]$, then for the mixture probability assignment q_{mix} as defined in* (3.7)*, with choice of mixture as in* (3.25)*, we have*

$$\max_f R_{n,P_X}(q_{\text{mix}}, f) \leq (2d+1)\log n + d\log(48d) + \log\frac{\pi^2}{8}. \tag{3.26}$$

3.3.3 Hypothesis Class of Axis-Aligned Rectangles

In this subsection, we will consider the case when $\mathscr{G}$ is the class of axis-aligned rectangles, defined precisely as follows. For[2] $\mathbf{a} := \{a_i\}_{i=1}^d$ and $\mathbf{b} := \{b_i\}_{i=1}^d$ that are such that $0 \leq a_i \leq b_i \leq 1, i \in [d]$ define the function $g_{\mathbf{a},\mathbf{b}} : \mathbb{R}^d \to \{0,1\}$ as $g_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = \prod_{i=1}^d \mathbb{1}\{a_i \leq x_i \leq b_i\}$. Then the hypothesis class $\text{RECT}_d := \{g_{\mathbf{a},\mathbf{b}}, \mathbf{a}, \mathbf{b} \in [0,1]^d, a_i \leq b_i\}$ is known as the class of axis aligned

[2] In this subsection, for clarity we will use boldface to denote vectors.

rectangles. It is well-known that $\text{VCdim}(\text{RECT}_d) = 2d$ [83]. Consider $\mathbf{X}_1^n \sim \text{Uniform}[0,1]^d$ iid. We will then evaluate the regret of the probability assignment q_{mix} in (3.7).

As before, characterizing q_{mix} requires specifying a distribution over the hypothesis class $\mathscr{F}$, which in turn requires specifying a distribution over the function class RECT_d (recall that Θ_0 and Θ_1 are chosen to be $\text{Beta}(1/2, 1/2)$ independently of each other and of G). We will chose $(A_i, B_i) \sim \text{Uniform}\{(a,b) \in [0,1] \times [0,1], b \geq a\}$, and $(A_i, B_i) \perp\!\!\!\perp (A_j, B_j)$ for $i \neq j$. Denoting $\mathbf{A} := (A_1, \ldots, A_d)$ and $\mathbf{B} := (B_1, \ldots, B_d)$, for the aforementioned choice of distribution over $\mathscr{F}$, we have

$$\mathbb{E}_F[p_F(y^i|\mathbf{x}^i)] = \mathbb{E}_{\mathbf{A},\mathbf{B}}[\mathbb{E}_{\Theta_0,\Theta_1}[p_{\mathbf{A},\mathbf{B},\Theta_0,\Theta_1}(y^i|\mathbf{x}^i)]] \tag{3.27}$$

Now, using the notation

$$\begin{aligned} &q_{\mathbf{a},\mathbf{b},\text{mix}}(y^i|\mathbf{x}^i) \\ &\quad := \mathbb{E}_{\Theta_0,\Theta_1}[p_{\mathbf{a},\mathbf{b},\Theta_0,\Theta_1}(y^i|\mathbf{x}^i)] = \int_0^1 \int_0^1 p_{\mathbf{a},\mathbf{b},\theta_0,\theta_1}(y^i|\mathbf{x}^i) w(\theta_0) w(\theta_1) d\theta_0 d\theta_1 \end{aligned} \tag{3.28}$$

we see that

$$\mathbb{E}_F[p_F(y^i|\mathbf{x}^i)] = \mathbb{E}_{\mathbf{A},\mathbf{B}}[q_{\mathbf{A},\mathbf{B},\text{mix}}(y^i|\mathbf{x}^i)]. \tag{3.29}$$

We can then make the following assertion.

Proposition 4. *If $P_X = \text{Uniform}[0,1]^d$, then for the probability assignment q_{mix} as defined in* (3.7)*, with choice of mixture as in* (3.29)*, we have*

$$\max_{f \in \mathscr{F}} R_{n,P_X}(q_{\text{mix}}, f) \leq (2d+1)\log(n+1) + \log\frac{\pi^2}{8}. \tag{3.30}$$

Remark 4. *In Sections 3.3.2 and 3.3.3, we have fixed P_X to be the uniform distribution. Considering the proofs, it appears to be a reasonable guess that the mixture probability assignment*

q_{mix} employed to prove the regret guarantees would work for other distributions P_X that are sufficiently "smooth". Thus, finding non-uniform P_X for which the given q_{mix} achieves logarithmic regret is an intriguing question.

3.4 Proof of Theorem 1

In this section, we prove Theorem 1. To motivate the main proof idea, recall the case discussed in Section 3.2.3 when noncausal side information is available. In that case, using the Sauer–Shelah lemma we argued that given X^n, the (possibly infinite) class of functions $\mathscr{G}$ could be effectively reduced to a class of at most $\left(\frac{en}{d}\right)^d$ functions, and using the mixture probability assignment that took a uniform mixture over these functions yielded an $O(d\log n)$ regret. This leads to us considering the following alternative to noncausal side information being available: what if *another* sequence $\widetilde{X}^n \overset{(d)}{=} X^n$ is available noncausally? The sequence $\widetilde{X}^n$ also reduces the class $\mathscr{G}$ to at most $\left(\frac{en}{d}\right)^d$ functions (albeit not the same reduction as that of $\mathscr{G}$ by X^n). We establish in Section 3.4.1 that a uniform mixture over the finite reduction of $\mathscr{G}$ induced by $\widetilde{X}^n$ achieves an $O(\sqrt{dn}\log n)$ regret. We then use this result in Section 3.4.2 to establish a general $O(\sqrt{nd}\log n)$ regret when the side information X^n is available sequentially.

For clarity, throughout this section we use $g(Z^n)$ to denote $(g(Z_1),\ldots,g(Z_n)) \in \{0,1\}^n$.

3.4.1 Sequence $\widetilde{X}^n$ available noncausally

Consider a sequence $\widetilde{X}^n \overset{(d)}{=} X^n, \widetilde{X}^n \perp\!\!\!\perp X^n$, with X_i having distribution P_X iid. In this subsection we consider the regret

$$\widetilde{R}_{n,P_X}(q,f) := \mathbb{E}_{X^n,Y^n,\widetilde{X}^n}\left[\sum_{i=1}^{n}\log\frac{1}{q(Y_i|X^i,Y^{i-1},\widetilde{X}^n)} - \sum_{i=1}^{n}\log\frac{1}{p_f(Y_i|X_i)}\right] \tag{3.31}$$

and in particular the worst-case regret attained by a probability assignment q

$$\widetilde{R}_n(q) := \max_{f\in\mathscr{F},P_X} \widetilde{R}_{n,P_X}(q,f). \tag{3.32}$$

Now, using the same notation as in Section 3.2.3, let $\mathscr{P}_n(\widetilde{X}^n) = \{g(\widetilde{X}^n), g \in \mathscr{G}\} \subseteq \{0,1\}^n$ with $|\mathscr{P}_n(\widetilde{X}^n)| \leq \left(\frac{en}{d}\right)^d$ by the Sauer–Shelah lemma. Pick $\widetilde{g}_1, \ldots, \widetilde{g}_{|\mathscr{P}_n(\widetilde{X}^n)|} \in \mathscr{G}$ such that $\widetilde{g}_j(\widetilde{X}^n) \in \mathscr{P}_n(\widetilde{X}^n)$, and $\widetilde{g}_i(\widetilde{X}^n) \neq \widetilde{g}_j(\widetilde{X}^n)$ if $i \neq j$. Thus, for every $g \in \mathscr{G}$, there exists a $j \in \left[|\mathscr{P}_n(\widetilde{X}^n)|\right]$ such that $g(\widetilde{X}^n) = \widetilde{g}_j(\widetilde{X}^n)$. Therefore, the class $\mathscr{G}$ has been effectively reduced to $|\mathscr{P}_n(\widetilde{X}^n)|$ functions by $\widetilde{X}^n$. Consider now a mixture probability assignment, akin to (3.7), as

$$\widetilde{q}_{\text{mix}}(y_i|x^i, y^{i-1}, \widetilde{x}^n) := \frac{\frac{1}{|\mathscr{P}(\widetilde{x}^n)|}\sum_{j=1}^{|\mathscr{P}(\widetilde{x}^n)|}\int_0^1\int_0^1 p_{\widetilde{g}_j,\theta_0,\theta_1}(y^i|x^i)d\theta_0 d\theta_1}{\frac{1}{|\mathscr{P}(\widetilde{x}^n)|}\sum_{j=1}^{|\mathscr{P}(\widetilde{x}^n)|}\int_0^1\int_0^1 p_{\widetilde{g}_j,\theta_0,\theta_1}(y^{i-1}|x^{i-1})d\theta_0 d\theta_1}. \tag{3.33}$$

Note that this is indeed a mixture probability assignment in the sense of (3.7)—$F = (G, \Theta_0, \Theta_1)$ has the distribution where $G \sim \text{Uniform}\{\widetilde{g}_1, \ldots, \widetilde{g}_{|\mathscr{P}_n(\widetilde{X}^n)|}\}$, $\Theta_0, \Theta_1 \sim \text{Uniform}[0,1]$[3] and G, Θ_0, and Θ_1 are mutually independent. We can now state the following.

Lemma 6. *For $\widetilde{q}_{\text{mix}}$ defined in* (3.33)*, we have for an absolute constant $C \leq 250$,*

$$\widetilde{R}_n(\widetilde{q}_{\text{mix}}) \leq d\log(en/d) + 16C\sqrt{nd}\log(6n+2) \tag{3.34}$$

and moreover, for any P_X and h we have

$$\sum_{i=1}^{n}\log\frac{1}{\widetilde{q}_{\text{mix}}(Y_i|X^i, Y^{i-1}, \widetilde{X}^n)} - \sum_{i=1}^{n}\log\frac{1}{p_f(Y_i|X_i)} \\ \leq d\log(en/d) + 16\sqrt{n}\log(6n+2)\left(C\sqrt{d} + \sqrt{2\log\frac{2}{\delta}}\right). \tag{3.35}$$

Remark 5 (Empirical covering)**.** *The probability assignment $\widetilde{q}_{\text{mix}}$ can also be motivated by considering the scenario in Section 3.2.4 where P_X is known. Recall that there, we took a uniform mixture over a $1/n$-covering of $\mathscr{G}$ in the metric d with $d(g_1, g_2) = \mathbb{P}(g_1(X) \neq g_2(X))$. If we have $\widetilde{X}^n$, as an alternative to a mixture over a covering in the metric d, we can take an empirical $1/n$ covering of $\mathscr{G}$, i.e. a covering in the metric $\tilde{\Delta}_n(g_1, g_2) = \frac{1}{n}d_H(g_1(\widetilde{X}^n), g_2(\widetilde{X}^n))$. Indeed, the*

[3] The choice of taking a uniform prior for Θ_0 and Θ_1 instead of the Jeffreys prior is simply because using the uniform prior (which, recalling Remark 3, corresponds to the Laplace probability assignment) makes some calculations far simpler in the proof at just the cost of a worse constant factor in the regret.

functions $\widetilde{g}_1, \widetilde{g}_2, \ldots, \widetilde{g}_{|\mathscr{P}_n(\widetilde{X}^n)|}$ form not just a $1/n$ covering but a 0-covering of $\mathscr{G}$ in the metric $\tilde{\Delta}_n$.

3.4.2 Epoch-based mixture probability

In this subsection, we use Lemma 6 to construct a general probability assignment when side information is available sequentially. In this scenario, we do not have access to another sequence $\widetilde{X}^n$. However, at time step $i+1$, we have access to the past sequence X^i which could be used, as done in [84], in lieu of $\widetilde{X}^i$. We now precisely define and analyze this probability assignment.

For simplicity, assume that $n = 2^k$ for some integer k. The analysis is easily extended to any arbitrary n. We will split the n time steps into $\log n$ "epochs". Starting from $j = 1$, define the $j-$the epoch to consist of the time steps $2^{j-1} + 1 \leq i \leq 2^j$. So, the first epoch consists of X_2, the second epoch consists of X_3^4, the third epoch consists of X_5^8 and so on. Consider the the following probability assignment q^*.

1. $q^*(Y_1|X_1) = 1/2$
2. For $i \geq 2$, if $2^{j-1} + 1 \leq i \leq 2^j$, i.e. if the time step i falls within the $j-$th epoch, then

$$q^*(Y_i|X^i, Y^{i-1}) = \frac{q_{\text{mix},j}(Y_{2^{j-1}+1}^i|X_{2^{j-1}+1}^i)}{q_{\text{mix},j}(Y_{2^{j-1}+1}^{i-1}|X_{2^{j-1}+1}^{i-1})} \tag{3.36}$$

   where

$$\begin{aligned} & q_{\text{mix},j}((Y_{2^{j-1}+1}^i|X_{2^{j-1}+1}^i)) \\ & := \frac{1}{|\mathscr{P}(X^{2^{j-1}})|} \sum_{k=1}^{|\mathscr{P}(X^{2^{j-1}})|} \int_0^1 \int_0^1 p_{\theta_0,\theta_1,g_k}(Y_{2^{j-1}+1}^i|X_{2^{j-1}+1}^i) d\theta_0 d\theta_1 \end{aligned} \tag{3.37}$$

   is the finite mixture over the $|\mathscr{P}(X^{2^{j-1}})|$ partition of $\mathscr{G}$ induced by $X^{2^{j-1}}$. This is the same probability assignment as in (3.33).

Using Lemma 6 and an epoch-wise analysis of q^* we can establish Theorem 1.

3.5 Proof of Theorem 2

In this section we prove Theorem 2. A key component of the proof is the redundancy-capacity theorem [92].

First, note that the class of probability assignments that utilize the side information X^n causally is a subset of the set of probability assignments that utilize the side information X^n noncausally. This implies

$$\begin{aligned} R_n &= \min_{q} \max_{P_X, f} \mathbb{E}_{X^n,Y^n}\left[\sum_{i=1}^{n} \log \frac{1}{q(Y_i|X^i,Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)}\right] \\ &\geq \min_{q} \max_{P_X, f} \mathbb{E}_{X^n,Y^n}\left[\sum_{i=1}^{n} \log \frac{1}{q(Y_i|X^n,Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)}\right] \qquad (3.38) \end{aligned}$$

and therefore

$$\begin{aligned} R_n &\geq \min_{q} \max_{P_X, f} \mathbb{E}_{X^n,Y^n}\left[\log \frac{p_f(Y^n|X^n)}{q(Y^n|X^n)}\right] \\ &= \min_{q} \max_{P_X, P_F} \mathbb{E}_{F,X^n,Y^n}\left[\log \frac{p_F(Y^n|X^n)}{q(Y^n|X^n)}\right] \qquad (3.39) \\ &\geq \max_{P_X, P_F} \min_{q} \mathbb{E}_{F,X^n,Y^n}\left[\log \frac{p_F(Y^n|X^n)}{q(Y^n|X^n)}\right] \qquad (3.40) \end{aligned}$$

where P_F denotes a distribution over $\mathscr{F}$ in (3.39), and (3.40) follows since

$$\min\max(\cdot) \geq \max\min(\cdot).$$

By a conditional variant of the redundancy-capacity theorem we have for a fixed P_X and P_F (recall that $F = (\Theta_0, \Theta_1, G)$)

$$\min_{q} \mathbb{E}_{F,X^n,Y^n}\left[\log \frac{p_F(Y^n|X^n)}{q(Y^n|X^n)}\right] = I(F;Y^n|X^n) \qquad (3.41)$$

and so

$$R_n \geq \max_{P_X, P_F} I(F; Y^n | X^n) \tag{3.42}$$

where recall $F = (\Theta_0, \Theta_1, G)$.

Remark 6. *The result in* (3.42) *holds for any class of conditional distributions $\mathscr{F}$, not just the VC class under consideration.*

We will first provide a lower bound on R_n when $|\mathscr{X}| < \infty$ which we will then use to achieve a general lower bound on R_n.

Lemma 7. *If $|\mathscr{X}| = m < \infty$ and $\mathscr{G} = 2^{[m]}$ so that $|\mathscr{G}| = 2^m$, we have*

$$R_n \geq m + \log(n+1) - \log(\pi e) - 2\sqrt{e} m^2 e^{-3n/100m}. \tag{3.43}$$

Lemma 7 is proved by choosing a particular P_X and P_F and analyzing the right hand side of (3.42).

Remark 7 (Tightness for finite $\mathscr{X}$)**.** *Combining Lemma 7 and Lemma 2 with $|\mathscr{G}| = 2^m$, we see that for $\mathscr{X} = m, \mathscr{G} = 2^{[m]}$, we can obtain a tight characterization of the regret R_n on n and m.*

Consider now the case when $\mathscr{X}$ is possibly infinite. Since $\mathrm{VCdim}(\mathscr{G}) = d$, there exist $x_1, \ldots, x_d \in \mathscr{X}$ such that $|\{(g(x_1), \ldots, g(x_d)), g \in \mathscr{G}\}| = 2^d$. Theorem 2 then follows as a corollary to Lemma 7 by substituting $m = d$ and choosing the distributions of P_X, P_F as in the proof of Lemma 7.

3.6 Discussion

We considered the problem of sequential prediction under log-loss with side information. This can be considered as an extension of the well-studied information-theoretic problem of universal compression of an i.i.d. binary source, and the regret incurred can be characterized via the value of a minmax game. We provided upper bounds on the regret via construction of a probability assignment, and lower bounds by the redundancy-capacity theorem. There are several open directions. Previous results established an $O(d\log n)$ upper bound via minmax duality. Even though our upper and lower bounds are off by a $\sqrt{n}$ factor, we suspect that a variant of the mixture probability assignment from information theory can achieve the optimal $O(d\log n)$ upper bound. We provided some special cases and a probability assignment where $O(d\log n)$ redundancy is achieved to provide evidence for this. Recently, [93] provided a mixture based probability assignment, and used a similar epoch-based covering method to achieve $O(d\log^2 n)$ regret for this problem. It would also be interesting to answer the weaker question of whether the current upper bound on R_n can be improved upon (constructively) under certain further restrictions on the class of functions $\mathscr{G}$. Moreover, even though the lower bound cannot be improved in order, it may be possible to get a better dependence on d. Finally, we have not considered complexity concerns for actual algorithmic implementation. Computing the coverings may be probihitively expensive in several cases, so finding efficient algorithms for sequential probability assignment is yet another avenue to be explored. All these directions are promising for further study.

3.7 Skipped proofs

3.7.1 Skipped Proofs from Section 3.2

Proof of Proposition 1

$$\begin{aligned}\sum_{y_i\in\mathscr{Y}} q_{\text{mix}}(y_i|x^i,y^{i-1}) &= \frac{\sum_{y_i\in\mathscr{Y}}\mathbb{E}[p_F(y^i|x^i)]}{\mathbb{E}[p_F(y^{i-1}|x^{i-1}))} \\ &= \frac{\mathbb{E}[\sum_{y_i\in\mathscr{Y}} p_F(y^i|x^i)]}{\mathbb{E}[p_F(y^{i-1}|x^{i-1}))}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathbb{E}[p_F(y^{i-1}|x^{i-1})\sum_{y_i\in\mathscr{Y}} p_F(y_i|x_i)]}{\mathbb{E}[p_F(y^{i-1}|x^{i-1}))} \\
&= \frac{\mathbb{E}[p_F(y^{i-1}|x^{i-1})]}{\mathbb{E}[p_F(y^{i-1}|x^{i-1}))} = 1
\end{aligned}
$$

and so $q_{\text{mix}}(y_i|x^i,y^{i-1})$ is a valid probability assignment.

Proof of Lemma 1

Since the function g^* is known and the range of g^* is only 0 and 1, we can assume without loss of generality that the side information is binary, i.e. $\mathscr{X}=\{0,1\}$. Now define

$$
n_l := \sum_{i=1}^{n} \mathbb{1}\{x_i = l\}, l \in \{0,1\} \tag{3.44}
$$

$$
k_l := \sum_{i=1}^{n} \mathbb{1}\{y_i = 1, x_i = l\}, l \in \{0,1\}. \tag{3.45}
$$

Note that

$$
\begin{aligned}
&\log \frac{p_{\theta_0,\theta_1,g^*}(Y^n|X^n)}{\int_0^1\int_0^1 p_{\theta_0,\theta_1,g^*}(y^n|x^n)w(\theta_0)w(\theta_1)d\theta_0} \\
&\quad = \sum_{i=1}^{n} \log \frac{1}{q_{\text{KT}}(Y_i|X^i,Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)}
\end{aligned} \tag{3.46}
$$

and

$$
\begin{aligned}
&\sum_{i=1}^{n} \log \frac{1}{q_{\text{KT}}(Y_i|X^i,Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)} \\
&\qquad = \sum_{l=0}^{1} \sum_{i:X_i=l} \left[\log \frac{1}{q_{\text{KT}}(Y_i|X^i,Y^{i-1})} - \log \frac{1}{p_f(Y_i|X_i)} \right] \\
&\qquad = \sum_{l=0}^{1} \left[\log \frac{\prod_{i:X_i=l} p_f(Y_i|X_i)}{\prod_{i:X_i=l} q_{\text{KT}}(Y_i|X^i,Y^{i-1})} \right].
\end{aligned} \tag{3.47}
$$

Now, if $n_l = 0$, we have

$$\log \frac{\prod_{i:X_i=l} p_f(Y_i|X_i)}{\prod_{i:X_i=l} q_{\mathrm{KT}}(Y_i|X^i,Y^{i-1})} = 0 \tag{3.48}$$

And if $n_l \geq 1$, we have

$$\log \frac{\prod_{i:X_i=l} p_f(Y_i|X_i)}{\prod_{i:X_i=l} q_{\mathrm{KT}}(Y_i|X^i,Y^{i-1})} = \log \frac{\theta_l^{k_l}(1-\theta_l)^{n_l-k_l}}{\frac{1}{4^{n_l}}\frac{\binom{n_l}{k_l}\binom{2n_l}{n_l}}{\binom{2n_l}{2k_l}}} \tag{3.49}$$

where (3.49) follows from properties of the KT sequential probability assignment. Moreover, we have

$$\theta_l^{k_l}(1-\theta_l)^{n_l-k_l} \leq \left(\frac{k_l}{n_l}\right)^{k_l}\left(1-\frac{k_l}{n_l}\right)^{n_l-k_l} = 2^{-n_l h\left(\frac{k_l}{n_l}\right)} \tag{3.50}$$

which can be established by noting that the binary KL divergence

$$d\left(\frac{k_l}{n_l}||\theta_l\right) = \frac{1}{n_l}\log\frac{\left(\frac{k_l}{n_l}\right)^{k_l}\left(1-\frac{k_l}{n_l}\right)^{n_l-k_l}}{\theta_{g(l)}^{k_l}(1-\theta_{g(l)})^{n_l-k_l}} \geq 0,$$

and furthermore, using a Sterling approximation we can establish

$$\frac{1}{4^{n_l}}\frac{\binom{n_l}{k_l}\binom{2n_l}{n_l}}{\binom{2n_l}{2k_l}} \leq \sqrt{\frac{8}{\pi^2}}\frac{2^{-n_l h\left(\frac{k_l}{n_l}\right)}}{\sqrt{n_l}}. \tag{3.51}$$

Plugging (3.50) and (3.51) into (3.49) yields

$$\log \frac{\prod_{i:X_i=l} p_f(Y_i|X_i)}{\prod_{i:X_i=l} q_{\mathrm{KT}}(Y_i|X^i,Y^{i-1})} \leq \log\sqrt{\frac{\pi^2}{8}}\frac{2^{-n_l h\left(\frac{k_l}{n_l}\right)}\sqrt{n_l}}{2^{-n_l h\left(\frac{k_l}{n_l}\right)}} = \frac{1}{2}\log n_l + \frac{1}{2}\log\frac{\pi^2}{8} \tag{3.52}$$

when $n_l \geq 1$. Combining (3.48) and (3.52) we can establish

$$\log \frac{\prod_{i:X_i=l} p_f(Y_i|X_i)}{\prod_{i:X_i=l} q_{\text{KT}}(Y_i|X^i, Y^{i-1})} \leq \frac{1}{2}\log(n_l+1) + \frac{1}{2}\log\frac{\pi^2}{8} \tag{3.53}$$

for all $n_l \geq 0$. Plugging the upper bound (3.53) into (3.47) yields

$$\begin{aligned}
\sum_{l=0}^{1}\left[\log \frac{\prod_{i:X_i=l} p_f(Y_i|X_i)}{\prod_{i:X_i=l} q_{\text{KT}}(Y_i|X^i, Y^{i-1})}\right] &\leq \sum_{l=0}^{1}\frac{1}{2}\log(n_l+1) + \frac{1}{2}\log\frac{\pi^2}{8} \\
&= \frac{1}{2}\log\prod_{l=0}^{1}(n_l+1) + \log\frac{\pi^2}{8} \\
&\leq \frac{1}{2}\log\left(\frac{n}{2}+1\right)^2 + \log\frac{\pi^2}{8} \qquad (3.54) \\
&= \log\left(\frac{n}{2}+1\right) + \log\frac{\pi^2}{8} \qquad (3.55)
\end{aligned}$$

where the inequality (3.54) follows by noting that $\sum_{l=0}^{m-1}(n_l+1) = n+m$ and the using the AM-GM inequality. We have now established

$$\sum_{i=1}^{n}\log\frac{1}{q_{\text{KT}}(Y_i|X^i, Y^{i-1})} - \sum_{i=1}^{n}\log\frac{1}{p_f(Y_i|X_i)} \leq \log\left(\frac{n}{2}+1\right) + \log\frac{\pi^2}{8}$$

and monotonicity of expectation followed by taking supremum over θ_0, θ_1 then yields the result.

Proof of Lemma 2

By Proposition 2, we see that for a fixed $f^* = (g^*, \theta_0^*, \theta_1^*)$ and P_X, the regret achieved by the probability assignment q_{mix} characterized by (3.13) is

$$R_{n,P_X}(q_{\text{mix}}, f^*) = \mathbb{E}_{X^n,Y^n}\left[\log\frac{p_{f^*}(Y^n|X^n)}{\mathbb{E}_F\left[p_F(Y^n|X^n)\right]}\right]$$

and we have

$$\mathbb{E}_{X^n,Y^n}\left[\log\frac{p_{f^*}(Y^n|X^n)}{\mathbb{E}_F\left[p_F(Y^n|X^n)\right]}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\log \frac{p_{f^*}(Y^n|X^n)}{\frac{1}{|\mathscr{G}|}\sum_{g\in\mathscr{G}}\int_0^1\int_0^1 p_{g,\theta_0,\theta_1}(Y^n|X^n)w(\theta_0)w(\theta_1)d\theta_0 d\theta_1}\right] \\
&= \log|\mathscr{G}| + \mathbb{E}\left[\log \frac{p_{f^*}(Y^n|X^n)}{\sum_{g\in\mathscr{G}}\int_0^1\int_0^1 p_{g,\theta_0,\theta_1}(Y^n|X^n)w(\theta_0)w(\theta_1)d\theta_0 d\theta_1}\right] \\
&\le \log|\mathscr{G}| + \mathbb{E}\left[\log \frac{p_{f^*}(Y^n|X^n)}{\int_0^1\int_0^1 p_{g^*,\theta_0,\theta_1}(Y^n|X^n)w(\theta_0)w(\theta_1)d\theta_0 d\theta_1}\right] \qquad (3.56) \\
&\le \log|\mathscr{G}| + \log\left(\frac{n}{2}+1\right) + \log\frac{\pi^2}{8} \qquad (3.57)
\end{aligned}
$$

where (3.56) follows since each of the summands in the denominator of the second term are nonnegative, and (3.57) is a consequence of Lemma 1.

Proof of Lemma 3

Following the proof of Lemma 2 up to (3.56), for any fixed $f^* = (g^*, \theta_0^*, \theta_1^*)$ the probability assignment q_{mix} characterized by the mixture (3.17) has

$$
\begin{aligned}
&\mathbb{E}_{X^n,Y^n}\left[\sum_{i=1}^n \log\frac{1}{q_{\text{mix}}(Y_i|X^n,Y^{i-1})} - \sum_{i=1}^n \log\frac{1}{p_{f^*}(Y_i|X_i)}\right] \\
&\qquad \le \mathbb{E}[|\mathscr{P}_n|] + \mathbb{E}\left[\log \frac{p_{f^*}(Y^n|X^n)}{\int_0^1\int_0^1 p_{g_j,\theta_0,\theta_1}(Y^n|X^n)w(\theta_0)w(\theta_1)d\theta_0 d\theta_1}\right] \qquad (3.58)
\end{aligned}
$$

Where $j \in [|\mathscr{P}_n|]$ is such that

$$(g^*(X_1),\ldots,g^*(X_n)) = (g_j(X_1),\ldots,g_j(X_n)).$$

If VCdim$(\mathscr{G}) = d < \infty$, we can control $|\mathscr{P}_n|$ using the following standard result [94, Chapter 8].

Lemma 8 (Sauer–Shelah)**.** *If VCdim($\mathscr{G}$) $= d < \infty$, then* $|\mathscr{P}_n| \le \left(\frac{en}{d}\right)^d$.

Finally, using Lemma 1 and Lemma 8 in (3.58) yields

$$R_{n,\text{nc}}(q_{\text{mix}}) \le d\log(en/d) + \log\left(\frac{n}{2}+1\right) + \log\frac{\pi^2}{8}. \qquad (3.59)$$

Proof of Lemma 5

By Proposition 2, we have for a fixed $f^* = (g^*, \theta_0^*, \theta_1^*)$

$$
\begin{aligned}
R_{n,P_X}(q_{\text{mix}}, f^*) &= \mathbb{E}\left[\frac{p_{f^*}(Y^n|X^n)}{\mathbb{E}_H[p_F(Y^n|X^n)]}\right] \\
&= \mathbb{E}\left[\log \frac{p_{f^*}(Y^n|X^n)}{\frac{1}{\lfloor (e^4 n)^d \rfloor} \sum_{i=1}^{\lfloor (e^4 n)^d \rfloor} \int_0^1 \int_0^1 p_{g_i,\theta_0,\theta_1}(Y^n|X^n) d\theta_0 d\theta_1}\right] \\
&\leq d \log(e^4 n) + \mathbb{E}\left[\log \frac{p_{f^*}(Y^n|X^n)}{\sum_{i=1}^{\lfloor (e^4 n)^d \rfloor} \int_0^1 \int_0^1 p_{g_i,\theta_0,\theta_1}(Y^n|X^n) d\theta_0 d\theta_1}\right] \qquad (3.60)
\end{aligned}
$$

Let $\widetilde{g} \in \{g_1, g_2, \ldots, g_{\lfloor (e^4 n)^d \rfloor}\}$ be such that $\mathbb{P}(\widetilde{g}(X) \neq g^*(X)) = d(\widetilde{g}, g^*) \leq 1/n$. Such a $\widetilde{g}$ exists since $g_1, \ldots, g_{\lfloor (e^4 n)^d \rfloor}$ form a $1/n$ covering of $\mathscr{G}$ in the metric d. We then have from (3.60)

$$
R_{n,P_X}(q_{\text{mix}}, f^*) \leq d \log(e^4 n) + \mathbb{E}\left[\log \frac{p_{f^*}(Y^n|X^n)}{\int_0^1 \int_0^1 p_{\widetilde{g},\theta_0,\theta_1}(Y^n|X^n) d\theta_0 d\theta_1}\right]. \qquad (3.61)
$$

Now, defining

$$
\begin{aligned}
N_j &:= \sum_{i=1}^{n} \mathbb{1}\{g^*(X_i) = j\}, j \in \{0,1\} \\
K_j &:= \sum_{i=1}^{n} \mathbb{1}\{g^*(X_i) = j, Y_i = 1\}, j \in \{0,1\} \\
\widetilde{N}_j &:= \sum_{i=1}^{n} \mathbb{1}\{\widetilde{g}(X_i) = j\}, j \in \{0,1\} \\
\widetilde{K}_j &:= \sum_{i=1}^{n} \mathbb{1}\{\widetilde{g}(X_i) = j, Y_i = 1\}, j \in \{0,1\}
\end{aligned}
$$

we have

$$
p_{f^*}(Y^n|X^n) = p_{g^*,\theta_0^*,\theta_1^*}(Y^n|X^n) = \theta_0^{*K_0}(1-\theta_0^*)^{N_0-K_0}\theta_1^{*K_1}(1-\theta_1^*)^{N_1-K_1} \qquad (3.62)
$$

and

$$\int_0^1 \int_0^1 p_{\widetilde{g},\theta_0,\theta_1}(Y^n|X^n) d\theta_0 d\theta_1 = \int_0^1 \int_0^1 \theta_0^{\widetilde{K}_0}(1-\theta_0)^{\widetilde{N}_0-\widetilde{K}_0} \theta_1^{\widetilde{K}_1}(1-\theta_1)^{\widetilde{N}_1-\widetilde{K}_1} d\theta_0 d\theta_1$$
$$= \int_0^1 \theta_0^{\widetilde{K}_0}(1-\theta_0)^{\widetilde{N}_0-\widetilde{K}_0} d\theta_0 \int_0^1 \theta_1^{\widetilde{K}_1}(1-\theta_1)^{\widetilde{N}_1-\widetilde{K}_1} d\theta_1$$
$$= \frac{1}{(\widetilde{N}_0+1)\binom{\widetilde{N}_0}{\widetilde{K}_0}(\widetilde{N}_1+1)\binom{\widetilde{N}_1}{\widetilde{K}_1}} \tag{3.63}$$

where (3.63) follows from properties of the Laplace probability assignment. Now, from (3.62) and (3.63), we have

$$\frac{p_{f^*}(Y^n|X^n)}{\int_0^1 \int_0^1 p_{\widetilde{g},\theta_0,\theta_1}(Y^n|X^n) d\theta_0 d\theta_1}$$
$$= (\widetilde{N}_0+1)(\widetilde{N}_1+1)\binom{\widetilde{N}_0}{\widetilde{K}_0} \theta_0^{*K_0}(1-\theta_0^*)^{N_0-K_0} \binom{\widetilde{N}_1}{\widetilde{K}_1} \theta_1^{*K_1}(1-\theta_1^*)^{N_1-K_1}$$
$$\leq (n+1)^2 \binom{\widetilde{N}_0}{\widetilde{K}_0} \theta_0^{*K_0}(1-\theta_0^*)^{N_0-K_0} \binom{\widetilde{N}_1}{\widetilde{K}_1} \theta_1^{*K_1}(1-\theta_1^*)^{N_1-K_1} \tag{3.64}$$
$$\leq (n+1)^2 \frac{\binom{\widetilde{N}_0}{\widetilde{K}_0}}{\binom{N_0}{K_0}} \frac{\binom{\widetilde{N}_1}{\widetilde{K}_1}}{\binom{N_1}{K_1}} \tag{3.65}$$

where (3.64) follows because $\widetilde{N}_0, \widetilde{N}_1 \leq n$, and (3.65) follows since $\binom{n}{k} x^k (1-x)^{n-k} \leq 1$ for any $x \in [0,1]$. Substituting (3.65) into (3.61) yields

$$R_{n,P_X}(q_{\text{mix}}, f^*) \leq d \log(e^4 n) + \mathbb{E}_{X^n,Y^n}\left[\log \frac{\binom{\widetilde{N}_0}{\widetilde{K}_0}}{\binom{N_0}{K_0}}\right] + \mathbb{E}_{X^n,Y^n}\left[\log \frac{\binom{\widetilde{N}_1}{\widetilde{K}_1}}{\binom{N_1}{K_1}}\right] \tag{3.66}$$

Recall that $d_H(\cdot,\cdot)$ denotes the Hamming distance. We can then easily verify the following proposition.

Proposition 5. *We have*

$$|\widetilde{N}_j - N_j| \leq d_H(g^*(X^n), \widetilde{g}(X^n)), j \in \{0,1\}$$

and

$$|\widetilde{K}_j - K_j| \leq d_H(g^*(X^n), \widetilde{g}(X^n)), j \in \{0,1\}.$$

We now wish to use Proposition 5, to obtain a bound on $\log \frac{\binom{\widetilde{N}_0}{\widetilde{K}_0}}{\binom{N_0}{K_0}}$. For this, we will need an additional proposition.

Proposition 6. *For any two nonnegative integers a, b, we have*

$$\log \frac{(a+b)!}{a!} \leq b \log(a+b+1) + b + 1 \tag{3.67}$$

Proof. By the Stirling approximation, for any positive integer m, we have

$$\sqrt{2\pi} m^{m+1/2} e^{-m} \leq m! \leq e m^{m+1/2} e^{-m}. \tag{3.68}$$

We now use this to claim that when $a, b \geq 1$

$$\begin{aligned}
\ln(a+b)! - \ln a! &\leq \ln\left(e(a+b)^{a+b+1/2} e^{-(a+b)}\right) - \ln\left(\sqrt{2\pi} a^{a+1/2} e^{-a}\right) \\
&= \ln \frac{e}{\sqrt{2\pi}} + (a+b+1/2)\ln(a+b) - (a+1/2)\ln a + a - (a+b) \\
&= \ln \frac{e}{\sqrt{2\pi}} + b\ln(a+b) + (a+1/2)\ln(1+b/a) - b \\
&\leq \ln \frac{e}{\sqrt{2\pi}} + b\ln(a+b) + (a+1/2)\ln(1+b/a) - b \\
&\leq \ln \frac{e}{\sqrt{2\pi}} + b\ln(a+b) + b/2a \qquad (3.69) \\
&\leq \ln \frac{e}{\sqrt{2\pi}} + b\ln(a+b) + b/2 \qquad (3.70)
\end{aligned}$$

where (3.69) follows since for $x \geq 0, \ln(1+x) \leq x$ and (3.70) follows since $a \geq 1$. When $a = b = 0$ and when $b = 0, a \geq 1$, the proposition is immediate. Finally, when $a = 0$ and $b \geq 1$, we have by the upper bound on $b!$ in (3.68) that

$$\ln b! \leq \left(b + \frac{1}{2}\right) \ln b + 1 - b \tag{3.71}$$

and after some algebraic manipulations we can see that the proposition holds in this case as well. □

For convenience, define $\delta_n := d_H(g^*(X^n), \widetilde{g}(X^n))$. Note that

$$\begin{aligned}
\log \frac{\binom{\widetilde{N}_0}{\widetilde{K}_0}}{\binom{N_0}{K_0}} &= \log \frac{\widetilde{N}_0! K_0! (N_0 - K_0)!}{N_0! \widetilde{K}_0! (\widetilde{N}_0 - \widetilde{K}_0)!} \\
&= \log \frac{\widetilde{N}_0!}{N_0!} + \log \frac{K_0!}{\widetilde{K}_0!} + \log \frac{(N_0 - K_0)!}{(\widetilde{N}_0 - \widetilde{K}_0)!}. \qquad (3.72)
\end{aligned}$$

We will now bound each of the three terms in the RHS of (3.72). We have

$$\begin{aligned}
\log \frac{\widetilde{N}_0!}{N_0!} &\leq \log \frac{(N_0 + \delta_n)!}{N_0!} && (3.73) \\
&\leq \delta_n \log(N_0 + \delta_n + 1) + \delta_n + 1 && (3.74) \\
&\leq \delta_n \log(2n + 1) + \delta_n + 1 && (3.75)
\end{aligned}$$

where (3.73) follows from Proposition 5, (3.74) from Proposition 6 and (3.75) since $N_0, \delta_n \leq n$. Using the same reasoning, we conclude

$$\log \frac{K_0!}{\widetilde{K}_0!} \leq \delta_n \log(2n + 1) + \delta_n + 1. \tag{3.76}$$

and

$$\log \frac{(N_0 - K_0)!}{(\widetilde{N}_0 - \widetilde{K}_0)!} \leq 2\delta_n \log(3n + 1) + 2\delta_n + 1 \tag{3.77}$$

where in (3.77) we additionally use the fact that $|(N_0 - K_0) - (\widetilde{N}_0 - \widetilde{K}_0)| \leq |N_0 - \widetilde{N}_0| + |K_0 - \widetilde{K}_0| \leq$

$2\delta_n$. Substituting (3.75)— (3.77) into (3.72) yields

$$\log \frac{\binom{\widetilde{N}_0}{\widetilde{K}_0}}{\binom{N_0}{K_0}} \leq 4\delta_n \log(3n+1) + 4\delta_n + 3. \tag{3.78}$$

Similarly, we have

$$\log \frac{\binom{\widetilde{N}_1}{\widetilde{K}_1}}{\binom{N_1}{K_1}} \leq 4\delta_n \log(3n+1) + 4\delta_n + 3 \tag{3.79}$$

and substituting (3.78) and (3.79) into (3.66) yields

$$\begin{aligned} R_{n,P_X}(q_{\text{mix}}, f^*) &\leq d \log\left(e^4 n\right) + \mathbb{E}_{X^n,Y^n}\left[8\delta_n \log(3n+1) + 8\delta_n + 6\right] \\ &= d \log\left(e^4 n\right) + 8 \log(6n+2) \mathbb{E}_{X^n,Y^n}[\delta_n] + 6 \end{aligned} \tag{3.80}$$

Now, we have

$$\mathbb{E}_{X^n,Y^n}[\delta_n] = \mathbb{E}_{X^n}\left[\sum_{i=1}^{n} \mathbb{1}\{g^*(X_i) \neq \widetilde{g}(X_i)\}\right] = n\mathbb{P}(g^*(X_1) \neq \widetilde{g}(X_1)) \leq 1$$

Since by design $d(\widetilde{g}, g^*) = \mathbb{P}(\widetilde{g}(X) \neq g^*(X)) \leq 1/n$ where X is distributed as P_X. Substituting this into (3.80) yields

$$R_{n,P_X}(q_{\text{mix}}, f^*) \leq (d+8) \log\left(e^4 n\right) + 6. \tag{3.81}$$

3.7.2 Skipped proofs from Section 3.3

Proof of Proposition 3

By Proposition 2, we have for a fixed $f^* = (a^*, \theta_0^*, \theta_1^*)$

$$R_{n,P_X}(q_{\text{mix}}, f^*) = \mathbb{E}\left[\frac{p_{f^*}(Y^n|X^n)}{\mathbb{E}_F[p_F(Y^n|X^n)]}\right]$$

$$= \mathbb{E}\left[\log \frac{p_{f^*}(Y^n|X^n)}{\mathbb{E}_A[q_{A,\mathrm{mix}}(Y^n|X^n)]}\right] \tag{3.82}$$

We will need the following claim.

Claim 1. *Let $a^* \in \mathbb{S}^{d-1}$ denote the function picked by the adversary, and $\delta := \min_i |a^{*T}X_i|$. Then, for all $a \in \mathbb{S}^{d-1}$ such that $\|a - a^*\| < \delta$, we have $q_{a,\mathrm{mix}}(Y^i|X^i) = q_{a^*,\mathrm{mix}}(Y^i|X^i)$.*

Proof. Note that by definition of $q_{a,\mathrm{mix}}(Y^i|X^i)$, showing that

$$(g_a(X_1), \ldots, g_a(X_n)) = (g_{a^*}(X_1), \ldots, g_{a^*}(X_n))$$

or equivalently that

$$(\mathrm{sign}(a^T X_1), \ldots, \mathrm{sign}(a^T X_n)) = (\mathrm{sign}(a^{*T} X_1), \ldots, \mathrm{sign}(a^{*T} X_n)) \tag{3.83}$$

for all $\{a : \|a - a^*\| < \delta\}$ suffices to prove the claim. Observe now that for all $a \in \mathbb{S}^{d-1}$ we have $\mathrm{sign}(a^T X_i) = \mathrm{sign}(a^{*T} X_i + (a - a^*)^T X_i)$, and if $\|a^* - a\| < \delta$, we have $|(a - a^*)^T X_i| \leq \|a - a^*\| < \delta$, and therefore $\mathrm{sign}(w^T X_i) = \mathrm{sign}(w^{*T} X_i)$ for all $i = 1, \ldots, n$ (since $|w^{*T} X_i| \geq \delta$). This proves (3.83) and consequently the claim. □

We now have

$$\begin{aligned} q_{A,\mathrm{mix}}(Y^n|X^n) &\geq q_{A,\mathrm{mix}}(Y^n|X^n) \mathbb{1}\{|a^* - A| < \delta\} \\ &= q_{a^*,\mathrm{mix}}(Y^n|X^n) \mathbb{1}\{|a^* - A| < \delta\} \end{aligned} \tag{3.84}$$

where (3.84) follows from Claim 1. Then,

$$\begin{aligned} \mathbb{E}_A[q_{A,\mathrm{mix}}(Y^n|X^n)] &\geq \mathbb{E}_A[q_{a^*,\mathrm{mix}}(Y^n|X^n) \mathbb{1}\{|a^* - A| < \delta\}] \\ &= q_{a^*,\mathrm{mix}}(Y^n|X^n) \mathbb{E}_A[\mathbb{1}\{|a^* - A| < \delta\}] \end{aligned}$$

$$= q_{a^*,\mathrm{mix}}(Y^n|X^n)\frac{\mathrm{Area}(\{\|a-a^*\| \leq \delta\} \cap \mathbb{S}^{d-1})}{\mathrm{Area}(\mathbb{S}^{d-1})} \tag{3.85}$$

where (3.85) follows since $A \sim \mathrm{Uniform}[\mathbb{S}^{d-1}]$.

We now bound $\frac{\mathrm{Area}(\{\|a-a^*\| \leq \delta\} \cap \mathbb{S}^{d-1})}{\mathrm{Area}(\mathbb{S}^{d-1})}$ by a simple covering number argument explained next. Consider $\mathscr{N}(d,\delta)$ to be a $\delta-$covering of $\mathbb{S}^{d-1}$, consisting of the points $z_1, \ldots, z_{|\mathscr{N}(d,\delta)|}$. Then by definition of a covering,

$$\mathbb{S}^{d-1} = \cup_{i=1}^{|\mathscr{N}(d,\delta)|}\left(\{\|z-z_i\| \leq \delta\} \cap \mathbb{S}^{d-1}\right)$$

and subsequently,

$$\begin{aligned}\mathrm{Area}(\mathbb{S}^{d-1}) &= \mathrm{Area}\left(\cup_{i=1}^{|\mathscr{N}(d,\delta)|}\left(\{\|z-z_i\| \leq \delta\} \cap \mathbb{S}^{d-1}\right)\right)\\ &\leq \sum_i \mathrm{Area}\left(\{\|z-z_i\| \leq \delta\} \cap \mathbb{S}^{d-1}\right)\\ &= |\mathscr{N}(d,\delta)|\mathrm{Area}(\{\|z-a^*\| \leq \delta\} \cap \mathbb{S}^{d-1}) \qquad (3.86)\end{aligned}$$

where (3.86) follows by symmetry of $\mathbb{S}^{d-1}$, which implies that any for each point $z \in \mathbb{S}^{d-1}$ the $\delta-$neighbourhood is isomorphic. This establishes that

$$\frac{\mathrm{Area}(\{\|a-a^*\| \leq \delta\} \cap \mathbb{S}^{d-1})}{\mathrm{Area}(\mathbb{S}^{d-1})} \geq \frac{1}{|\mathscr{N}(d,\delta)|}. \tag{3.87}$$

Finally, we can show that when $\delta \leq 1$,

$$|\mathscr{N}(d,\delta)| \leq \left(\frac{3}{\delta}\right)^d$$

since $\mathscr{N}(d,\delta) \leq \mathscr{N}(\mathbb{B}_d,\delta)$, the covering number of the unit ball, and $\mathscr{N}(\mathbb{B}_d,\delta) \leq \left(\frac{3}{\delta}\right)^d$ [94, Chapter 4]. This implies that $\frac{\mathrm{Area}(\{\|a-a^*\| \leq \delta\} \cap \mathbb{S}^{d-1})}{\mathrm{Area}(\mathbb{S}^{d-1})} \geq \left(\frac{\delta}{3}\right)^d$. Now, substituting this back in (3.85) yields

$$\mathbb{E}_A[q_{A,\mathrm{mix}}(Y^n|X^n)] \geq q_{a^*,\mathrm{mix}}(Y^n|X^n)\left(\frac{\delta}{3}\right)^d \tag{3.88}$$

and by substituting (3.88) into (3.82), we have

$$R_n(f^*, q_{\mathrm{mix}}) \leq \mathbb{E}\left[\log\frac{p_{f^*}(Y^n|X^n)}{q_{a^*,\mathrm{mix}}(Y^n|X^n)}\right] + d\mathbb{E}\left[\log\frac{3}{\delta}\right]. \tag{3.89}$$

We now consider $\mathbb{E}\left[\log\frac{1}{\delta}\right]$. Recall that we have $\delta = \min_i |a^{*T}X_i|$, where $X_i \sim \mathrm{Uniform}(\mathbb{S}^{d-1})$ i.i.d. By symmetry, for any $a_1, a_2 \in \mathbb{S}^{d-1}$ we have

$$(|a_1^T X_1|, \ldots, |a_1^T X_n|) \overset{(d)}{=} (|a_1^T X_1|, \ldots, |a_2^T X_n|).$$

In particular, choosing $a_1 = a^*$ and $a_2 = \begin{bmatrix} 1 & 0 \cdots 0 \end{bmatrix}$ we have

$$(|a^{*T} X_1|, \ldots, |a^{*T} X_n|) \overset{(d)}{=} (|X_{1,1}|, \ldots, |X_{n,1}|)$$

where $X_{i,1}$ denotes the first co-ordinate of X_i. Now, for $X_i \sim \mathrm{Uniform}(\mathbb{S}^{d-1})$, $X_{i,1} = 2Z - 1$ where $Z \sim \mathrm{Beta}(d/2, d/2)$ (this follows directly from the formula for the surface area of the hyperspherical cap, see for example [95]). So, $X_{1,1}, \ldots, X_{n,1}$ are i.i.d. samples from a shifted and rescaled beta distribution. Thus, we can explicitly calculate $\mathbb{E}_{X^n}[-\log\delta]$, which is simply $\mathbb{E}_{Z^n}[-\log(\min|2Z_i - 1|)] = \mathbb{E}_{Z^n}[\max_i -\log|2Z_i - 1|]$ where $Z_i \sim \mathrm{Beta}(d/2, d/2)$. We will next show that $\mathbb{E}[-\log\delta] \leq 2\ln(n) + o(1)$.

Let $Z \sim \mathrm{Beta}(d/2, d/2)$, and $W := -\ln|2Z - 1|$. Since $Z \in [0, 1]$, we have $W \geq 0$. Recalling that the density of Z is $f_Z(z) = \frac{(z(1-z))^{d/2-1}}{\mathrm{B}(d/2,d/2)}, 0 \leq z \leq 1$, we can then calculate the density $f_W(w)$ as follows. We have, for any $w \geq 0$,

$$1 - F_W(w) = \mathbb{P}(W > w)$$

$$
\begin{aligned}
&= \mathbb{P}(-\ln|2Z-1| > w) \\
&= \mathbb{P}\left(\frac{1-e^{-w}}{2} < Z < \frac{1+e^{-w}}{2}\right) \\
&= F_Z\left(\frac{1+e^{-w}}{2}\right) - F_Z\left(\frac{1-e^{-w}}{2}\right).
\end{aligned}
$$

Since $f_W(w) = \frac{dF_W(w)}{dw}$, taking derivative with respect to w on both sides of (3.90) yields

$$
\begin{aligned}
f_W(w) &= \frac{dF_Z\left(\frac{1-e^{-w}}{2}\right)}{dw} - \frac{dF_Z\left(\frac{1+e^{-w}}{2}\right)}{dw} \\
&= \frac{1}{\mathrm{B}(d/2,d/2)} e^{-w}\left(\frac{1-e^{-2w}}{4}\right)^{d/2-1} \qquad (3.90)
\end{aligned}
$$

Since W is sub-exponential, we expect the scaling of $\mathbb{E}[\max\{W_1,\ldots,W_n\}]$ with n to be $O(\log n)$ (i.e. similar to the dependence on n of expected maximum for an exponential distribution). We next formalize this using a standard technique for bounding maximum of independent random variables. First, we provide a useful claim.

Claim 2. *For all $w \geq 0$, we have $\frac{1}{\mathrm{B}(d/2,d/2)}\left(\frac{1-e^{-2w}}{4}\right)^{d/2-1} \leq c_d$ where $c_d := 2\sqrt{d}$, and subsequently $f_W(w) \leq c_d e^{-w}$.*

Proof. Uses simple properties of the beta function and a Stirling approximation. □

For n i.i.d. samples from W, denoted W^n, we show that $\mathbb{E}[\max\{W_1,\ldots,W_n\}] \leq 2\ln(2c_d n)$. Note that

$$
\begin{aligned}
\mathbb{E}[\max\{W_1,\ldots,W_n\}] &= 2\mathbb{E}\left[\ln\max\{e^{W_1/2},\ldots,e^{W_n/2}\}\right] \\
&\leq 2\ln\left(\mathbb{E}\left[\max\{e^{W_1/2},\ldots,e^{W_n/2}\}\right]\right) \qquad (3.91) \\
&\leq 2\ln\left(\mathbb{E}\left[\sum_{i=1}^{n} e^{W_i}/2\right]\right) \\
&= 2\ln\left(n\mathbb{E}\left[e^{W_1/2}\right]\right) \\
&= 2\ln\left(n\int_0^{\infty} e^{w/2} f_W(w)dw\right)
\end{aligned}
$$

$$\leq 2\ln\left(n\int_0^{\infty} e^{w/2} c_d e^{-w} dw\right) \tag{3.92}$$

$$\leq 2\ln\left(c_d n\int_0^{\infty} e^{-w/2} dw\right) = 2\ln(2c_d n). \tag{3.93}$$

where (3.91) follows from the Jensen inequality and (3.92) follows from Claim 2. Therefore,

$$\mathbb{E}[-\log\delta] \leq 2\ln n + 2\ln(4\sqrt{d}). \tag{3.94}$$

Going back to (3.89), and using Lemma 1 and (3.94) yields

$$\max_{f\in\mathscr{F}} R_{n,P_X}(q_{\text{mix}}, f) \leq (2\ln 2d + 1)\log n + d\log(48d) + \log\frac{\pi^2}{8}. \tag{3.95}$$

Proof of Proposition 4

The flow of this proof is almost the same as that of Proposition 3. By Proposition 2, we have for a fixed $f^* = (\mathbf{a}^*, \mathbf{b}^*, \theta_0^*, \theta_1^*)$

$$R_{n,P_X}(q_{\text{mix}}, f^*) = \mathbb{E}\left[\frac{p_{f^*}(Y^n|X^n)}{\mathbb{E}_F[p_F(Y^n|\mathbf{X}^n)]}\right] \tag{3.96}$$

Fix some $f^* = (\mathbf{a}^*, \mathbf{b}^*, \theta_0^*, \theta_1^*)$. Now, recall that $\mathbf{X}_j \in \mathbb{R}^d, j \in [n]$. Denote the $i-$th coordinate of $\mathbf{X}_j$ by $\mathbf{X}_{j,i}$. Now, given the n real numbers $\mathbf{X}_{1,i}, \ldots, \mathbf{X}_{n,i}$ (i.e. the $i-$th coordinates of $\mathbf{X}_1, \ldots, \mathbf{X}_n$) we can arrange these in order as $\mathbf{X}^i_{(1)} \leq \mathbf{X}^i_{(2)} \leq \ldots \leq \mathbf{X}^i_{(n)}$. Thus, $\mathbf{X}^i_{(1)}, \ldots, \mathbf{X}^i_{(n)}$ denote the order statistics of the $i-$th component of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Now, clearly, there exist unique $k_i, l_i \in 0, \ldots, n, l_i \geq k_i$ such that $\mathbf{X}^i_{(k_i)} \leq a_i^* \leq \mathbf{X}^i_{(k_i+1)}$ and $\mathbf{X}^i_{(l_i)} \leq b_i^* \leq \mathbf{X}^i_{(l_i+1)}$. This holds for all $i \in [d]$. We then make the following claim.

Claim 3. *For any* $\mathbf{a}, \mathbf{b}$ *that are such that for all* $i \in [d]$,

$$\mathbf{X}^i_{(k_i)} \leq a_i \leq \mathbf{X}^i_{(k_i+1)}, \mathbf{X}^i_{(l_i)} \leq b_i \leq \mathbf{X}^i_{(l_i+1)}$$

and $a_i \leq b_i$, we have $q_{\mathbf{a},\mathbf{b},\mathrm{mix}}(Y^i|\mathbf{X}^i) = q_{\mathbf{a}^*,\mathbf{b}^*,\mathrm{mix}}(Y^i|\mathbf{X}^i)$.

Proof. To prove this claim, note first that from the definition of $q_{\mathbf{a},\mathbf{b},\mathrm{mix}}$ in (3.28), we have that if

$$(g_{\mathbf{a}^*,\mathbf{b}^*}(\mathbf{X}_1),\ldots,g_{\mathbf{a}^*,\mathbf{b}^*}(\mathbf{X}_n)) = (g_{\mathbf{a},\mathbf{b}}(\mathbf{X}_1),\ldots,g_{\mathbf{a},\mathbf{b}}(\mathbf{X}_n)) \tag{3.97}$$

the claim holds. Then, for $\mathbf{X}_1$, we have $g_{\mathbf{a}^*,\mathbf{b}^*}(\mathbf{X}_1) = \prod_{i=1}^d \mathbb{1}\{a_i^* \leq \mathbf{X}_{1,i} \leq b_i^*\}$. Now, for any $i \in [d]$, since $\mathbf{X}^i_{(k_i)} \leq a_i^* \leq \mathbf{X}^i_{(k_i+1)}$, this implies that if $\mathbf{X}^i_{(k_i)} \leq a_i \leq \mathbf{X}^i_{(k_i+1)}$ we have that $\{j : \mathbf{X}_{j,i} \geq a_i^*\} = \{j : \mathbf{X}_{j,i} \geq a_i\}$ and therefore $\mathbb{1}\{a_i^* \leq \mathbf{X}_{1,i}\} = \mathbb{1}\{a_i \leq \mathbf{X}_{1,i}\}$. Similarly $\{j : \mathbf{X}_{j,i} \leq b_i^*\} = \{j : \mathbf{X}_{j,i} \leq b_i\}$ if $\mathbf{X}^i_{(l_i)} \leq b_i \leq \mathbf{X}^i_{(l_i+1)}$, and consequently $\mathbb{1}\{\mathbf{X}_{1,i} \leq b_i^*\} = \mathbb{1}\{\mathbf{X}_{1,i} \leq b_i\}$. This implies that for any $\mathbf{a},\mathbf{b}$ satisfying the conditions of the claim, we have $\mathbb{1}\{a_i^* \leq \mathbf{X}_{1,i} \leq b_i^*\} = \mathbb{1}\{a_i \leq \mathbf{X}_{1,i} \leq b_i\}$ for all $i \in [d]$, thereby implying that $g_{\mathbf{a}^*,\mathbf{b}^*}(\mathbf{X}_1) = g_{\mathbf{a},\mathbf{b}}(\mathbf{X}_1)$. The same argument applied to $\mathbf{X}_2,\ldots,\mathbf{X}_n$ implies (3.97) and therefore the claim. □

Now, we have

$$\begin{aligned}
&q_{\mathbf{A},\mathbf{B},\mathrm{mix}}(Y^n|\mathbf{X}^n)\\
&\geq q_{\mathbf{A},\mathbf{B},\mathrm{mix}}(Y^n|\mathbf{X}^n)\prod_{i=1}^d \mathbb{1}\{\mathbf{X}^i_{(k_i)} \leq A_i \leq \mathbf{X}^i_{(k_i+1)}\}\mathbb{1}\{\mathbf{X}^i_{(l_i)} \leq B_i \leq \mathbf{X}^i_{(l_i+1)}\}\\
&= q_{\mathbf{a}^*,\mathbf{b}^*,\mathrm{mix}}(Y^n|\mathbf{X}^n)\prod_{i=1}^d \mathbb{1}\{\mathbf{X}^i_{(k_i)} \leq A_i \leq \mathbf{X}^i_{(k_i+1)}\}\mathbb{1}\{\mathbf{X}^i_{(l_i)} \leq B_i \leq \mathbf{X}^i_{(l_i+1)}\} \qquad (3.98)
\end{aligned}$$

where (3.98) follows from Claim 3. Now, we have

$$\begin{aligned}
&\mathbb{E}_F\left[p_F(Y^n|\mathbf{X}^n)\right]\\
&= \mathbb{E}_{\mathbf{A},\mathbf{B}}\left[q_{\mathbf{A},\mathbf{B},\mathrm{mix}}(Y^n|\mathbf{X}^n)\right]\\
&\geq \mathbb{E}_{\mathbf{A},\mathbf{B}}\left[q_{\mathbf{a}^*,\mathbf{b}^*,\mathrm{mix}}(Y^n|\mathbf{X}^n)\prod_{i=1}^d \mathbb{1}\{\mathbf{X}^i_{(k_i)} \leq A_i \leq \mathbf{X}^i_{(k_i+1)}\}\mathbb{1}\{\mathbf{X}^i_{(l_i)} \leq B_i \leq \mathbf{X}^i_{(l_i+1)}\}\right] \qquad (3.99)\\
&= q_{\mathbf{a}^*,\mathbf{b}^*,\mathrm{mix}}(Y^n|\mathbf{X}^n)\mathbb{E}_{\mathbf{A},\mathbf{B}}\left[\prod_{i=1}^d \mathbb{1}\{\mathbf{X}^i_{(k_i)} \leq A_i \leq \mathbf{X}^i_{(k_i+1)}\}\mathbb{1}\{\mathbf{X}^i_{(l_i)} \leq B_i \leq \mathbf{X}^i_{(l_i+1)}\}\right]
\end{aligned}$$

$$= q_{\mathbf{a}^*,\mathbf{b}^*,\text{mix}}(Y^n|\mathbf{X}^n)\prod_{i=1}^{d}\mathbb{E}_{A_i,B_i}\left[\mathbb{1}\{\mathbf{X}^i_{(k_i)} \le A_i \le \mathbf{X}^i_{(k_i+1)}\}\mathbb{1}\{\mathbf{X}^i_{(l_i)} \le B_i \le \mathbf{X}^i_{(l_i+1)}\}\right] \tag{3.100}$$

$$\ge q_{\mathbf{a}^*,\mathbf{b}^*,\text{mix}}(Y^n|\mathbf{X}^n)\prod_{i=1}^{d}(\mathbf{X}^i_{(k_i+1)} - \mathbf{X}^i_{(k_i)})(\mathbf{X}^i_{(l_i+1)} - \mathbf{X}^i_{(l_i)})/2 \tag{3.101}$$

where (3.99) follows from (3.98), (3.100) follows since the (A_i, B_i) are all mutually independent, and (3.101) follows since

$$\mathbb{E}_{A_i,B_i}\left[\mathbb{1}\{\mathbf{X}^i_{(k_i)} \le A_i \le \mathbf{X}^i_{(k_i+1)}\}\mathbb{1}\{\mathbf{X}^i_{(l_i)} \le B_i \le \mathbf{X}^i_{(l_i+1)}\}\right]$$
$$= \begin{cases} (\mathbf{X}^i_{(k_i+1)} - \mathbf{X}^i_{(k_i)})(\mathbf{X}^i_{(l_i+1)} - \mathbf{X}^i_{(l_i)})/2 & \text{for } l_i = k_i \\ (\mathbf{X}^i_{(k_i+1)} - \mathbf{X}^i_{(k_i)})(\mathbf{X}^i_{(l_i+1)} - \mathbf{X}^i_{(l_i)}) & \text{for } l_i > k_i. \end{cases} \tag{3.102}$$

By substituting (3.101) into (3.96), we get

$$R_n(q_{\text{mix}}) \le \mathbb{E}\left[\log\frac{p_{\mathbf{a}^*,\mathbf{b}^*,\theta_0,\theta_1}(Y^n|\mathbf{X}^n)}{q_{\mathbf{a}^*,\mathbf{b}^*,\text{mix}}(Y^n|\mathbf{X}^n)}\right] + \sum_{i=1}^{d}\mathbb{E}_{\mathbf{X}^n}\left[\log\frac{2}{\mathbf{X}^i_{(k_i+1)} - \mathbf{X}^i_{(k_i)}}\right]$$
$$+\sum_{i=1}^{d}\mathbb{E}_{\mathbf{X}^n}\left[\log\frac{2}{\mathbf{X}^i_{(l_i+1)} - \mathbf{X}^i_{(l_i)}}\right] \tag{3.103}$$

Now, consider $\mathbb{E}_{\mathbf{X}^n}\left[\log\frac{2}{\mathbf{X}^i_{(k_i+1)} - \mathbf{X}^i_{(k_i)}}\right]$. Clearly, this quantity depends only on the $i-$th coordinates of $\mathbf{X}^n$, $\mathbf{X}_{1,i},\dots,\mathbf{X}_{n,i}$. Since $\mathbf{X}^n \sim \text{Uniform}[0,1]^d$ i.i.d, we can see that $\mathbf{X}_{1,i},\dots,\mathbf{X}_{n,i} \sim \text{Uniform}[0,1]$ i.i.d. Now, it is known that for $Z^n \sim \text{Uniform}[0,1]$ i.i.d., $Z_{(k+1)} - Z_{(k)} \sim \text{Beta}(1,n)$ for all $k \in \{0,\dots,n\}$. Moreover, for $Z' \sim \text{Beta}(\alpha,\beta)$, it can be shown that $\mathbb{E}[-\log Z'] \le \log(\alpha+\beta)$. Using these two results, we can conclude that

$$\mathbb{E}_{\mathbf{X}^n}\left[\log\frac{1}{\mathbf{X}^i_{(k_i+1)} - \mathbf{X}^i_{(k_i)}}\right], \mathbb{E}_{\mathbf{X}^n}\left[\log\frac{1}{\mathbf{X}^i_{(l_i+1)} - \mathbf{X}^i_{(l_i)}}\right] \le \log(n+1), i \in [d]. \tag{3.104}$$

Finally, using Lemma 1 and (3.104) in (3.103) yields

$$\max_{f\in\mathscr{F}} R_{n,P_X}(q_{\text{mix}}, f) \le (2d+1)\log(n+1) + \log\frac{\pi^2}{8} \tag{3.105}$$

as required.

3.7.3 Skipped Proofs from Section 3.4

Proof of Lemma 6

We have

$$\begin{aligned}
&\log\frac{1}{\widetilde{q}_{\text{mix}}(Y_i|X^i,Y^{i-1},\widetilde{X}^n)}-\sum_{i=1}^{n}\log\frac{1}{p_f(Y_i|X_i)}\\
&\qquad=\log\frac{p_{f^*}(Y^n|X^n)}{\frac{1}{|\mathscr{P}(\widetilde{X}^n)|}\sum_{j=1}^{|\mathscr{P}(\widetilde{X}^n)|}\int_0^1\int_0^1 p_{\widetilde{g}_j,\theta_0,\theta_1}(Y^n|X^n)d\theta_0 d\theta_1}\\
&\qquad\le d\log(en/d)+\log\frac{p_{f^*}(Y^n|X^n)}{\sum_{j=1}^{|\mathscr{P}(\widetilde{X}^n)|}\int_0^1\int_0^1 p_{\widetilde{g}_j,\theta_0,\theta_1}(Y^n|X^n)d\theta_0 d\theta_1}. \qquad (3.106)
\end{aligned}$$

So far, the construction and analysis of $\widetilde{q}_{\text{mix}}$ has paralleled the analysis of the mixture q_{mix} in Section 3.2.3. There, the next step was to claim that since $\exists j\in[|\mathscr{P}_n(X^n)|]$ such that $g_j(X^n)=g^*(X^n)$, invoking Lemma 1 yielded an $O(\log n)$ upper bound for the second term in (3.106). Unfortunately we cannot claim the same in the current case. However, we can claim that there exists $\tilde{j}\in\left[|\mathscr{P}_n(\widetilde{X}^n)|\right]$ such that

$$\widetilde{g}_{\tilde{j}}(\widetilde{X}^n)=g^*(\widetilde{X}^n).$$

Since $d_H(\widetilde{g}_{\tilde{j}}(\widetilde{X}^n),g^*(\widetilde{X}^n))=0$ and $\widetilde{X}^n\overset{(d)}{=}X^n$, we would expect $d_H(\widetilde{g}_{\tilde{j}}(X^n),g^*(X^n))$ to not be too large. We now quantify this intuition more precisely. For brevity, denote

$$\widetilde{g}:=\widetilde{g}_{\tilde{j}}.$$

We have from (3.106)

$$\begin{aligned}
&\log\frac{1}{\widetilde{q}_{\text{mix}}(Y_i|X^i,Y^{i-1},\widetilde{X}^n)}-\sum_{i=1}^{n}\log\frac{1}{p_f(Y_i|X_i)}\\
&\qquad\qquad\le d\log(en/d)+\log\frac{p_{f^*}(Y^n|X^n)}{\int_0^1\int_0^1 p_{\widetilde{g},\theta_0,\theta_1}(Y^n|X^n)d\theta_0 d\theta_1} \qquad (3.107)
\end{aligned}$$

$$\leq d\log(en/d)+8\log(6n+2)d_H(g^*(X^n),\widetilde{g}(X^n))+6 \tag{3.108}$$

Where to get from (3.107) to (3.108) we follow the exact same steps employed in the proof of Lemma 5 from (3.61) to (3.80).

We now focus on $d_H(g^*(X^n),\widetilde{g}(X^n))$ and establish that

$$\mathbb{E}_{X^n,\widetilde{X}^n}[d_H(g^*(X^n),\widetilde{g}(X^n))] \leq 2C\sqrt{dn} \tag{3.109}$$

$$d_H(g^*(X^n),\widetilde{g}(X^n)) \leq 2C\sqrt{dn}+2\sqrt{2n\log\frac{2}{\delta}}\text{ , with probability } \geq 1-\delta \tag{3.110}$$

for an absolute constant $C \leq 250$.

For any $g_1,g_2 \in \mathscr{G}$ define

$$\begin{aligned}\Delta_n(g_1,g_2) &:= \frac{1}{n}d_H(g_1(X^n),g_2(X^n))\\ \tilde{\Delta}_n(g_1,g_2) &:= \frac{1}{n}d_H(g_1(\widetilde{X}^n),g_2(\widetilde{X}^n))\\ \Delta(g_1,g_2) &:= \mathbb{P}(g_1(X) \neq g_2(X))\end{aligned}$$

for $X \overset{(d)}{=} X_1 \overset{(d)}{=} \widetilde{X}_1$. Recall that $\tilde{\Delta}_n(\widetilde{g},g^*)=0$ by design, and $\Delta(g_1,g_2)=\mathbb{E}_{X^n}[\Delta_n(g_1,g_2)] = \mathbb{E}_{\widetilde{X}^n}[\tilde{\Delta}_n(g_1,g_2)]$. We then have

$$\begin{aligned}\Delta_n(g^*,\widetilde{g}) &= \Delta_n(g^*,\widetilde{g})-\tilde{\Delta}_n(g^*,\widetilde{g})\\ &\leq \sup_{g_1,g_2\in\mathscr{G}}\left|\Delta_n(g_1,g_2)-\tilde{\Delta}_n(g_1,g_2)\right|\\ &\leq \sup_{g_1,g_2\in\mathscr{G}}\left|\Delta_n(g_1,g_2)-\Delta(g_1,g_2)\right|+\sup_{g_1,g_2\in\mathscr{G}}\left|\tilde{\Delta}_n(g_1,g_2)-\Delta(g_1,g_2)\right|. \end{aligned}\tag{3.111}$$

We first establish (3.109). Taking expectations on both sides of (3.111).

$$\mathbb{E}_{X^n,\widetilde{X}^n}[\Delta_n(g^*,\widetilde{g})]$$

$$\leq \mathbb{E}_{X^n,\widetilde{X}^n}\left[\sup_{g_1,g_2\in\mathscr{G}} |\Delta_n(g_1,g_2)-\Delta(g_1,g_2)| + \sup_{g_1,g_2\in\mathscr{G}} \left|\tilde{\Delta}_n(g_1,g_2)-\Delta(g_1,g_2)\right|\right]$$

$$= 2\mathbb{E}_{X^n}\left[\sup_{g_1,g_2\in\mathscr{G}} |\Delta_n(g_1,g_2)-\Delta(g_1,g_2)|\right] \tag{3.112}$$

where (3.112) follows by linearity of expectation and since $X^n \overset{(d)}{=} \widetilde{X}^n$. Finally, we note that

$$\Delta_n(g_1,g_2) = \frac{d_H(g_1(X^n),g_2(X^n))}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{g_1(X_i)\neq g_2(X_i)\}$$

$$\Delta(g_1,g_2) = \mathbb{E}[\mathbb{1}\{g_1(X)\neq g_2(X)\}]$$

and the class of boolean functions $\{x \mapsto \mathbb{1}\{g_1(x)\neq g_2(x)\}, (g_1,g_2)\in\mathscr{G}\times\mathscr{G}\}$ has VC dimension $\leq 2d$. Thus, we can now invoke [94, Theorem 8.3.23], [91, Theorem 13.7] to claim that

$$\mathbb{E}_{X^n}\left[\sup_{g_1,g_2\in\mathscr{G}} |\Delta_n(g_1,g_2)-\Delta(g_1,g_2)|\right] \leq C\sqrt{\frac{d}{n}} \tag{3.113}$$

for a universal constant $C \leq 250$. Consequently, taking expectations on both sides of (3.108) and substituting (3.113), followed by a supremum over f^* and P_X yields

$$\tilde{R}_n(\widetilde{q}_{\text{mix}}) \leq d\log(en/d) + 16C\sqrt{nd}\log(6n+2) \tag{3.114}$$

as required.

To establish (3.110), we invoke Theorem 12.1 of [91] to assert

$$\sup_{g_1,g_2\in\mathscr{G}} |\Delta_n(g_1,g_2)-\Delta(g_1,g_2)| \leq \mathbb{E}\left[\sup_{g_1,g_2\in\mathscr{G}} |\Delta_n(g_1,g_2)-\Delta(g_1,g_2)|\right] + \sqrt{\frac{2}{n}\log\frac{2}{\delta}} \tag{3.115}$$

with probability $1-\delta/2$. The same high-probability bound for

$$\sup_{g_1,g_2\in\mathscr{G}} \left|\tilde{\Delta}_n(g_1,g_2)-\Delta(g_1,g_2)\right|$$

along with a union bound and (3.113) yields (3.110). Using (3.110) in (3.108) yields the second part of the lemma.

Proof of Theorem 1

We have

$$\begin{aligned}
&\sum_{i=1}^{n} \log \frac{1}{q^*(Y_i|X^i,Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)} \\
&\quad \leq \sum_{i=2}^{n} \log \frac{1}{q^*(Y_i|X^i,Y^{i-1})} - \sum_{i=2}^{n} \log \frac{1}{p_f(Y_i|X_i)} + 1 \\
&\quad = \sum_{j=1}^{\log n} \left[\sum_{i=2^{j-1}+1}^{2^j} \log \frac{1}{q^*(Y_i|X^i,Y^{i-1})} - \sum_{i=2^{j-1}+1}^{2^j} \log \frac{1}{p_f(Y_i|X_i)} \right] \qquad (3.116)
\end{aligned}$$

Taking expectation on both sides of (3.116), we have

$$\begin{aligned}
R_{n,P_X}(q^*,f) &\leq \sum_{j=1}^{\log n} \mathbb{E} \left[\sum_{i=2^{j-1}+1}^{2^j} \log \frac{1}{q^*(Y_i|X^i,Y^{i-1})} - \sum_{i=2^{j-1}+1}^{2^j} \log \frac{1}{p_f(Y_i|X_i)} \right] + 1 \\
&\leq \sum_{j=1}^{\log n} \tilde{R}_{2^{j-1}}(\widetilde{q}_{\text{mix}}) \qquad (3.117)
\end{aligned}$$

where recall $\mathbb{E}$ in the first inequality are w.r.t. $X_{2^{j-1}+1}^{2^j}, Y_{2^{j-1}+1}^{2^j}, X_1^{2^{j-1}}$, and (3.117) follows since q^* is exactly $\widetilde{q}_{\text{mix}}$. Using Lemma 6, we have for any $n' \geq 2$

$$\tilde{R}_{n'}(\widetilde{q}_{\text{mix}}) \leq d\log(en'/d) + 16C\sqrt{dn'}\log(6n'+2) \leq d\log n' + 64C\sqrt{dn'}\log(n')$$

and therefore, from (3.117)

$$\begin{aligned}
R_{n,P_X}(q^*,f) &\leq \sum_{j=2}^{\log n} \Big(d(j-1) + 64C\sqrt{d}2^{(j-1)/2}(j-1) \Big) + 2 \\
&\leq d(\log n)^2 + 64C\sqrt{d} \int_1^{\log n+1} x2^{x/2} dx + 2 \qquad (3.118) \\
&\leq d(\log n)^2 + 125C\sqrt{dn}\log(2n) + 2 \qquad (3.119)
\end{aligned}$$

and finally taking supremum over f and P_X concludes the first part of the proof.

For the second part, we have from Lemma 6 for any $j \geq 2$,

$$\sum_{i=2^{j-1}+1}^{2^j} \log \frac{1}{q^*(Y_i|X^i, Y^{i-1})} - \sum_{i=2^{j-1}+1}^{2^j} \log \frac{1}{p_f(Y_i|X_i)} \\ \leq d(j-1) + 642^{(j-1)/2}(j-1)\left(C\sqrt{d} + \sqrt{2\log\frac{2\log n}{\delta}}\right) \tag{3.120}$$

with probability $1 - \delta/\log n$. Then from (3.116), a union bound and the calculations in (3.118) and (3.119) we have

$$\sum_{i=1}^{n} \log \frac{1}{q^*(Y_i|X^i, Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)} \\ \leq d(\log n)^2 + 125C\sqrt{dn}\log(2n)\left(C\sqrt{d} + \sqrt{2\log\frac{2\log n}{\delta}}\right) + 2 \tag{3.121}$$

with probability $\geq 1 - \delta$ as required.

3.7.4 Skipped Proofs from Section 3.5

Proof of Lemma 7

We have from (3.41)

$$R_n \geq \max_{P_X, P_F} I(\Theta_0, \Theta_1, G; Y^n | X^n) \tag{3.122}$$

and therefore a lower bound on $I(\Theta_0, \Theta_1, G; Y^n|X^n)$ for any choice of P_X and P_F provides a lower bound on R_n. We will choose P_X to be the uniform distribution on $\{1, \ldots, m\}$ so that

$$X \sim \text{Uniform}([m]).$$

Consider now the following distribution P_F over hypothesis class $F = (\Theta_0, \Theta_1, G)$ that has

$$(\Theta_0, \Theta_1) \sim \text{Uniform}\left(\theta_0, \theta_1 \in [0,1] \times [0,1] \cap \{\theta_1 - \theta_0 \geq 1/2\}\right) \tag{3.123}$$

$$G \perp\!\!\!\perp (\Theta_0, \Theta_1) \text{ and } G \sim \text{Uniform}\left\{2^{[m]}\right\} \tag{3.124}$$

We then have

$$\begin{aligned}
&I(\Theta_0, \Theta_1, G; Y^n | X^n) \\
&= H(\Theta_0, \Theta_1, G | X^n) - H(\Theta_0, \Theta_1, G | X^n, Y^n) \\
&= h(\Theta_0, \Theta_1) + H(G) - H(\Theta_0, \Theta_1, G | X^n, Y^n) && (3.125) \\
&\geq h(\Theta_0, \Theta_1) + H(G) - H(G | X^n, Y^n) - H(\Theta_0 | G, X^n, Y^n) - H(\Theta_1 | G, X^n, Y^n) && (3.126) \\
&= 3 + m - H(G | X^n, Y^n) - h(\Theta_0 | G, X^n, Y^n) - h(\Theta_1 | G, X^n, Y^n) && (3.127)
\end{aligned}$$

where (3.125) follows since the $G \perp\!\!\!\perp (\Theta_0, \Theta_1)$ and both $(\Theta_0, \Theta_1), G \perp\!\!\!\perp X^n$, (3.126) follows from the chain rule of entropy and because conditioning reduces entropy, and (3.127) follows since by the distribution of (Θ_0, Θ_1) and G in (3.123), (3.124), we have (Θ_0, Θ_1) is uniform over a set with area $\frac{1}{8}$, and $G \sim \text{Uniform}(\mathscr{G})$ with $|\mathscr{G}| = 2^m$.

We also have

$$\begin{aligned}
h(\Theta_0 | X^n, Y^n, G) &= \sum_{g \in \mathscr{G}} h(\Theta_0 | X^n, Y^n, G = g) \mathbb{P}(G = g) \\
&= \frac{1}{2^m} \sum_{g \in \mathscr{G}} h(\Theta_0 | X^n, Y^n, G = g) && (3.128)
\end{aligned}$$

Now, define the estimator

$$\widehat{\Theta}_0(X^n, Y^n, g) = \frac{\sum_{i=1}^n \mathbb{1}\{g(X_i) = 0, Y_i = 1\} + 1/2}{\sum_{i=1}^n \mathbb{1}\{g(X_i) = 0\} + 1}. \tag{3.129}$$

Defining $N_0 := \sum_{i=1}^n \mathbb{1}\{g(X_i)=0\}$ and $K_0 = \sum_{i=1}^n \mathbb{1}\{g(X_i)=0, Y_i=1\}$, we have

$$\widehat{\Theta}_0 = \frac{K_0 + 1/2}{N_0+1}.$$

Now, going back to (3.128), we have

$$\begin{aligned}
\frac{1}{2^m}\sum_{g\in\mathscr{G}} h(\Theta_0|X^n,Y^n,G=g) &\le \frac{1}{2^m}\sum_{g\in\mathscr{G}} h(\Theta_0|\widehat{\Theta}_0,g) && (3.130)\\
&= \frac{1}{2^m}\sum_{g\in\mathscr{G}} h(\Theta_0-\widehat{\Theta}_0|\widehat{\Theta}_0,g)\\
&\le \frac{1}{2^m}\sum_{g\in\mathscr{G}} h(\Theta_0-\widehat{\Theta}_0|g)\\
&\le \frac{1}{2^m}\sum_{g\in\mathscr{G}} \frac{1}{2}\log(2\pi e\mathrm{Var}(\Theta_0-\widehat{\Theta}_0|g)) && (3.131)\\
&\le \frac{1}{2^m}\sum_{g\in\mathscr{G}} \frac{1}{2}\log(2\pi e\mathbb{E}[(\Theta_0-\widehat{\Theta}_0)^2|g]) && (3.132)
\end{aligned}$$

where (3.130) follows from the data processing inequality, (3.131) follows since the Gaussian random variable of a given variance maximizes entropy, and (3.132) follows since for any random variable $Z, \mathrm{Var}[Z] \le \mathbb{E}[Z^2]$.

We now have

$$\begin{aligned}
\mathbb{E}[(\Theta_0-\widehat{\Theta}_0)^2|g] &= \mathbb{E}_{\Theta_0,X^n,Y^n|g}(\Theta_0-\widehat{\Theta}_0)^2\\
&= \mathbb{E}_{\Theta_0,X^n,Y^n|g}\left(\Theta_0 - \frac{K_0+1/2}{N_0+1}\right)^2\\
&= \mathbb{E}_{\Theta_0,N_0,K_0|g}\left(\Theta_0 - \frac{K_0+1/2}{N_0+1}\right)^2\\
&= \mathbb{E}_{\Theta_0|g}\mathbb{E}_{N_0|\Theta_0,g}\mathbb{E}_{K_0|N_0,\Theta_0,g}\left[\left(\Theta_0 - \frac{K_0+1/2}{N_0+1}\right)^2 \middle| N_0,\Theta_0\right] && (3.133)
\end{aligned}$$

Since

$$K_0|N_0,\Theta_0,g \sim \mathrm{Binomial}(N_0,\Theta_0)$$

we can calculate

$$\mathbb{E}_{K_0|N_0,\Theta_0,g}\left[\left(\Theta_0-\frac{K_0+1/2}{N_0+1}\right)^2\Big|N_0,\Theta_0\right]=\frac{(\Theta_0-1/2)^2+N_0\Theta_0(1-\Theta_0)}{(N_0+1)^2}$$
$$\leq\frac{1}{4(N_0+1)} \tag{3.134}$$

where (3.134) follows since $x(1-x)\leq\frac{1}{4},(x-1/2)^2\leq\frac{1}{4}$ for $x\in[0,1]$. Substituting (3.134) back into (3.133) we obtain

$$\mathbb{E}(\Theta_0-\widehat{\Theta}_0)^2\leq\mathbb{E}_{\Theta_0|g}\mathbb{E}_{N_0|\Theta_0,g}\left[\frac{1}{4(N_0+1)}\right]$$
$$=\mathbb{E}_{N_0|g}\left[\frac{1}{4(N_0+1)}\right] \tag{3.135}$$

where (3.135) follows since $N_0|g \perp\!\!\!\perp \Theta_0$ with distribution

$$N_0\sim\text{Binomial}\left(n,\sum_{i:g(i)=1}\mathbb{P}(X=i)\right).$$

Defining

$$p_g:=\sum_{i:g(i)=1}\mathbb{P}(X=i),$$

we can the see that when $p_g\neq 0$ by a simple binomial calculation

$$\mathbb{E}_{N_0|g}\left[\frac{1}{4(N_0+1)}\right]=\frac{1-(1-p_g)^{n+1}}{4(n+1)p_g} \tag{3.136}$$

and $\mathbb{E}_{N_0|g}\left[\frac{1}{4(N_0+1)}\right]=\frac{1}{4}$ when $p_g=0$. Now, we have

$$h(\Theta_0|X^n,Y^n,G)=\frac{1}{2^m}\sum_{g\in\mathscr{G}}h(\Theta_0|X^n,Y^n,G=g)$$
$$\leq\frac{1}{2^m}\sum_{g\in\mathscr{G}}\frac{1}{2}\log\left(2\pi e\mathbb{E}_{N_0|g}\left[\frac{1}{4(N_0+1)}\right]\right) \tag{3.137}$$

$$= \frac{1}{2}\log(\pi e/2) + \sum_{g\in\mathscr{G}} \frac{1}{2}\log\left(\mathbb{E}_{N_0|g}\left[\frac{1}{N_0+1}\right]\right) \tag{3.138}$$

where (3.137) follows from (3.132).

In the exact same way, we can upper-bound $h(\Theta_1|X^n,Y^n,G)$ as

$$h(\Theta_1|X^n,Y^n,G) \le \frac{1}{2}\log(\pi e/2) + \sum_{g\in\mathscr{G}} \frac{1}{2}\log\left(\mathbb{E}_{N_1|g}\left[\frac{1}{N_1+1}\right]\right). \tag{3.139}$$

From (3.138) and (3.139) we get

$$\begin{aligned} &h(\Theta_0|X^n,Y^n,G) + h(\Theta_1|X^n,Y^n,G) \\ &\qquad \le \log(\pi e/2) + \sum_{g\in\mathscr{G}} \frac{1}{2}\log\left(\mathbb{E}_{N_0|g}\left[\frac{1}{N_0+1}\right]\mathbb{E}_{N_1|g}\left[\frac{1}{N_1+1}\right]\right) \end{aligned} \tag{3.140}$$

Now, from (3.136) we have, when $p_g \neq 0, 1$

$$\begin{aligned} \mathbb{E}_{N_0|g}\left[\frac{1}{N_0+1}\right]\mathbb{E}_{N_1|g}\left[\frac{1}{N_1+1}\right] &= \frac{1-(1-p_g)^{n+1}}{(n+1)p_g}\cdot\frac{1-p_g^{n+1}}{(n+1)(1-p_g)} \\ &\le \frac{1}{n+1} \end{aligned} \tag{3.141}$$

where (3.141) follows from noting that the function $\frac{1-(1-x)^{n+1}}{x}\cdot\frac{1-x^{n+1}}{1-x} \le n+1$ for all $0 < x < 1$. Moreover, when p_g is either 0 or 1 we have $\mathbb{E}_{N_0|g}\left[\frac{1}{N_0+1}\right]\mathbb{E}_{N_1|g}\left[\frac{1}{N_1+1}\right] = \frac{1}{n+1}$, and putting the aforementioned two cases together we have

$$\begin{aligned} h(\Theta_0|X^n,Y^n,G) + h(\Theta_1|X^n,Y^n,G) &\le \log(\pi e/2) + \sum_{g\in\mathscr{G}} \frac{1}{2}\log\left(\frac{1}{n+1}\right) \\ &\le \log(\pi e/2) - \frac{1}{2}\log(n+1) \end{aligned} \tag{3.142}$$

Substituting the bound (3.142) into (3.127) yields

$$I(\Theta_0,\Theta_1,G;Y^n|X^n) \ge m + \log(n+1) - H(G|X^n,Y^n) - \log(4\pi e). \tag{3.143}$$

Now, we have for any estimator $\widehat{G}(X^n,Y^n)$ of G,

$$H(G|X^n,Y^n) \leq H(G|\widehat{G}(X^n,Y^n)) \tag{3.144}$$

$$\leq \mathbb{P}(G \neq \widehat{G}(X^n,Y^n))m+1 \tag{3.145}$$

where (3.144) follows by the data processing inequality, and (3.145) follows from the Fano inequality [96]. We now provide an estimator $\widehat{G}(X^n,Y^n)$ for which the error probability $\mathbb{P}(G \neq \widehat{G}(X^n,Y^n)) = o(1)$. Given X^n,Y^n, we define

$$\widehat{p}_l := \frac{\sum_{i=1}^n \mathbb{1}\{X_i = l, Y_i = 1\} + 1/2}{\sum_{i=1}^n \mathbb{1}\{X_i = l\} + 1},\ l \in \{1,\ldots,m\}. \tag{3.146}$$

Let $\widehat{p}_{\min} := \min_l \widehat{p}_l$ and $\widehat{p}_{\max} := \max_l \widehat{p}_l$. The estimator $\widehat{G}(X^n,Y^n) \in \mathscr{G}$ is then defined as

$$\widehat{G}(l) = \begin{cases} 0 & \text{if } \widehat{p}_l \leq \frac{\widehat{p}_{\max}+\widehat{p}_{\min}}{2} \\ 1 & \text{otherwise.} \end{cases}$$

The probability of error of this estimator can now be bounded as follows.

Lemma 9. *We have*

$$\mathbb{P}(\widehat{G}(X^n,Y^n) \neq G) \leq \frac{2}{2^m} + \left(1 - \frac{2}{2^m}\right) 2\sqrt{e}me^{-3n/100m}. \tag{3.147}$$

The proof of Lemma 9 is provided in the next subsection of Appendix D.

Using Lemma 9 in (3.145) and substituting this into (3.143), since $\frac{2m}{2^m} \leq 1$, we have

$$I(\Theta_0,\Theta_1,G;Y^n|X^n) \geq m + \log(n+1) - 2\sqrt{e}m^2e^{-3n/100m} - \log(\pi e) \tag{3.148}$$

as required.

Proof of Lemma 9

We will denote $\widehat{G}(X^n,Y^n)$ simply by $\widehat{G}$ for convenience.

Let $g=0$ and $g=1$ denote the all-0 and all-1 functions respectively (i.e. $g(x)=0/1$ for all $x\in[m]$). We have

$$\begin{aligned}\mathbb{P}(\widehat{G}\neq G) &= \frac{1}{2^m}\sum_{g\in\mathscr{G}}\mathbb{P}(\widehat{G}\neq g|G=g)\\ &= \frac{1}{2^m}\Big(\mathbb{P}(\widehat{G}\neq 0|G=0)+\mathbb{P}(\widehat{G}\neq 1|G=1)\Big)+\frac{1}{2^m}\sum_{g\in\mathscr{G}}\mathbb{P}(\widehat{G}\neq g|G=g)\\ &\leq \frac{2}{2^m}+\frac{1}{2^m}\sum_{g\in\mathscr{G}\setminus\{g=0,g=1\}}\mathbb{P}(\widehat{G}\neq g|G=g) \qquad (3.149)\end{aligned}$$

Now, consider $\mathbb{P}(\widehat{G}\neq g|G=g)$ for $g\neq 0,1$ identically. Since

$$\mathbb{P}(\widehat{G}\neq g|G=g)=\mathbb{E}_{\Theta_0,\Theta_1}[\mathbb{P}(\widehat{G}\neq g|G=g,\Theta_0,\Theta_1)],$$

showing that for a *fixed* (g,θ_0,θ_1) with $g\neq 0,1$ identically and $\theta_1-\theta_0\geq\frac{1}{2}$, with $X_i\sim\text{Unif}\{[m]\}$ i.i.d. and $Y_i|(X_i=l)\sim\text{Bernoulli}(\theta_{g(l)}), i\in[n]$, $\mathbb{P}(\widehat{G}\neq g)\leq 2\sqrt{e}me^{-3n/100m}$ suffices to prove the lemma (recall that the $\theta_1-\theta_0\geq\frac{1}{2}$ condition arises due to the choice of P_H and more specifically the distribution of (Θ_0,Θ_1) in (3.123), which has zero density over the region $\theta_1-\theta_0<\frac{1}{2}$). We now prove this statement.

We claim that

$$\left\{\cap_{l=1}^m(|\widehat{p}_l-\theta_{g(l)}|\leq 1/8)\right\}\subseteq\left\{\widehat{G}=g\right\}. \qquad (3.150)$$

To see this, note that if the event $\left\{\cap_{l=1}^m(|\widehat{p}_l-\theta_{g(l)}|\leq 1/8)\right\}$ occurs, we have $\theta_1-1/8\leq\widehat{p}_{\max}\leq\theta_1+1/8$ and $\theta_0-1/8\leq\widehat{p}_{\min}\leq\theta_0+1/8$ (recall that there is at least one l such that $g(l)=0$, and similarly at least one l such that $g(l)=1$) and subsequently, adding these two inequalities,

$$\frac{\theta_0+\theta_1}{2}-1/8 \le \frac{\widehat{p}_{\max}+\widehat{p}_{\min}}{2} \le \frac{\theta_0+\theta_1}{2}+1/8 \tag{3.151}$$

But, since $\theta_1-\theta_0 \ge 1/2$, we have $\theta_0+1/8 \le \frac{\theta_0+\theta_1}{2}-1/8$ and similarly $\theta_1-1/8 \ge \frac{\theta_0+\theta_1}{2}+1/8$. This, together with (3.151) implies that

$$\theta_0+1/8 \le \frac{\widehat{p}_{\max}+\widehat{p}_{\min}}{2} \le \theta_1-1/8$$

Since the event $\left\{\cap_{l=1}^m(|\widehat{p}_l-\theta_{g(l)}| \le 1/8)\right\}$ occurring implies that if $g(l)=0, \widehat{p}_l \le \theta_0+1/8$, which implies that in this case $\widehat{p}_l \le \frac{\widehat{p}_{\max}+\widehat{p}_{\min}}{2}$ and so $\widehat{G}(l)=g(l)=0$. Similarly, when $g(l)=1$, $\widehat{G}(l)=g(l)=1$.

Going back to (3.150), we have

$$\begin{aligned}
&\mathbb{P}\left(\cap_{l=1}^m|\widehat{p}_l-\theta_{g(l)}| \le 1/8\right) \le \mathbb{P}(\widehat{G}=g) \\
&\qquad \implies \mathbb{P}(\widehat{G}\neq g) \le \mathbb{P}\left(\cup_{l=1}^m|\widehat{p}_l-\theta_{g(l)}| > 1/8\right) \\
&\qquad \implies \mathbb{P}(\widehat{G}\neq g) \le \sum_{l=1}^m \mathbb{P}\left(|\widehat{p}_l-\theta_{g(l)}| > 1/8\right) \qquad (3.152)
\end{aligned}$$

where (3.152) follows from the union bound. Consider now $\mathbb{P}\left(|\widehat{p}_m-\theta_{g(m)}| > 1/8\right)$. Without loss of generality, we may assume that $g(m)=1$. Introducing the notation[4]

$$N_l := \sum_{i=1}^n \mathbb{1}\{X_i=l\},\ l\in\{1,\dots,m\} \tag{3.153}$$

$$K_l := \sum_{i=1}^n \mathbb{1}\{Y_i=1, X_i=l\},\ l\in\{1,\dots,m\}. \tag{3.154}$$

[4] This notation is independent of and not to be confused with the definitions of N_0, K_0, N_1 and K_1 in the proof of Lemma 7.

we have

$$\mathbb{P}\left(|\widehat{p}_m - \theta_1| > 1/8\right) = \mathbb{P}\left(\left|\frac{K_m + 1/2}{N_m + 1} - \theta_1\right| > 1/8\right)$$
$$= \mathbb{E}_{N_m}\left[\mathbb{P}\left(\left|\frac{K_m + 1/2}{N_m + 1} - \theta_1\right| > 1/8 \middle| N_m\right)\right].$$

Recalling that $K_m|N_m \sim \text{Binomial}(N_m, \theta_1)$, a slight variation on the Hoeffding inequality yields

$$\mathbb{P}\left(\left|\frac{K_m + 1/2}{N_m + 1} - \theta_1\right| > 1/8 \middle| N_m\right) \leq 2\sqrt{e}e^{-N_m/32}. \tag{3.155}$$

Next, since $N_m \sim \text{Binomial}(n, \mathbb{P}(X = m))$ and $\mathbb{P}(X = m) = \frac{1}{m}$ by our choice of P_X, recalling the moment-generating function of the binomial random variable $\mathbb{E}[e^{tN_m}] = \left(1 - \frac{1}{m} + \frac{1}{m}e^t\right)^n$, we have

$$\mathbb{E}_{N_m}\left[\mathbb{P}\left(\left|\frac{K_m + 1/2}{N_m + 1} - \theta_1\right| > 1/8 \middle| N_m\right)\right] \leq \mathbb{E}_{N_m}[2\sqrt{e}e^{-N_m/32}]$$
$$\leq 2\sqrt{e}\left(1 - \frac{1}{m} + \frac{1}{m}e^{-1/32}\right)^n. \tag{3.156}$$

We can use the exact same procedure to establish

$$\mathbb{P}\left(|\widehat{p}_l - \theta_{g(l)}| > 1/8\right) \leq 2\sqrt{e}\left(1 - \frac{1}{m} + \frac{1}{m}e^{-1/32}\right)^n \tag{3.157}$$

for $l = 1, \ldots, m-1$. Substituting this bound into (3.152) yields

$$\begin{aligned}
\mathbb{P}(\widehat{G} \neq g) &\leq 2m\sqrt{e}\left(1 - \frac{1}{m} + \frac{1}{m}e^{-1/32}\right)^n \\
&= 2m\sqrt{e}\left(1 - \frac{(1 - e^{-1/32})}{m}\right)^n \\
&\leq 2\sqrt{e}me^{-n(1-e^{-1/32})/m} \\
&\leq 2\sqrt{e}me^{-3n/100m}.
\end{aligned} \tag{3.158}$$

3.8 Acknowledgement

This chapter, in part, includes the material in Alankrita Bhatt and Young-Han Kim, "Sequential prediction under log-loss with side information," Algorithmic Learning Theory conference, 2021. The dissertation author was a primary investigator and author of this paper.

Chapter 4

On Universal Portfolios With Continuous Side Information

4.1 Introduction

We study the classical problem of portfolio selection, formally defined as follows. Suppose that there exist $m \ge 2$ stocks in a stock market and let $\mathbf{x}_t = (x_{t1}, \ldots, x_{tm}) \in \mathbb{R}_{\ge 0}$ denote a market vector at time t, which encodes the *price relatives* of stocks on that day. That is, for each stock $i \in [m] \coloneqq \{1, \ldots, m\}$, $x_{ti} \ge 0$ is the ratio of the end price to the start price on day t. Concretely, an investment strategy a, at each day t, outputs a nonnegative weight vector $a(\cdot|\mathbf{x}^{t-1}) \in \Delta^{m-1}$ over the stocks $[m]$, upon which the investor distributes her wealth accordingly; hereafter, we use $\mathscr{B} \coloneqq \Delta^{m-1} \coloneqq \{(\widehat{\theta}_1, \ldots, \widehat{\theta}_m) \in \mathbb{R}^m_{\ge 0} \colon \sum_{i=1}^m \widehat{\theta}_i = 1\}$ to denote the standard m-simplex. That is, the multiplicative wealth gain on day t (i.e., the ratio of wealth on day t to the wealth on day $t-1$) is $\sum_{j \in [m]} a(j|\mathbf{x}^{t-1}) x_{tj}$. Thus, her cumulative wealth gain after n days becomes

$$S_n(a, \mathbf{x}^n) \coloneqq \prod_{t=1}^n \sum_{j \in [m]} a(j|\mathbf{x}^{t-1}) x_{tj} = \sum_{y^n \in [m]^n} \Big(\prod_{t=1}^n a(y_t|\mathbf{x}^{t-1}) \Big) \mathbf{x}(y^n), \tag{4.1}$$

where $\mathbf{x}(y^n) \coloneqq x_{1y_1} \cdots x_{ny_n}$ denotes the wealth gain of an extreme investment strategy that puts all money to the stock y_t on day t, and the second equality follows from the distributive law.

An investor's goal is to design an investment strategy that maximizes her cumulative wealth $S_n(a, \mathbf{x}^n)$. For a stock market where $\mathbf{x}^n$ are i.i.d., it is known that the log-optimal portfolio $\boldsymbol{\theta}^\star$ that maximizes $\mathsf{E}[\log \boldsymbol{\theta}^T \mathbf{X}]$ is asymptotically and competitively optimal. A similar result is well-established for stationary ergodic markets, see, *e.g.* [17, Chapter 16]. The log-optimal portfolio theory with stochastic market assumptions, however, is unrealistic, as modeling a stock market could be harder than predicting the market.

As a more realistic alternative, [18] presented *universal portfolios* that asymptotically achieve the best wealth, to first order in the exponent, attained by a certain class of reference portfolios, with *no statistical assumptions* on the stock market. For the reference class, Cover considered a class of constant rebalanced portfolios (CRPs), where a CRP parameterized by a weight vector $\boldsymbol{\theta} \in \mathscr{B}$ is defined to redistribute its wealth according to $\boldsymbol{\theta}$ on every day. Note that CRPs are optimal in an i.i.d. stock market when the distribution is known.

Later, [19] extended the theory to a setup where a discrete side information sequence is causally available to an investor; in practice, the side information sequence can be thought to encode an external information that may help predict the stock market. They proposed a variation of [18]'s universal portfolios that asymptotically achieves the best wealth attained by a class of *state-wise* CRPs that may play different weight vectors according to the side information.

Taking one step further, in this paper, we consider a more challenging scenario in which a side information sequence $z^n \in \mathscr{Z}^n$ is continuous-valued, which could even be the (truncated) market history itself. A reference portfolio we aim to compete with is parameterized by a state-wise CRP and a *state function* $g\colon \mathscr{Z} \to [S]$ for some $S \geq 2$ and plays the state-wise CRP according to the state sequence $g(z^n) \coloneqq g(z_1) \dots g(z_n)$, where we assume a class of state functions $\mathscr{G}$ from which g is drawn; note that larger the $\mathscr{G}$, the richer the reference class. This flexibility in the class $\mathscr{G}$ and the choice of continuous side information sequence may hugely enlarge the capacity of the competitor class since it can capture a variety of investment strategies. As a simple example, consider a portfolio strategy that selects the state based on whether the price relative of the first stock yesterday $\mathbf{x}_{t,1} \geq \beta$ or not for a (variable) threshold β. This falls into this enlarged class with $z_t = \mathbf{x}_{t-1}$.

As the main result, we propose a new investment strategy that asymptotically achieves the same wealth attained by the best state-constant rebalanced portfolios with a state function drawn from a class of functions of finite Natarajan dimension, under a mild regularity condition on the stochasticity of the side information sequence Z^n. The proposed strategy is based on a generalization of a universal probability assignment scheme recently proposed by [97]. Note that we assume no transaction costs and that the investor's actions do not affect the market.

The rest of the paper is organized as follows. In Section 4.2, we review universal portfolios without and with discrete side information, highlighting the connection between universal compression (or probability assignment) and universal portfolios. Section 4.3 described the proposed algorithm and a crude approximation algorithm for its simulation, together with some concrete examples of side information sequence. We present the proof of the main theorem in Section 4.4. We conclude with discussing related work in Section 4.5. All deferred proofs and technical discussions can be found in Appendices.

4.2 A Review of Universal Portfolio Theory

4.2.1 Universal Portfolios

In his seminal work, [18] set an ambitious goal that aims to design an investment strategy b to compete with the best strategy in a class $\mathscr{A}$ of investment strategies for any stock market $\mathbf{x}^n$, in the sense that it minimizes the worst-case regret

$$\mathsf{Reg}_n^{\mathsf{port}}(b,\mathscr{A}) := \sup_{\mathbf{x}^n} \sup_{a\in\mathscr{A}} \log \frac{S_n(a,\mathbf{x}^n)}{S_n(b,\mathbf{x}^n)}.$$

We call a portfolio b *universal with respect to* $\mathscr{A}$ if $\mathsf{Reg}_n^{\mathsf{port}}(b,\mathscr{A}) = o(n)$, i.e., in words, b achieves the same exponential wealth growth rate attained by the best strategy in $\mathscr{A}$ chosen in hindsight with observed market. Remarkably, Cover constructed a universal portfolio with respect to the class of CRPs and established its universality. Cover's theory is based on the key observation that competing against CRPs in portfolio optimization is equivalent to competing against i.i.d. Bernoulli models in log-loss prediction problem. In what follows, we describe this relationship in a general form beyond between i.i.d. probabilities and CRPs.

For any sequential probability assignment scheme $q(\cdot|y^{t-1}) \in \mathscr{B}$ (where $y_i \in [m]$) the *probability induced portfolio* $a = \phi(p)$ is defined as

$$a(j|\mathbf{x}^{t-1}) := \frac{\sum_{y^{t-1}\in[m]^{t-1}} p(y^{t-1}j)\mathbf{x}(y^{t-1})}{\sum_{y^{t-1}\in[m]^{t-1}} p(y^{t-1})\mathbf{x}(y^{t-1})}. \tag{4.2}$$

Note that if p is an i.i.d. probability, i.e., $p(\cdot|y^{t-1}) = \boldsymbol{\theta} \in \mathscr{B}$, it is easy to check from the expression (4.2) that the corresponding portfolio $\phi(p)$ is the CRP parameterized by $\boldsymbol{\theta}$; thus the class of CRPs $\mathscr{A}^{\mathsf{CRP}}$ is $\phi(\mathscr{P}^{\otimes})$, where we use $\mathscr{P}^{\otimes}$ to denote the class of i.i.d. probabilities.

A peculiar property of a probability induced portfolio $a = \phi(p)$ is that the daily gain can be written as

$$\sum_{y_t\in[m]} a(y_t|\mathbf{x}^{t-1})\mathbf{x}_t(y_t) = \frac{\sum_{y^t} p(y^t)\mathbf{x}(y^t)}{\sum_{y^{t-1}} p(y^{t-1})\mathbf{x}(y^{t-1})},$$

and thus by telescoping, the cumulative wealth gain (4.1) becomes

$$S_n(\phi(p), \mathbf{x}^n) = \sum_{y^n\in[m]^n} p(y^n)\mathbf{x}(y^n). \tag{4.3}$$

In view of this expression, a probability induced portfolio can be interpreted as a *fund-of-funds*, i.e., a mixture of the extremal portfolios with weights $p(y^n)$.

As alluded to earlier, there is an intimate connection between the portfolio optimization with respect to a class of probability induced portfolios and the corresponding log-loss prediction problem. In the log-loss prediction problem, given a class of probabilities $\mathscr{P}$, we define the worst-case regret of a probability q with respect to $\mathscr{P}$ as

$$\mathsf{Reg}_n^{\mathsf{prob}}(q, \mathscr{P}) = \sup_{y^n} \sup_{p\in\mathscr{P}} \log \frac{p(y^n)}{q(y^n)} \tag{4.4}$$

and call a probability q *universal* with respect to $\mathscr{P}$ if $\mathsf{Reg}_n^{\mathsf{prob}}(q, \mathscr{P}) = o(n)$. The following proposition shows that the portfolio optimization with respect to $\phi(\mathscr{P})$ is no more difficult than the corresponding log-loss prediction problem with respect to $\mathscr{P}$.

Proposition 7. *For any probability q and any class of probability assignments $\mathscr{P}$, we have*

$$\mathsf{Reg}_n^{\mathsf{port}}(\phi(q),\phi(\mathscr{P})) \le \mathsf{Reg}_n^{\mathsf{prob}}(q,\mathscr{P}).$$

Proof. We first recall (4.3) that the cumulative wealth of the probability induced portfolio $\phi(p)$ is written as $S_n(\phi(p),\mathbf{x}^n) = \sum_{y^n} p(y^n)\mathbf{x}(y^n)$. Hence, for any probability q, we can write

$$\begin{aligned}\mathsf{Reg}_n^{\mathsf{port}}(\phi(q),\phi(\mathscr{P})) = \sup_{\mathbf{x}^n}\sup_{p\in\mathscr{P}} \frac{S_n(\phi(p),\mathbf{x}^n)}{S_n(\phi(q),\mathbf{x}^n)} &= \sup_{\mathbf{x}^n}\sup_{p\in\mathscr{P}} \frac{\sum_{y^n} p(y^n)\mathbf{x}(y^n)}{\sum_{y^n} q(y^n)\mathbf{x}(y^n)} \\ &\overset{(a)}{\le} \sup_{p\in\mathscr{P}}\max_{y^n} \frac{p(y^n)}{q(y^n)} = \mathsf{Reg}_n^{\mathsf{prob}}(q,\mathscr{P}),\end{aligned}$$

where (a) follows by Lemma 10 below. □

Lemma 10 ([17], Lemma 16.7.1)**.** *Let $a_1,\ldots,a_n,b_1,\ldots,b_n$ be nonnegative real numbers. Then, defining $0/0=0$, we have $\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \le \max_{j\in[n]} \frac{a_j}{b_j}$.*

A direct implication of this statement is that if a probability assignment q is universal with respect to $\mathscr{P}$ for the log-loss prediction problem, then the induced portfolio $\phi(q)$ is universal with respect to $\phi(\mathscr{P})$. If we consider the class of all i.i.d. probabilities $\mathscr{P}^{\otimes}$, it is well known that the Laplace probability assignment $q_{\mathrm{L}}(y^n) := \int_{\mathscr{B}} \mu(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(y^n)\,\mathrm{d}\boldsymbol{\theta}$ is universal for $\mathscr{P}^{\otimes}$, where $\mu(\boldsymbol{\theta})$ is the uniform density over $\mathscr{B}$ and $p_{\boldsymbol{\theta}}(y^n)$ is the i.i.d. probability with parameter $\boldsymbol{\theta} = (\widehat{\theta}_1,\ldots,\widehat{\theta}_m) \in \mathscr{B}$, i.e., $p_{\boldsymbol{\theta}}(y^n) := \prod_{i=1}^n \widehat{\theta}_{y_n} = \prod_{j=1}^m \widehat{\theta}_j^{k_j}$ with $k_i = |\{t\colon y_t = i\}|$.[1] Indeed, we have:

[1] We remark that while the Krichevsky–Trofimov (KT) probability assignment q_{KT} is universal with an optimal constant in the regret, we consider q_{L} for simplicity throughout this paper.

Lemma 11 ([98], Chapter 9)**.**

$$\sup_{\boldsymbol{\theta}\in\mathscr{B}} \sup_{y^n\in[m]^n} \log\frac{p_{\boldsymbol{\theta}}(y^n)}{q_{\mathrm{L}}(y^n)} \le m\log n.$$

Hence, $\phi(q_{\mathrm{L}})$ is a universal portfolio for $\mathscr{A}^{\mathsf{CRP}} = \phi(\mathscr{P}^{\otimes})$—this is [18]'s universal portfolio. We remark that the universal portfolio $\phi(q_{\mathrm{L}})$ can be expressed as

$$\phi(q_{\mathrm{L}})(\cdot|\mathbf{x}^{t-1}) = \frac{\int_{\mathscr{B}} \boldsymbol{\theta} S_{t-1}(\boldsymbol{\theta},\mathbf{x}^{t-1})\mu(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{\int_{\mathscr{B}} S_{t-1}(\boldsymbol{\theta},\mathbf{x}^{t-1})\mu(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}},$$

and is thus also known as the μ-weighted portfolio.

4.2.2 Universal Portfolios with Discrete Side Information

Let us now consider a scenario at each time t, the investor is additionally given a discrete side information $w_t \in [S]$ for some $S \ge 1$ and chooses a portfolio $a(\cdot|\mathbf{x}^{t-1};w^t) \in \mathscr{B}$, as considered by [19]. Since the investor's multiplicative wealth gain is $\sum_{y\in[m]} a(y|\mathbf{x}^{t-1};w^t)\mathbf{x}_t(y)$, similar to the no-side-information setting, the cumulative wealth factor is

$$S_n(a,\mathbf{x}^n;w^n) \coloneqq \prod_{t=1}^{n} \sum_{j\in[m]} a(j|\mathbf{x}^{t-1};w^t)x_{tj} \tag{4.5}$$

and we define the worst-case regret as

$$\mathsf{Reg}_n^{\mathsf{port}}(b,\mathscr{A};w^n) \coloneqq \sup_{a\in\mathscr{A}} \sup_{\mathbf{x}^n} \log\frac{S_n(a,\mathbf{x}^n;w^n)}{S_n(b,\mathbf{x}^n;w^n)}$$

for a class $\mathscr{A}$ of portfolios that also adapt to w^n. Concretely, as a natural extension of CRPs, we consider a class of state-constant rebalanced portfolios (state-CRPs), denoted as $\mathscr{A}_S^{\mathsf{CRP}}$, where a state-CRP parameterized by a S-tuple $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_S) \in \mathscr{B}^S$ plays a portfolio $\boldsymbol{\theta}_{w_t}$ at each time t.

Paralleling the connection between probability and portfolio in the no-side-information case, we can also define a probability induced portfolio in this setting. In the log-loss prediction with a causal side information sequence, a learner is asked to assign a probability $p(\cdot|y^{t-1};w^t)$ over $[m]$ based on the causal information, i.e., past sequence y^{t-1} and the side information sequence w^t. Here, we use $p(y^n\|w^n) \coloneqq \prod_{t=1}^n p(y_t|y^{t-1};w^t)$ to denote the joint probability over y^n given w^n. The probability induced portfolio $a = \phi(p)$ is then defined as

$$a(j|\mathbf{x}^{t-1};w^t) \coloneqq \frac{\sum_{y^{t-1}} p(y^{t-1}j\|w^t)\mathbf{x}(y^{t-1})}{\sum_{y^{t-1}} p(y^{t-1}\|w^{t-1})\mathbf{x}(y^{t-1})}, \tag{4.6}$$

and as in the no-side information setting, we can write

$$S_n(\phi(p),\mathbf{x}^n;w^n) = \sum_{y^n} p(y^n\|w^n)\mathbf{x}(y^n).$$

For example, the class of S-state-CRPs $\mathscr{B}_S^{\mathsf{CRP}}$ is induced by the class of all S-state i.i.d. probabilities $\mathscr{P}_S^{\otimes}$, i.e., $\mathscr{B}_S^{\mathsf{CRP}} = \phi(\mathscr{P}_S^{\otimes})$. To see this, note that every S-state-CRP parameterized by $\boldsymbol{\theta}_{1:S} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_S)$ is the portfolio induced by the state-wise i.i.d. probability assignment $p_{\boldsymbol{\theta}_{1:S}}(y^n\|w^n) \coloneqq \prod_{t=1}^n p_{\boldsymbol{\theta}_{w_t}}(y_t)$. Moreover, as stated in Proposition 7, solving the log-loss prediction problem suffices for the probability optimization with side information with respect to a class of probability induced portfolios. The proof can be found in Appendix 4.6.2.

Proposition 8. *For any probability assignment q and any class of probability assignment schemes $\mathscr{P}$ with side information sequence w^n, we have*

$$\mathsf{Reg}_n^{\mathsf{port}}(\phi(q),\phi(\mathscr{P});w^n) \le \mathsf{Reg}_n^{\mathsf{prob}}(q,\mathscr{P};w^n),$$

where we define

$$\mathsf{Reg}_n^{\mathsf{prob}}(q,\mathscr{P};w^n) \coloneqq \sup_{p\in\mathscr{P}} \max_{y^n} \log\frac{p(y^n\|w^n)}{q(y^n\|w^n)}.$$

Note that for the class of S-state-wise i.i.d. distributions $\mathscr{P}_S^{\otimes}$, the state-wise extension of the Laplace probability assignment $q_{\mathrm{L};S}$ that assigns

$$q_{\mathrm{L};S}(y^n \| w^n) := \prod_{s=1}^{S} q_{\mathrm{L}}(y^n(s;w^n)), \tag{4.7}$$

where $y^n(s;w^n) = (y_i\colon w_i = s, i \in [n])$, is universal, and so $\phi(q_{\mathrm{L};S})$ is universal for $\mathscr{A}_S^{\mathsf{CRP}} = \phi(\mathscr{P}_S^{\otimes})$—this is [19]'s universal portfolio.

4.3 Main Results

4.3.1 Universal Portfolios with Continuous Side Information

We now consider our main setting where a side information sequence $z^n \in \mathscr{Z}^n$ is continuous-valued. For example, in this setup, one may take z_t as a suffix of the market history $\mathbf{x}_{t-k}^{t-1}$ for some $k \geq 1$. As described earlier in the introduction, we aim to design a universal portfolio that competes against a class of state-CRPs that adapts to the sequence $g(w^n)$, where g is a state function $g\colon \mathscr{Z} \to [S]$ assumed to belong to a class of functions $\mathscr{G}$. Note that a singleton $\mathscr{G} = \{g\}$ recovers the setting of [19]. Our goal is to design a portfolio that is universal for a largest possible $\mathscr{G}$ with a minimal assumption on the side information sequence. In this paper, we will assume that the *Natarajan dimension* [99] of $\mathscr{G}$, denoted as $\mathrm{Ndim}(\mathscr{G})$, is finite. The Natarajan dimension can be seen as a generalization of the classic VC dimension, when the function class under consideration is not binary—for completeness a formal definition is provided in Appendix 4.6.1.

Leveraging the established connection between probability and portfolio, we continue to view the class of state-wise CRPs $\mathscr{B}_S^{\mathsf{CRP}} = \phi(\mathscr{P}_S^{\otimes})$ as the class of portfolios induced by $\mathscr{P}_S^{\otimes}$ and describe the problem in an abstract setting. For a class of probability induced portfolios with (discrete) side information $\mathscr{A} = \phi(\mathscr{P})$ and a class of state functions $\mathscr{G}$, our goal is to design a strategy b that achieves a sublinear worst-case regret

$$\mathsf{Reg}_n^{\mathsf{port}}(b;\mathscr{A},\mathscr{G};\mathbf{x}^n,z^n) \coloneqq \sup_{g\in\mathscr{G}}\sup_{a\in\mathscr{A}} \log\frac{S(a,\mathbf{x}^n;g(z^n))}{S(b,\mathbf{x}^n;z^n)}.$$

Similar to the universal portfolios with discrete side information, a universal portfolio can be readily induced by a universal probability with respect to a continuous side information sequence with an unknown state function, based on the following statement, whose proof is deferred to Appendix 4.6.2.

Proposition 9. *For any $\mathbf{x}^n$ and z^n, we have*

$$\mathsf{Reg}_n^{\mathsf{port}}(\phi(q);\phi(\mathscr{P}),\mathscr{G};\mathbf{x}^n,z^n) \le \mathsf{Reg}_n^{\mathsf{prob}}(q;\mathscr{P},\mathscr{G};z^n),$$

where

$$\mathsf{Reg}_n^{\mathsf{prob}}(q;\mathscr{P},\mathscr{G};z^n) \coloneqq \sup_{g\in\mathscr{G}}\sup_{p\in\mathscr{P}}\max_{y^n}\log\frac{p(y^n\|g(z^n))}{q(y^n\|z^n)}.$$

In this work, we specifically plug-in an extended version of the universal probability assignment $q^*_{\mathscr{G}}$ proposed by [97], which was designed for $m=2,S=2$ with regret guarantee established when y^n is random and the side information sequence Z^n is i.i.d.. We will extend their scheme for arbitrary m and S with a guarantee for adversarial y^n and non-i.i.d. Z^n.

Below, we further assume that a side information sequence Z^n is stochastic with distribution P_{Z^n} which may be arbitrarily correlated with the stock market $\mathbf{X}^n$; the universality is established with respect to the expected worst-case regret

$$\overline{\mathsf{Reg}}_n^{\mathsf{port}}(b;\mathscr{A},\mathscr{G}) \coloneqq \mathsf{E}\big[\mathsf{Reg}_n^{\mathsf{port}}(b;\mathscr{A},\mathscr{G};\mathbf{X}^n,Z^n)\big],$$

where the expectation is over a joint distribution $\mathbf{P}_{\mathbf{X}^n,Z^n}$. We remark that it is unclear whether the required stochastic assumptions on Z^n in Theorem 3 are an artifact of our analysis or whether they can be completely removed and universality can be established for individual sequences z^n.

We leave this question for future work; see also Section 4.5.

Proposed Strategy.

Firstly, for any $\tilde{n} \in \mathbb{N}$ and any $\tilde{z}^{\tilde{n}} \in \mathscr{Z}^{\tilde{n}}$, let $\{\widetilde{g}_1, \ldots, \widetilde{g}_\ell\} \subset \mathscr{G}$ be a *minimal empirical covering* of $\mathscr{G}$ with respect to $\tilde{z}^{\tilde{n}}$, i.e., a set of functions such that $\{\widetilde{g}_i(\tilde{z}^{\tilde{n}}) \colon i \in [\ell]\} = \{g(\tilde{z}^{\tilde{n}}) \colon g \in \mathscr{G}\}$ with the minimum possible size $\ell = \ell(\tilde{z}^{\tilde{n}})$. Then, we define a mixture probability assignment

$$q_{\mathscr{G};\tilde{z}^{\tilde{n}}}(y^i \| z^i) := \frac{1}{\ell} \sum_{j=1}^{\ell} q_{\mathrm{L};S}(y^i \| \widetilde{g}_j(z^i)) \tag{4.8}$$

with respect to the empirical covering, and define the induced sequential probability assignment

$$q_{\mathscr{G};\tilde{z}^{\tilde{n}}}(y_i | y^{i-1}; z^i) := \frac{q_{\mathscr{G};\tilde{z}^{\tilde{n}}}(y^i \| z^i)}{q_{\mathscr{G};\tilde{z}^{\tilde{n}}}(y^{i-1} \| z^{i-1})}.$$

The proposed probability assignment $q^*_{\mathscr{G}}$ is then defined as follows. First, we split the n time steps into $\lceil \log_2 n \rceil$ epochs: starting from $j = 1$, define the j-the epoch to consist of the time steps $2^{j-1} + 1 \le i \le 2^j$. So, the first epoch consists of z_2, the second epoch consists of z_3^4, the third epoch consists of z_5^8 and so on. Then,

- For $i = 1$, $q^*_{\mathscr{G}}(\cdot | z_1) := 1/m$;
- For $i \ge 2$, if $2^{j-1} + 1 \le i \le 2^j$, i.e., if the time step i falls within the j-th epoch, then

  $$q^*_{\mathscr{G}}(y_i | y^{i-1}; z^i) := \frac{q_{\mathscr{G};z^{2^{j-1}}}(y^i_{2^{j-1}+1} \| z^i_{2^{j-1}+1})}{q_{\mathscr{G};z^{2^{j-1}}}(y^{i-1}_{2^{j-1}+1} \| z^{i-1}_{2^{j-1}+1})},$$

  where we define $q_{\mathscr{G};z^{2^{j-1}}}(\emptyset \| \emptyset) = 1$ by convention.

Concretely, the probability assigned over y^n given z^n for some $n \in (2^{J-1}, 2^J]$ is

$$\begin{aligned} q^*_{\mathscr{G}}(y^n \| z^n) &= \prod_{i=1}^{n} q^*_{\mathscr{G}}(y_i | y^{i-1}; z^i) \\ &= q_{\mathscr{G};\emptyset}(y_1 \| z_1) q_{\mathscr{G};z_1}(y_2 \| z_2) q_{\mathscr{G};z^2}(y_3^4 \| z_3^4) \cdots q_{\mathscr{G};z^{2^{J-1}}}(y^n_{2^{J-1}+1} \| z^n_{2^{J-1}+1}). \end{aligned} \tag{4.9}$$

Finally, we obtain a sequential portfolio $a = \phi(q^*_{\mathscr{G}})$ via the expression (4.6).

A note on implementation.

While the main focus of this paper is to construct a provably universal portfolio with continuous side information, we also include a discussion on its Monte Carlo based simulation and an example with real stock data in Appendix 4.6.6.

4.3.2 Performance Guarantee and Examples

Given a class of S-state functions $\mathscr{G}$, we need to impose a structural condition on the sequence $Z^n \sim P_{Z^n}$ as a stochastic process. For any binary function class $\mathscr{H} \subset \{\mathscr{Z} \to \{0,1\}\}$, define

$$\rho_{\mathscr{H}}(Z^n) = \sup_{h \in \mathscr{H}} \Big| \sum_{i=1}^{n} \big(h(Z_i) - \mathsf{E}[h(Z_i)]\big) \Big|, \tag{4.10}$$

which is a well-studied quantity in the empirical process theory. Specifically, we are interested in the binary function class

$$\mathbb{1}\mathscr{G} \times \mathscr{G} := \{h \colon \mathscr{Z} \to \{0,1\} \colon h(z) = \mathbb{1}(g(z) \neq g'(z)) \text{ for } g, g' \in \mathscr{G}\}.$$

With a slight abuse of notation, we use $\rho_{\mathscr{G}\times\mathscr{G}}(Z^n)$ to denote $\rho_{\mathbb{1}\mathscr{G}\times\mathscr{G}}(Z^n)$. We now state our main result.

Theorem 3 (Asymptotic universality)**.** *For any collection of functions $\mathscr{G}$ of finite Natarajan dimension and any stationary stochastic process Z^n such that*

$$\mathsf{E}[\rho_{\mathscr{G}\times\mathscr{G}}(Z^n)] = o\Big(\frac{n}{\log^2 n}\Big), \tag{4.11}$$

the induced portfolio $\phi(q^*_{\mathscr{G}})$ *satisfies*

$$\lim_{n\to\infty}\frac{1}{n}\overline{\mathsf{Reg}}^{\mathsf{port}}(\phi(q^*_{\mathscr{G}}),\mathscr{A}_S^{\mathsf{CRP}},\mathscr{G})=0.$$

In Theorem 3, the condition $\mathsf{E}[\rho_{\mathscr{G}\times\mathscr{G}}(Z^n)]\ll\frac{n}{\log^2 n}$ on the marginal distribution P_{Z^n} is crucial in ensuring consistency of the portfolio $\phi(q^*_{\mathscr{G}})$. We now provide a few example cases of side information sequences Z^n where this requirement is satisfied. In fact, by controlling $\mathsf{E}[\rho_{\mathscr{G}\times\mathscr{G}}(Z^n)]$ we can also bound the *nonasymptotic regret* for these particularly interesting cases.

Example 1 (i.i.d. processes)**.** *When the joint distribution* $P_{\mathbf{X}^n,Z^n}$ *is such that* Z^n *is i.i.d., it is well known that* $\mathsf{E}[\rho_{\mathscr{H}}(Z^n)]\le C\sqrt{VCdim(\mathscr{H})n}$ *(for absolute constant C) for any binary class* $\mathscr{H}$ *and distribution* P_{Z^n}*; see [3, Theorem 8.3.23]. Following the same logic*[2]*, it can be shown that* $\mathsf{E}[\rho_{\mathscr{G}\times\mathscr{G}}(Z^n)]\le C\sqrt{(d\log S)n}$ *and consequently* $\overline{\mathsf{Reg}}^{\mathsf{port}}=\widetilde{O}(\sqrt{n})$.

Example 2 (β-mixing processes)**.** *The quantity* $\mathsf{E}[\rho_{\mathscr{H}}(Z^n)]$ *has also been studied for classes beyond i.i.d. sequences—in particular, [100] studied the case when* Z^n *is* β-mixing*, which we now define. For the sigma-fields* $\sigma_l:=\sigma(Z_1,\dots,Z_\ell)$ *and* $\sigma'_{l+k}:=\sigma(Z_{\ell+k},Z_{\ell+k+1},\dots,)$*, we define* $\beta_k:=\frac{1}{2}\sup\{\mathsf{E}|P(B|\sigma_l)-P(B)|:B\in\sigma'_{\ell+k},\ell\ge 1\}$ *and if* $\beta_k=O(k^{-r_\beta})$ *as* $k\to\infty$*,* r_β *is called the* β-mixing exponent*; a larger* r_β *guarantees faster mixing. We can restate the main result of [100] for the case when* $\mathscr{H}$ *has a finite VC dimension.*

Theorem 4 ([100, Corollary 3.2 and Remark (i)])**.** *Assume that a class of binary functions* $\mathscr{H}$ *is of finite VC dimension. Let* Z^n *be a stationary* β*-mixing sequence with* β*-mixing exponent* $r_\beta\in(0,1]$. *Let* $\xrightarrow{p}$ *denote convergence in probability. Then, for any given* $s\in(0,r_\beta)$*, we have*

$$n^{s/(1+s)}\frac{\rho_{\mathscr{H}}(Z^n)}{n}\xrightarrow{p}0\quad as\ n\to\infty. \tag{4.12}$$

[2] The only change to be made in the proof is in the growth function—rather than $\left(\frac{en}{d}\right)^d$, the growth function in this case is $\le(S^2n)^{2d}$ by Natarajan's Lemma; see Section 4.4.1.

This theorem immediately implies that $\frac{1}{n}\mathsf{Reg}^{\mathsf{port}} \xrightarrow{p} 0$*, i.e.,* $\phi(q^*_{\mathscr{G}})$ *is universal* in probability. *We can also establish its universality* in expectation *via Theorem 3, by showing* (4.11) *under the same assumption. The proof requires an additional technical argument and thus deferred to Appendix 4.6.3.*

Example 3 (Market history $z_t = \mathbf{x}_{t-k}^{t-1}$)**.** *A canonical example of side information is the market history* $z_t = \mathbf{x}^{t-1}$ *or a truncated version of it with memory size k, i.e.,* $z_t = \mathbf{x}_{t-k}^{t-1}$*. In this case, if the stock market* $(\mathbf{x}_t)$ *itself is k-th order Markov, then under an additional mild regularity condition, we can show a faster rate* $\overline{\mathsf{Reg}}^{\mathsf{port}} \le \widetilde{O}(\sqrt{n})$ *than implied by the previous example; see Appendix 4.6.5.*

4.4 Proofs

In this section, we prove Theorem 3. We first note that the probability assignment $q^*_{\mathscr{G}}$ used to derive the proposed portfolio guarantees the following regret bound.

Theorem 5. *For probability assignment* $q^*_{\mathscr{G}}$ *if the Natarajan dimension (denoted by* $\mathrm{Ndim}(\mathscr{G}) = d$*) of* $\mathscr{G}$ *is finite and* $Z^n \sim P_{Z^n}$ *is stationary, we have*[3]

$$\mathsf{E}\Big[\sup_{g\in\mathscr{G}}\sup_{p\in\mathscr{P}_S^{\otimes}}\sup_{y^n\in[m]^n}\log\frac{p(y^n\|g(Z^n))}{q^*_{\mathscr{G}}(y^n\|Z^n)}\Big] \tag{4.13}$$

$$\le S(d+m)(\log^2 n) + 2.5Sm\sum_{j=0}^{\log n-1} j\mathsf{E}[\rho_{\mathscr{G}\times\mathscr{G}}(Z^{2^j})].$$

We will first prove Theorem 5; Theorem 3 then follows as a corollary of Theorem 5 via the established connection between a probability and the induced portfolio in Proposition 9.

4.4.1 Proof of Theorem 5

Note that the key building block of the proposed probability assignment scheme $q^*_{\mathscr{G}}$ is $q_{\tilde{z}^n}(y^i\|z^i)$ defined in (4.8), the uniform mixture based on a minimal empirical covering of $\mathscr{G}$ with

[3] Here, $\log n$ is assumed to be an integer for simplicity, which can be easily rectified at the cost of an absolute constant factor in the regret; see Section 4.4.1.

respect to $\tilde{z}^n$. The proof consists of three steps. In Step 1, we first consider the simplest case where the whole side information sequence z^n is provided *noncausally* by an oracle, where we can use z^n as $\tilde{z}^n$ to build the empirical covering. We then analyze the performance of $q_{\tilde{z}^n}(y^i \| z^i)$ for an arbitrary auxiliary sequence $\tilde{z}^n$ in Step 2. Finally, in Step 3, we analyze $q^*_{\mathscr{G}}$ based on the analysis of $q_{\tilde{z}^n}(y^i \| z^i)$.

Step 1. Side Information Given Noncausally

Suppose that z^n is available noncausally so that it can be used to construct a minimal empirical covering in $q_{z^n}(y^i \| z^i)$ for $i \in [n]$. First, note that since $|\{(g(z^n)\colon g \in \mathscr{G}\}| \le S^n$, we can construct an empirical covering $\{g_1, \ldots, g_\ell\}$ of $\mathscr{G}$ with respect to z^n with $\ell \le S^n$. Assuming $\mathrm{Ndim}(\mathscr{G}) = d < \infty$, however, we can even do so with $\ell \le (S^2 n)^d$ by Natarajan's Lemma [99, Lemma 29.4]. Hence, for the mixture probability assignment $q_{\tilde{z}^n}(y^i \| z^i)$ defined in (4.8) with $\tilde{z}^n \leftarrow z^n$, i.e.,

$$q_{z^n}(y^i \| z^i) = \frac{1}{\ell} \sum_{j=1}^{\ell} q_{\mathrm{L};S}(y^i \| g_j(z^i)),$$

it readily follows that for any $g \in \mathscr{G}$,

$$\sup_{p \in P_S^{\otimes}} \sup_{y^n \in [m]^n} \log \frac{p(y^n \| g(z^n))}{q_{z^n}(y^n \| z^n)} \le d \log(S^2 n) + Sm \log n \tag{4.14}$$

by invoking that $\ell \le (S^2 n)^d$ and applying the regret bound for the m-ary Laplace probability assignment in Lemma 11 for each state.

Step 2. Auxiliary Side Information Given Noncausally

We now analyze the mixture probability $q_{\tilde{z}^n}(y^n \| z^n)$ for an arbitrary auxiliary sequence $\tilde{z}^n$, possibly being different from z^n. Intuitively, the sequence $\tilde{z}^n$ will also reduce the class $\mathscr{G}$ to at most $(S^2 n)^d$ functions, and if z^n and $\tilde{z}^n$ are "not too far apart", the two reductions each obtained by z^n and $\tilde{z}^n$ may be also close. The following lemma provides the performance of the mixture probability $q_{\tilde{z}^n}(y^n \| z^n)$ with respect to the auxiliary sequence $\tilde{z}^n$, capturing the expected gap from the intuition by the Hamming distance (denoted by $\mathrm{d_H}$) between $g(z^n)$ and $\widetilde{g}(z^n)$.

Lemma 12. *For any $\tilde{z}^n$, z^n, and $g \in \mathscr{G}$ with* $\mathrm{Ndim}(\mathscr{G}) = d < \infty$, *we have*

$$\sup_{p \in \mathscr{P}_S^{\otimes}} \sup_{y^n \in [m]} \log \frac{p(y^n \| g(z^n))}{q_{\tilde{z}^n}(y^n \| z^n)} \leq d \log(S^2 n) + Sm(\log n)(1 + 2.5 d_H(g(z^n), \widetilde{g}(z^n)))$$

$$\leq S(\log n)(d + m + 2.5 m \, \mathrm{d_H}(g(z^n), \widetilde{g}(z^n))). \tag{4.15}$$

Note that setting $\mathrm{d_H}(g(z^n), \widetilde{g}(z^n)) = 0$ recovers (4.14) as expected.

Proof. Let $p_{\boldsymbol{\theta}_{1:S}}$ be a state-wise i.i.d. probability assignment characterized by $\boldsymbol{\theta}_{1:S} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$ $\in \mathscr{B}^S$, where $\boldsymbol{\theta}_i = (\widehat{\theta}_{i1}, \widehat{\theta}_{i2}, \ldots, \widehat{\theta}_{im}) \in \mathscr{B}$ for each $i \in [S]$. For any state function $g \in \mathscr{G}$, by definition of the empirical covering, there exists a function $\widetilde{g} \in \{\widetilde{g}_1, \ldots, \widetilde{g}_\ell\}$ such that $\widetilde{g}(\tilde{z}^n) = g(\tilde{z}^n)$. Hence, we first have

$$\log \frac{p_{\boldsymbol{\theta}_{1:S}}(y^n \| g(z^n))}{q_{\tilde{z}^n}(y^n \| z^n)} \leq d \log(S^2 n) + \log \frac{p_{\boldsymbol{\theta}_{1:S}}(y^n \| g(z^n))}{q_{\mathrm{L};S}(y^n \| \widetilde{g}(z^n))}. \tag{4.16}$$

It only remains to analyze $q_{\mathrm{L};S}(y^n \| \widetilde{g}(z^n))$. For each $i \in [S]$ and $j \in [m]$, we define $n_i \coloneqq |t \colon g(Z_t) = i|$ and $k_{ij} \coloneqq |t \colon g(Z_t) = i, y_t = j|$. Moreover let $\tilde{n}_i, \tilde{k}_{ij}$ be defined in a similar way as $\tilde{n}_i \coloneqq |t \colon \widetilde{g}(Z_t) = i|$ and $\tilde{k}_{ij} \coloneqq |t \colon \widetilde{g}(Z_t) = i, y_t = j|$). We can then write

$$p_{\boldsymbol{\theta}_{1:S}}(y^n \| g(z^n)) = \prod_{s=1}^{S} \widehat{\theta}_{s1}^{k_{s1}} \ldots \widehat{\theta}_{sm}^{k_{sm}}.$$

Further, we can explicitly write the expression for the Laplace probability assignment as $q_{\mathrm{L}}(y^n) = (\binom{n+m-1}{m-1}\binom{n}{k_1,\ldots,k_m})^{-1}$, where $k_i = |\{t \colon y_t = i\}|$, and thus its state-wise extension as

$$q_{\mathrm{L};S}(y^n \| \widetilde{g}(z^n)) = \Big(\prod_{s=1}^{S} \binom{\tilde{n}_s + m - 1}{m-1} \binom{\tilde{n}_s}{\tilde{k}_{s1}, \ldots, \tilde{k}_{s,m-1}}\Big)^{-1}.$$

Now, consider

$$\begin{aligned}\log \frac{p_{\boldsymbol{\theta}_{1:S}}(y^n \| g(z^n))}{q_{\mathrm{L};S}(y^n \| \widetilde{g}(\tilde{z}^n))} &= \sum_{i=1}^{S} \log \binom{\tilde{n}_i + m - 1}{m-1} \binom{\tilde{n}_i}{\tilde{k}_{i1}, \ldots, \tilde{k}_{i,m-1}} \widehat{\theta}_{i1}^{k_{i1}} \ldots \widehat{\theta}_{im}^{k_{im}} \\ &\le Sm \log n + \sum_{i=1}^{S} \log \frac{\binom{\tilde{n}_i}{\tilde{k}_{i1},\ldots,\tilde{k}_{i,m-1}}}{\binom{n_i}{k_{i1},\ldots,k_{i,m-1}}} && (4.17) \\ &= Sm \log n + \sum_{i=1}^{S} \log \frac{\tilde{n}_i!}{n_i!} + \sum_{i=1}^{S} \sum_{j=1}^{m} \log \frac{k_{ij}!}{\tilde{k}_{ij}}, && (4.18)\end{aligned}$$

where (4.17) follows since $\binom{n_i}{k_{i1},\ldots,k_{i,m-1}} \widehat{\theta}_{i1}^{k_{i1}} \ldots \widehat{\theta}_{im}^{k_{im}} \le 1$.

Now, since for all $i \in [S]$ and $j \in [m]$, we have $|n_i - \tilde{n}_i| \le \mathrm{d_H}(g(z^n), \widetilde{g}(z^n))$ and $|k_{ij} - \tilde{k}_{ij}| \le \mathrm{d_H}(g(z^n), \widetilde{g}(z^n))$, we have $\tilde{n}_i \le n_i + \mathrm{d_H}(g(z^n), \widetilde{g}(z^n))$ and consequently $\frac{\tilde{n}_i!}{n_i!} \le \frac{(n_i + \mathrm{d_H}(g(z^n), \widetilde{g}(z^n)))!}{n_i!}$. Thus, we can invoke the exact same calculations as in [97, Propositions 5 and 6] to bound the second and third terms in (4.18) as

$$\begin{aligned}\log \frac{p_{\boldsymbol{\theta}_{1:S}}(y^n \| g(z^n))}{q_{\mathrm{L};S}(y^n \| \widetilde{g}(\tilde{z}^n))} &\le Sm \log n + S(m+3)\, \mathrm{d_H}(g(z^n), \widetilde{g}(z^n)) \log n \\ &\le Sm(\log n)(1 + 2.5 d_H(g(z^n), \widetilde{g}(z^n))), && (4.19)\end{aligned}$$

since $m \ge 2$. Plugging this into (4.16) establishes the first bound. The second bound follows by observing $\log(S^2 n) \le S \log n$. □

When Z^n is stationary as a stochastic process and if $\tilde{Z}^n$ is a statistical copy of Z^n, the following lemma shows that the Hamming distance can be bounded by $\rho_{\mathscr{G}\times\mathscr{G}}(Z^n)$, which can be controlled in expectation as $o(n/\log^2 n)$ under mild regularity conditions on P_{Z^n} and $\mathscr{G}$. The proof is deferred to Appendix 4.6.4.

Lemma 13. *If Z^n is stationary, $\tilde{Z}^n \overset{(d)}{=} Z^n$, and $\widetilde{g}(\tilde{Z}^n) = \widetilde{g}(\tilde{Z}^n)$, then*

$$\mathrm{d_H}(g(Z^n), \widetilde{g}(Z^n)) \le \rho_{\mathscr{G}\times\mathscr{G}}(Z^n) + \rho_{\mathscr{G}\times\mathscr{G}}(\tilde{Z}^n).$$

Step 3. Side Information Given Causally

In view of Lemma 13, provided that Z^n is stationary, we can *bootstrap* the history sequence to construct such an auxiliary sequence, which motivates the epoch-based construction of $q^*_{\mathscr{G}}$. That is, we split the n time steps into $\log n$ epochs[4], and define the j-the epoch to consist of the time steps $2^{j-1}+1 \leq i \leq 2^j$ starting from $j=1$, while we define $q^*_{\mathscr{G}}(\cdot|Z_1) = 1/m$ for the 0-th epoch. For $i \geq 2$, if the time step i falls within the j-th epoch, i.e., $2^{j-1}+1 \leq i \leq 2^j$, then

$$q^*_{\mathscr{G}}(y_i|y^{i-1};Z^i) = \frac{q_{Z^{2^{j-1}}}(y^i_{2^{j-1}+1}\|Z^i_{2^{j-1}+1})}{q_{Z^{2^{j-1}}}(y^{i-1}_{2^{j-1}+1}\|Z^{i-1}_{2^{j-1}+1})} \tag{4.20}$$

where we can recall the definition of $q_{Z^{2^{j-1}}}$ from (4.8). For any $p \in \mathscr{P}_S^{\otimes}$, we then have

$$\begin{aligned}
\sum_{i=1}^{n} \log \frac{p(y_i|g(Z_i))}{q^*_{\mathscr{G}}(y_i|y^{i-1};Z^i)} &\leq \sum_{i=2}^{n} \log \frac{p(y_i|g(Z_i))}{q^*_{\mathscr{G}}(y_i|y^{i-1};Z^i)} + \log m \\
&= \sum_{j=1}^{\log n} \sum_{i=2^{j-1}+1}^{2^j} \log \frac{p(y_i|g(Z_i))}{q^*_{\mathscr{G}}(y_i|y^{i-1};Z^i)} \\
&= \sum_{j=1}^{\log n} \log \frac{p(y^{2^j}_{2^{j-1}+1}\|g(Z^{2^j}_{2^{j-1}+1}))}{q_{Z^{2^{j-1}}}(y^{2^j}_{2^{j-1}+1}\|Z^{2^j}_{2^{j-1}+1})} &(4.21)\\
&\leq S(d+m)(\log^2 n) + 2.5Sm \sum_{j=0}^{\log n-1} j d_H(g(Z_1^{2^j}), g(Z_{2^j+1}^{2^{j+1}})), &(4.22)
\end{aligned}$$

where (4.21) follows by (4.20) and (4.22) follows from Lemma 12. Finally, taking supremum over y^n, p and g and expectation over Z^n leads to the desired inequality by Lemma 13. □

4.4.2 Proof of Theorem 3

By Proposition 9 and Theorem 5, we have

$$\begin{aligned}
\overline{\mathsf{Reg}}^{\mathsf{port}}(\phi(q^*_{\mathscr{G}}), \mathscr{A}_S^{\mathsf{CRP}}, \mathscr{G}) &= \mathsf{E}[\mathsf{Reg}_n^{\mathsf{port}}(\phi(q^*_{\mathscr{G}}), \mathscr{A}_S^{\mathsf{CRP}}, \mathscr{G}; \mathbf{X}^n, Z^n)] \\
&\leq \mathsf{E}[\mathsf{Reg}_n^{\mathsf{prob}}(q; \mathscr{P}, \mathscr{G}; Z^n)]
\end{aligned}$$

[4] For simplicity, we assume that $\log n$ is an integer; if not, we may "extend" the horizon of the game from n to $2^{\lceil \log n \rceil} < 2n$, and follow the same analysis incurring at most a constant factor extra in the regret bound.

$$= \mathsf{E}\Big[\sup_{g\in\mathscr{G}}\sup_{p\in\mathscr{P}}\max_{y^n}\log\frac{p(y^n\|g(Z^n))}{q(y^n\|Z^n)}\Big]$$
$$\leq S(d+m)(\log^2 n)+2.5Sm\sum_{j=0}^{\log n-1} j\mathsf{E}[\rho(Z^{2^j})],$$

where we omit the subscript in $\rho_{\mathscr{G}\times\mathscr{G}}(\cdot)$ for brevity. Since the first term in the bound is sublinear in n when d and S are fixed, it then suffices to show that $\sum_{j=0}^{\log n-1} j\mathsf{E}[\rho(Z^{2^j})]=o(n)$. Using the change of variables $n'=\log n$, observe

$$\sum_{j=0}^{\log n-1} j\mathsf{E}[\rho(Z^{2^j})]=\frac{1}{n'}\sum_{j=0}^{n'-1} j\mathsf{E}[\rho(Z^{2^j})]\frac{n'}{2^{n'}}\leq\frac{1}{n'}\sum_{j=0}^{n'-1}\frac{j^2}{2^j}\mathsf{E}[\rho(Z^{2^j})],$$

where the inequality follows since $\frac{n'}{2^{n'}}\leq\frac{j}{2^j}$ for all $j\leq n'$. Now, since

$$\frac{(\log n)^2}{n}\mathsf{E}[\rho(Z^n)]=\frac{n'^2}{2^{n'}}\mathsf{E}[\rho(Z^{2^{n'}})]\to 0$$

as $n\to\infty$ is assumed, we also have $\frac{1}{n'}\sum_{j=0}^{n'-1}\frac{j^2}{2^j}\mathsf{E}[\rho(Z^{2^j})]\to 0$ as $n'\to\infty$, by the Cesàro mean Theorem. A final change of variables concludes the proof. □

4.5 Related Work and Discussion

Portfolio selection has been a closely studied topic in information theory since the seminal work of [18] and [19], both of which established close connections between portfolio selection and the classically studied information theoretic problem of universal compression [6, 24, 101, 102]. A number of variations have been considered since, for example incorporating transaction costs [103, 104] using other probability assignments than i.i.d. [105, 106], and considering space complexity issues [107]. [108] and [109] proposed portfolio selection techniques incorporating continuous side information; however, the competitor classes considered in both are disparate from ours making the problems different.

As demonstrated, portfolio selection with side information is closely related to sequential prediction with side information and log-loss. This problem has attracted recent interest [97,110–112], with the first two focused on obtaining fundamental limits via the sequential complexities approach of [113]. More recently, the preprint of [93] proposed a mixture-based conditional density estimator, which specifically achieves $\mathsf{E}[\mathsf{Reg}^{\mathsf{prob}}] = O(\log^2 n)$ for the binary probability assignment problem with i.i.d. side information with a VC class, which tightens the regret $\tilde{O}(\sqrt{n})$ established in [97]. Therefore, it is natural to consider applying the probability assignment of [93] in hoping to relax the technical condition (4.11) and establish Theorem 3 for *all* stationary ergodic Z^n—it is known that $\mathsf{E}[\rho_{\mathscr{G}\times\mathscr{G}}(Z^n)] = o(n)$ for any stationary ergodic process (see for example [114]).

We note, however, that analyzing their method in our setting of non-i.i.d. side information sequences seems to involve a significant amount of additional work. More precisely, their analysis needs to be extended to (1) individual-sequence y^n and (2) stationary ergodic side information with a dependence of the regret on $\rho_{\mathscr{H}}(Z^n)$ similar to that of the method of [97]. In their words, we would need to relax the assumption of the data being *well-specified*. At a high level, they use a similar covering approach (with respect to the Hellinger metric over distributions) as well as a smoothing of probabilities in order to avoid unbounded likelihood ratios (we, in contrast, have used the Laplace/KT probability assignment). Using a similar epoch-based analysis they establish regret bounds in [93, Appendix D] by first upper bounding the KL divergence in terms of the Hellinger divergence and then leveraging local Rademacher complexities in conjunction with an inequality of [115]. In order to extend their method to individual-sequence y^n and stationary ergodic Z^n, one would need to either extend the aforementioned inequality to these cases, or to bypass the step of upper-bounding the KL divergence in terms of the Hellinger divergence altogether. We leave these directions of extension for future work.

4.6 Appendix

4.6.1 Definition of Natarajan Dimension

We use the definitions from [99, Definitions 29.1, 29.2].

Definition 2 (Shattering). *Let $\mathscr{G} \subset \{\mathscr{Z} \to [S]\}$. Then, a set $C \subset \mathscr{Z}$ is said to be* shattered *by the function class $\mathscr{G}$ if there exist two functions $g_0, g_1 \in \mathscr{G}$ such that*

- *For each $z \in C, g_0(z) \neq g_1(z)$, and*
- *For each $B \subset C$ there exists a function $g \in \mathscr{G}$ such that*

$$\forall z \in B, g(x) = g_0(x) \text{ and } \forall z \in C \setminus B, g(x) = g_1(x).$$

We can now define the Natarajan dimension.

Definition 3 (Natarajan dimension). *For any function class $\mathscr{G} \subset \{\mathscr{Z} \to [S]\}$ the* Natarajan dimension *of $\mathscr{G}$ is the maximal size of a shattered set $C \subset \mathscr{Z}$.*

4.6.2 Proofs of Propositions 8 and 9

It suffices to prove Proposition 9, since Proposition 8 follows from it by taking $z^n = w^n$ and taking $|\mathscr{G}| = 1$ with the function $g \in \mathscr{G}$ being simply $g(z) = z$.

Recall that for a probability assignment $q(y_i|y^{i-1};z^i)$, we have the probability induced portfolio $a = \phi(q)$ defined as

$$a(j|\mathbf{x}^{t-1};z^t) := \frac{\sum_{y^{t-1}} q(y^{t-1}j\|z^t)\mathbf{x}(y^{t-1})}{\sum_{y^{t-1}} q(y^{t-1}\|z^{t-1})\mathbf{x}(y^{t-1})},$$

where recall for $t \in [n], q(y^t\|z^t) = \prod_{i=1}^t q(y_i|y^{i-1};z^i)$. We then have

$$\sum_{y_t \in [m]} a(y_t|\mathbf{x}^{t-1};z^t)\mathbf{x}_t(y_t) = \frac{\sum_{y^{t-1}} q(y^t\|z^t)\mathbf{x}(y^t)}{\sum_{y^{t-1}} q(y^{t-1}\|z^{t-1})\mathbf{x}(y^{t-1})},$$

and consequently using a telescoping argument,

$$S_n(\phi(q), \mathbf{x}^n; z^n) = \sum_{y^n \in [m]^n} q(y^n \| z^n) \mathbf{x}(y^n).$$

From this we can see that

$$\begin{aligned}
\mathsf{Reg}_n^{\mathsf{port}}(\phi(q); \phi(\mathscr{P}), \mathscr{G}; \mathbf{x}^n, z^n) &= \sup_{g \in \mathscr{G}} \sup_{p \in \mathscr{P}} \log \frac{S_n(\phi(p), \mathbf{x}^n; g(z^n))}{S_n(\phi(q), \mathbf{x}^n; z^n)} \\
&= \sup_{g \in \mathscr{G}} \sup_{p \in \mathscr{P}} \log \frac{\sum_{y^n \in [m]^n} p(y^n \| g(z^n)) \mathbf{x}(y^n)}{\sum_{y^n \in [m]^n} q(y^n \| z^n) \mathbf{x}(y^n)} \\
&\le \sup_{g \in \mathscr{G}} \sup_{p \in \mathscr{P}} \max_{y^n \in [m]^n} \frac{p(y^n \| g(z^n))}{q(y^n \| z^n)} \qquad (4.23) \\
&= \mathsf{Reg}_n^{\mathsf{prob}}(q; \mathscr{P}, \mathscr{G}; z^n),
\end{aligned}$$

where (4.23) follows from Lemma 10. □

4.6.3 Proof of Universality in Expectation in Example 2

Recall that by Theorem 3, it suffices to show that

$$\mathsf{E}[\rho_{\mathscr{G} \times \mathscr{G}}(Z^n)] = o\Big(\frac{n}{\log^2 n}\Big) \qquad (4.11)$$

to establish that the induced portfolio $\phi(q^*_{\mathscr{G}})$ is universal in expectation. Indeed, for a β-mixing process Z^n with β-mixing coefficient β_k and β-mixing exponent $r > 0$, i.e., $\beta_k = O(k^{-r})$ as $k \to \infty$, we can prove a stronger statement:

$$\mathsf{E}[\rho_{\mathscr{G} \times \mathscr{G}}(Z^n)] = O(n^{(3+r)/(3+2r)}). \qquad (4.24)$$

The argument below to show (4.24) is based on the techniques of [116] and [117].

Pick $k \ge 1$ which divides n for simplicity; the divisibility can be easily lifted by elongating the game from n steps to the next number divisible by k. We will choose k as a function of n at the end of proof. We define the nonoverlapping k subsequences $Z^{(1)}, \dots, Z^{(k)}$ of length n/k as

$$\begin{aligned}
Z^{(1)}{}_1^{n/k} &= Z_1, Z_{k+1}, Z_{2k+1} \dots, Z_{(n/k-1)+1}, \\
Z^{(2)}{}_1^{n/k} &= Z_2, Z_{k+2}, Z_{2k+2} \dots, Z_{(n/k-1)k+2}, \\
&\vdots \\
Z^{(k)}{}_1^{n/k} &= Z_k, Z_{2k}, Z_{3k} \dots, Z_{(n/k)k}.
\end{aligned}$$

We will invoke the classical result on β-mixing processes that states that

$$d_{\mathrm{TV}}\left(P_{Z^{(j)}{}_1^{n/k}}, \prod_{i=1}^{n/k} P_{Z_i^{(j)}}\right) \leq \left(\frac{n}{k} - 1\right) \beta_k \tag{4.25}$$

for each $j \in [k]$, where $d_{\mathrm{TV}}(\cdot, \cdot)$ denotes the total variation distance; see, for example, [117, Lemma 1] and the references therein.

Now, we consider

$$\begin{aligned}
\mathsf{E}[\rho_{\mathscr{H}}(Z^n)] &= \mathsf{E}\Big[\sup_{h \in \mathscr{H}} \Big|\sum_{i=1}^{n} (h(Z_i) - E[h(Z_i)])\Big|\Big] \\
&\leq \mathsf{E}\Big[\sum_{j=1}^{k} \sup_{h \in \mathscr{H}} \Big|\sum_{i=1}^{n/k} (h(Z_i^{(j)}) - \mathsf{E}[h(Z_i^{(j)})])\Big|\Big] \\
&= \sum_{j=1}^{k} \mathsf{E}\Big[\sup_{h \in \mathscr{H}} \Big|\sum_{i=1}^{n/k} (h(Z_i^{(j)}) - \mathsf{E}[h(Z_i^{(j)})])\Big|\Big]. \qquad (4.26)
\end{aligned}$$

Let $Z'_1, \dots, Z'_{n/k}$ be an i.i.d. process with the same marginal distribution of the stationary process Z^n, i.e., $P_{Z'_1} = P_{Z_1}$. Continuing from the summand in (4.26), we then have

$$\begin{aligned}
&\mathsf{E}\Big[\sup_{h \in \mathscr{H}} \Big|\sum_{i=1}^{n/k} (h(Z_i^{(1)}) - \mathsf{E}[h(Z_i^{(1)})])\Big|\Big] \qquad (4.27) \\
&= \mathsf{E}\Big[\sup_{h \in \mathscr{H}} \Big|\sum_{i=1}^{n/k} (h(Z_i^{(1)}) - h(Z'_i) + h(Z'_i) - \mathsf{E}[h(Z_i^{(1)})])\Big|\Big]
\end{aligned}$$

$$\begin{aligned}
&= \mathsf{E}\left[\sup_{h\in\mathscr{H}}\Big|\sum_{i=1}^{n/k}(h(\mathrm{Z}_i^{(1)})-h(\mathrm{Z}_i')+h(\mathrm{Z}_i')-\mathsf{E}[h(\mathrm{Z}_i')])\Big|\right] && (4.28)\\
&\leq \mathsf{E}\left[\sup_{h\in\mathscr{H}}\Big|\sum_{i=1}^{n/k}(h(\mathrm{Z}_i^{(1)})-h(\mathrm{Z}_i'))\Big|\right]+\mathsf{E}\left[\sup_{h\in\mathscr{H}}\Big|\sum_{i=1}^{n/k}(h(\mathrm{Z}_i')-E[h(\mathrm{Z}_i')])\Big|\right] && (4.29)\\
&\leq \mathsf{E}\left[\sup_{h\in\mathscr{H}}\Big|\sum_{i=1}^{n/k}(h(\mathrm{Z}_i^{(1)})-h(\mathrm{Z}_i'))\Big|\right]+C\sqrt{\frac{dn}{k}} && (4.30)\\
&\leq \frac{n}{k}\sup_{h\in\mathscr{H}}\mathsf{E}\left|\frac{k}{n}\sum_{i=1}^{n/k}h(\mathrm{Z}_i^{(1)})-\frac{k}{n}\sum_{i=1}^{n/k}h(\mathrm{Z}_i')\right|+C\sqrt{\frac{dn}{k}}\\
&\leq \frac{n}{k}d_{\mathrm{TV}}\left(P_{\mathrm{Z}(j)_1^{n/k}},\prod_{i=1}^{n/k}P_{\mathrm{Z}_i^{(j)}}\right)+C\sqrt{\frac{dn}{k}} && (4.31)\\
&\leq \frac{n^2\beta_k}{k^2}+C\sqrt{\frac{dn}{k}}. && (4.32)
\end{aligned}$$

Here, (4.28) follows since the marginal distribution $Z_i' \overset{(d)}{=} Z_i^{(1)}$, (4.30) follows since the distribution Z'^n is i.i.d. and from [3, Theorem 8.3.23], (4.31) follows from the following variational form of the total variation distance $d_{\mathrm{TV}}(P,P')$ between two measures P and P' defined over the same measure space, i.e.,

$$d_{\mathrm{TV}}(P,P') = \sup_{f:|f|\leq 1}|\mathsf{E}_{X\sim P}[f(X)]-\mathsf{E}_{X\sim P'}[f(X)]|,$$

and lastly (4.32) follows from (4.25). Substituting (4.32) into (4.26) yields that

$$\mathsf{E}[\rho_{\mathscr{H}}(Z^n)] \leq \frac{n^2\beta_k}{k}+C\sqrt{dnk} \leq \frac{C'n^2k^{-r}}{k}+C\sqrt{dnk}$$

for k sufficiently large with some $C'>0$, where we use the definition of the β-mixing exponent r in the second inequality. Finally, choosing $k=O(n^{\frac{3}{3+2r}})$ yields the claimed rate $\mathsf{E}[\rho_{\mathscr{H}}(Z^n)] = O(n^{\frac{3+r}{3+2r}})$. □

4.6.4 Proof of Lemma 13

Note that for any Z^n and $\tilde{Z}^n$, we can write

$$\begin{aligned}
\mathrm{d_H}(g(Z^n),\widetilde{g}(Z^n)) &= \mathrm{d_H}(g(Z^n),\widetilde{g}(Z^n)) - \mathrm{d_H}(g(\tilde{Z}^n),\widetilde{g}(\tilde{Z}^n)) \qquad (4.33)\\
&\le \sup_{g_1,g_2} \left|\mathrm{d_H}(g_1(Z^n),g_2(Z^n)) - \mathrm{d_H}(g_1(\tilde{Z}^n),g_2(\tilde{Z}^n))\right|\\
&\le \sup_{g_1,g_2} \left|\mathrm{d_H}(g_1(Z^n),g_2(Z^n)) - n\mathsf{P}(g_1(Z_1)\neq g_2(Z_2))\right|\\
&\quad + \sup_{g_1,g_2} \left|\mathrm{d_H}(g_1(\tilde{Z}^n),g_2(\tilde{Z}^n)) - n\mathsf{P}(g_1(\tilde{Z}_1)\neq g_2(\tilde{Z}_1))\right|\\
&= \rho_{\mathscr{G}\times\mathscr{G}}(z^n) + \rho_{\mathscr{G}\times\mathscr{G}}(\tilde{z}^n) \qquad (4.34)
\end{aligned}$$

where (4.33) follows since $\mathrm{d_H}(g(\tilde{Z}^n),\widetilde{g}(\tilde{Z}^n)) = 0$ by design and (4.34) follows since by stationarity of $Z^n \overset{(d)}{=} \tilde{Z}^n$, we have $n\mathsf{P}(g_1(Z_1)\neq g_2(Z_1)) = n\mathsf{P}(g_1(\tilde{Z}_1)\neq g_2(\tilde{Z}_1)) = \sum_{i=1}^n \mathsf{P}(g_1(\tilde{Z}_i)\neq g_2(\tilde{Z}_i)) = \sum_{i=1}^n \mathsf{E}[\mathbb{1}g_1(\tilde{Z}_i)\neq g_2(\tilde{Z}_i)]$. Finally, substituting (4.34) into (4.19) yields the lemma.

□

4.6.5 A Detailed Discussion on Example 3

For the side information $z_t = \mathbf{x}_{t-k}^{t-1}$ in Example 3, if the market $(\mathbf{X}_t)$ itself is k-th order Markov, then we can establish the following guarantee.

Lemma 14. *Let $\mathbf{X}^n$ be a stationary k-th order Markov process and let $Z_t = \mathbf{X}_{t-k}^{t-1} \in (\mathbb{R}_+^m)^k$. Suppose that (1) the density of $Z_0 = \mathbf{X}_{-k}^{-1}$ exists and is bounded and supported over a bounded, convex set $E \subset (\mathbb{R}_+^m)^k$ with nonempty interior and (2) there exist $b > 0$ and $\varepsilon > 0$ such that the time-invariant conditional density satisfies*

$$p_{\mathbf{X}_{t-k+1}^{t}|\mathbf{X}_{t-k}^{t-1}}(z'|z) \ge b1_{B(z,\varepsilon)}(z')$$

for any $z \in (\mathbb{R}_+^m)^k$, where $B(z,\varepsilon)$ denotes the open ball of radius ε centered at $z \in (\mathbb{R}_+^m)^k$ with respect to Euclidean distance. Then, we have $\mathsf{E}[\rho_{\mathscr{H}}(Z^n)] = \widetilde{O}(\sqrt{n})$.

Proof. This is a direct consequence of [118, Proposition 11], which establishes an upper bound on $\mathsf{E}[\rho_{\mathscr{H}}(Z'^n)]$ for a *Metropolis–Hastings* (MH) walk Z'^n. First, note that Z^n forms a Markov chain due to the k-th order Markovity of $\mathbf{X}^n$. To apply the proposition over the Markov chain Z^n, we set the proposal distribution q in the MH algorithm to be the actual transition kernel of the Markov chain Z^n, so that the MH walk becomes the process Z^n of our interest. Then, under the assumptions above, we can apply the result of [118] and conclude that $\mathsf{E}[\rho_{\mathscr{H}}(Z^n)] = \tilde{O}(\sqrt{n})$ for a VC-class $\mathscr{H}$. □

4.6.6 Simulation Based on Monte Carlo Approximation

The (maybe the only) downside of the universal portfolio algorithms is their computational complexity. It is not hard to see that the *exact* computation of Cover's universal portfolio requires, on the T-th day of investment over m stocks, $O(T^m)$ time complexity, and the computation quickly become infeasible for a long investment period; see [19] for a detailed argument. An efficient implementation of universal portfolios is a decades-old open problem and still remains as an active area of research [119, 120]. Hence, in this paper, we consider a Monte Carlo simulation of the universal portfolio algorithms based on the cumulative wealth expression of a probability induced portfolio (4.3). While it is a very crude approximation for large m, S, or $\mathrm{Ndim}(\mathscr{G})$, this at least provides a way to demonstrate the performance of the ideas.

First, note that from (4.3), the cumulative wealth achieved by Cover's universal portfolio $\phi(q_{\mathrm{L}})$ can be written as

$$S_n(\phi(q_{\mathrm{L}}),\mathbf{x}^n) = \sum_{y^n\in[m]^n} q_{\mathrm{L}}(y^n)\mathbf{x}(y^n) = \int_{\mathscr{B}} S_n(\boldsymbol{\theta},\mathbf{x}^n)\mu(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta},$$

since the Laplace probability assignment $q_{\mathrm{L}}(y^n) = \int_{\mathscr{B}} \mu(\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(y^n)\,\mathrm{d}\boldsymbol{\theta}$ is a mixture with respect to a uniform density $\mu(\boldsymbol{\theta})$ over the simplex $\mathscr{B}$. Hence, if we draw N CRPs $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N$ from μ and *buy-and-hold* uniformly over the CRPs, we will attain approximately similar wealth and the approximation will get better as N becomes larger. Note, however, that this naive approximation requires $N = \Omega(\frac{1}{\varepsilon^m})$ to achieve an approximation error ε and thus may not be feasible when the number of stocks m is large.

A similar crude approximation can be performed for the universal portfolio $\phi(q_{\mathrm{L};S})$ with discrete side information w^n, since

$$
\begin{aligned}
S_n(\phi(q_{\mathrm{L};S}), \mathbf{x}^n; w^n) &= \sum_{y^n \in [m]^n} q_{\mathrm{L};S}(y^n \| w^n)\mathbf{x}(y^n) \\
&= \prod_{s=1}^{S} S_{|\mathbf{x}^n(s;w^n)|}(\phi(q_{\mathrm{L}}), \mathbf{x}^n(s;w^n)),
\end{aligned}
$$

where $\mathbf{x}^n(s;w^n) = (\mathbf{x}_i \colon w_i = s, i \in [n])$, since $q_{\mathrm{L};S}(y^n \| w^n) = \prod_{s=1}^{S} q_{\mathrm{L}}(y^n(s;w^n))$. That is, we can crudely approximate the performance of $\phi(q_{\mathrm{L};S})$ by simply drawing many state-wise CRPs $\boldsymbol{\theta}_{1:S}$ according to $\mu(\boldsymbol{\theta}_{1:S}) \coloneqq \mu(\boldsymbol{\theta}_1) \cdots \mu(\boldsymbol{\theta}_S)$ and running the buy-and-hold strategy.

We can now consider an approximation of the proposed strategy $\phi(q_{\mathscr{G}}^*)$. By the epoch-wise construction of $q_{\mathscr{G}}^*$ as explicitly shown in (4.9), the cumulative wealth can be also factorized as

$$
\begin{aligned}
S_n(\phi(q_{\mathscr{G}}^*), \mathbf{x}^n; z^n) &= \prod_{j=1}^{J} \sum_{y_{2^{j-1}+1}^{2^j} \in [m]^{2^{j-1}}} q_{\mathscr{G};z^{2^{j-1}}}(y_{2^{j-1}+1}^{2^j} \| z_{2^{j-1}+1}^{2^j})\mathbf{x}_{2^{j-1}+1}^{2^j}(y_{2^{j-1}+1}^{2^j}) \\
&= \prod_{j=1}^{J} S_{2^{j-1}}(\phi(q_{\mathscr{G};z^{2^{j-1}}}), \mathbf{x}_{2^{j-1}+1}^{2^j}; z_{2^{j-1}+1}^{2^j}).
\end{aligned}
$$

where we assume $n = 2^J$ for simplicity. Here, for each $j \in [J]$, if $\{\widetilde{g}_1, \ldots, \widetilde{g}_{\ell_j}\}$ is a minimal empirical covering of $\mathscr{G}$ with respect to $z^{2^{j-1}}$, we can write

$$
S_{2^{j-1}}(q_{\mathscr{G};z^{2^{j-1}}}, \mathbf{x}_{2^{j-1}+1}^{2^j}; z_{2^{j-1}+1}^{2^j}) = \frac{1}{\ell_j} \sum_{k=1}^{\ell_j} S_{2^{j-1}}(\phi(q_{\mathrm{L};S}), \mathbf{x}_{2^{j-1}+1}^{2^j}; \widetilde{g}_k(z_{2^{j-1}+1}^{2^j})).
$$

For each state function $\widetilde{g}_k$, the summand is the cumulative wealth of the UP with the side information $\widetilde{g}_k(z_{2^{j-1}+1}^{2^j})$ and thus can be approximated by the same argument from the previous paragraph.

This leads to the following Monte Carlo simulation of the proposed algorithm. Let N be the number of Monte Carlo samples used in the approximation.

For each epoch $j = 1, 2, \ldots$:

1. Find an empirical covering $\{\widetilde{g}_1, \ldots, \widetilde{g}_{\ell_j}\} \subseteq \mathscr{G}$ with respect to $z^{2^{j-1}}$.
2. For each $k \in [\ell_j]$, draw N state-wise CRPs $(\boldsymbol{\theta}_{1:S,i}^{(k)})_{i=1}^{N}$ from $\mu(\boldsymbol{\theta}_{1:S})$ at random.
3. During the j-th investment epoch, i.e., $t \in (2^{j-1}, 2^j]$, run the buy-and-hold strategy uniformly over all sampled CRPs $(\boldsymbol{\theta}_{1:S,i}^{(k)})_{i=1}^{N}$ for each $k \in [\ell_j]$.
4. At the end of the epoch, sell all stocks.

We stress that this only simulates the cumulative wealth of the universal portfolio algorithm by directly estimating the cumulative wealth expression, rather than approximating actions of the algorithm for each round. Note that a more sophisticated Monte Carlo Markov Chain based approximation for Cover's universal portfolio was proposed and analyzed by [121]. It is left as a future direction to extend their method for our algorithm with continuous side information.

In the following, we study a simple example for concreteness, which admits an easy construction of minimal empirical coverings. Note that, for a richer class of state functions, finding a minimal empirical covering may be another computational bottleneck.

Example 4. *As a simple case of the canonical side information considered in Example 3, we choose the price relative of the stock 1 on the previous day as the continuous side information, i.e.,* $z_t = \mathbf{x}_{t-1,1}$*, and a class of 1D threshold functions* $\mathscr{G} = \{x \mapsto g_a(x) = 1\{x \geq a\} \colon a > 0\}$ *of* $\mathrm{Ndim}(\mathscr{G}) = 1$. *Note that we consider a binary state space (*$S = 2$*). In this case, it is easy to show that* $\{g_{x_{0,1}}, \ldots, g_{x_{t-1,1}}\}$ *is a minimal empirical covering given* $z^t = (x_{i,1})_{i=0}^{t-1}$.

In general, we can consider $z_t = \mathbf{x}_{t-1}$ *with a class of product of 1D threshold functions* $\mathscr{G} = \{x \mapsto g_{\mathbf{a}}(\mathbf{x}) = (1\{x_1 \geq a_1\}, \ldots, 1\{x_m \geq a_m\}) \colon \mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{R}_{++}^m\}$ *of* $\mathrm{Ndim}(\mathscr{G}) \leq m \log m$ *[99, Lemma 29.6] and* $S = 2^m$*. Given* $z^t = \mathbf{x}^{t-1}$*,* $\{g_{\mathbf{x}_0}, \ldots, g_{\mathbf{x}_{t-1}}\}$ *is a minimal empirical* *covering.*

A Toy Example.

We briefly demonstrate how the proposed portfolio performs on two real stocks. We collected the 6-year period from Jan-01-2012 to Dec-31-2017 (total 1508 trading days) of two stocks Ford (F) and Macy's (M). Over the period, Ford went up by a factor of 1.11, while Macy's went down by a factor of 0.77. The best CRP in hindsight, which turns out to be the buy-and-hold of Ford, achieves a growth factor of 1.11. The uniform CRP achieves a growth factor of 0.99. While the universal portfolio without side information achieves a growth factor of only, the proposed algorithm with the yesterday's prices and the class of thresholding functions achieves a growth factor of 1.15.

We note that there can exist more sophisticated, carefully chosen side information and state-function classes that may exhibit better performance in practice than the simple example above. We leave the problem of constructing good continuous side information and extensive experiments as future work.

4.7 Acknowledgement

This chapter includes the material in Alankrita Bhatt, Jongha Ryu, and Young-Han Kim "On universal portfolios with continuous side information", arXiv:2207.12382 to be submitted. The dissertation author was a primary investigator and author of this paper.

Appendix A

Omitted Proofs From Chapter 1

A.1 Proofs of Universality of q_{L} and q_{KT}

As mentioned in Chapter 1, both q_{L} and q_{KT} are point-wise (and therefore mean) universal for the class of binary i.i.d. processes, i.e. they satisfy

$$\max_{\theta, y^n} \log \frac{p_\theta(y^n)}{q(y^n)} = o(n)$$

where $p_\theta(y^n) = \theta^{\sum_{i=1}^n - \sum_{i=1}^n y_i}$. We use $k = \sum_{i=1}^n y_i$ for ease of notation.

Proof for q_{L}: In this case, we have

$$\log \frac{p_\theta(y^n)}{q_{\mathsf{L}}(y^n)} = \log(n+1) \binom{n}{k} t^k (1-t)^{n-k} \leq \log(n+1)$$

where the inequality follows since $\binom{n}{k} t^k (1-t)^{n-k} \leq 1$.

Proof for q_{KT}: In this case, we use the Stirling inequality to simplify the binomial terms. Firstly, recall that

$$\frac{\binom{n}{\sum_{i=1}^n y_i} \binom{2n}{n}}{4^n \binom{2n}{2\sum_{i=1}^n y_i}} \sim \frac{1}{\sqrt{2n}} 2^{-nh(k/n)}$$

where $h(\cdot)$ denotes, as usual, the binary entropy function. Moreover, we have

$$\max_\theta p_\theta(y^n) = 2^{-nh(k/n)}$$

achieved at $\theta = k/n$. Therefore, we see that $\max_{\theta, y^n} \log \frac{p_\theta(y^n)}{q_{\mathrm{L}}(y^n)} = \frac{1}{2}\log n + o(\log n)$.

A.2 Universality in Sequential Prediction

In this section, we consider the problem of sequential prediction, which encompasses the weather prediction problem motivated in Chapter 1.

Assuming that the (discrete) data Y^n is generated by a distribution p_θ, with $\theta \in \Theta$ being unknown, we look at the difference in losses (with loss function ℓ) suffered by the Bayes predictor

$$a_t^*(Y^{t-1}) = \arg\inf \mathsf{E}_{p_\theta}[\ell(a(Y^{t-1}), Y_t) | Y^{t-1}]. \tag{A.1}$$

and the *universal* predictor that pretends that the true data distribution is q (instead of the unknown p_θ) and plays

$$\widehat{a}_t(Y^{t-1}) = \arg\inf \mathsf{E}_q[\ell(a(Y^{t-1}), Y_t) | Y^{t-1}] \tag{A.2}$$

where q is assumed to be a distribution universal for the class Θ. More precisely, we look at the expected regret (with Y^{t-1} in the estimators $\widehat{a}_t, a_t^*$ suppressed for brevity)

$$\mathrm{Reg}_n = \mathsf{E}\left[\sum_{t=1}^{n} \ell(\widehat{a}_t, Y_t) - \sum_{t=1}^{n} \ell(a_t^*, Y_t)\right]. \tag{A.3}$$

We will also assume that the loss function $\ell \leq 1$.

We now show the following, as seen in [77].

Theorem 6. *For $\widehat{a}_t, a_t^*$ as defined in* (A.2) *and* (A.1) *respectively, if q is such that $\frac{1}{n}D(p_\theta \| q) \leq C_n$ for all $\theta \in \Theta$, then* $\mathrm{Reg}_n \leq \sqrt{2\ln 2 C_n}$.

Clearly, Theorem 6 established that if q is universal (i.e. $C_n = o(1)$), then $\mathrm{Reg}_n = o(1)$ as well. We now prove this statement, following the arguments of [77].

Proof. We have

$$
\begin{aligned}
&\frac{1}{n}\mathsf{E}\left[\sum_{t=1}^{n}\ell(\widehat{a}_t,Y_t)-\sum_{t=1}^{n}\ell(a_t^*,Y_t)\right]\\
&\quad=\frac{1}{n}\sum_{t=1}^{n}\mathsf{E}\left[\mathsf{E}\left[\ell(\widehat{a}_t,Y_t)-\ell(a_t^*,Y_t)|Y^{t-1}\right]\right]\\
&\quad=\frac{1}{n}\sum_{t=1}^{n}\sum_{y^{t-1}}p_\theta(y^{t-1})\sum_{y_t}(\ell(\widehat{a}_t,y_t)-\ell(a_t^*,y_t))p_\theta(y_t|y^{t-1})\\
&\quad\leq\frac{1}{n}\sum_{t=1}^{n}\sum_{y^{t-1}}p_\theta(y^{t-1})\sum_{y_t}(\ell(\widehat{a}_t,y_t)-\ell(a_t^*,y_t))\Big(q(y_t|y^{t-1})\\
&\qquad\qquad\qquad\qquad+|p_\theta(y_t|y^{t-1})-q(y_t|y^{t-1})|\Big) && \text{(A.4)}\\
&\quad\leq\frac{1}{n}\sum_{t=1}^{n}\sum_{y^{t-1}}p_\theta(y^{t-1})\sum_{y_t}(\ell(\widehat{a}_t,y_t)-\ell(a_t^*,y_t))\left|p_\theta(y_t|y^{t-1})-q(y_t|y^{t-1})\right| && \text{(A.5)}\\
&\quad\leq\frac{1}{n}\sum_{t=1}^{n}\sum_{y^{t-1}}p_\theta(y^{t-1})\sum_{y_t}\left|p_\theta(y_t|y^{t-1})-q(y_t|y^{t-1})\right| && \text{(A.6)}\\
&\quad\leq\sqrt{\frac{1}{n}\sum_{t=1}^{n}\sum_{y^{t-1}}p_\theta(y^{t-1})\left[\sum_{y_t}|p_\theta(y_t|y^{t-1})-q(y_t|y^{t-1})|\right]^2} && \text{(A.7)}\\
&\quad\leq\sqrt{\frac{2\ln 2}{n}\sum_{t=1}^{n}\sum_{y^{t-1}}p_\theta(y^{t-1})\sum_{y_t}p_\theta(y_t|y^{t-1})\log\frac{p_\theta(y_t|y^{t-1})}{q(y_t|y^{t-1})}} && \text{(A.8)}\\
&\quad=\sqrt{\frac{2\ln 2}{n}\sum_{t=1}^{n}D(p_\theta(y_t|y^{t-1})\|q(y_t|y^{t-1}))}\\
&\quad=\sqrt{\frac{2\ln 2}{n}D(p_\theta(y^n)\|q(y^n))} && \text{(A.9)}
\end{aligned}
$$

where (A.4) follows since $a\leq b+|a-b|$; (A.5) follows since by definition of $\widehat{a}_t$ in (A.2), we have

$$\mathsf{E}_{q(y_t|y^{t-1})}[\ell(\widehat{a}_t,Y_t])]\leq\mathsf{E}_{q(y_t|y^{t-1})}[\ell(a_t^*,Y_t])];$$

(A.6) follows since $\ell\leq 1$; (A.7) follows by the Jensen inequality (since $\mathsf{E}|Z|\leq\sqrt{E[Z^2]}$); (A.8) follows by the Pinsker inequality; and finally (A.9) follows from the chain rule of KL divergence.

□

A.2.1 The Weather Prediction Problem

We can see by substituting $\ell(a,y) = \mathbb{1}\{a \neq y\}$, that the setting of sequential prediction above encompasses the weather prediction problem from Chapter 1. We can clearly see that the Bayes responses for a distribution p_θ (or a universal distribution q) turn out to be $a_t^* = y_t^* = \mathbb{1}\{p_\theta(y_t) \geq 1/2\}$ and $\widehat{a}_t = \hat{y}_t = \mathbb{1}\{q(y_t|y^{t-1}) \geq 1/2\}$ respectively. By substituting $q = q_{\mathsf{KT}}$ or q_{L}, and by virtue of their universality for binary i.i.d. processes, we see that $C_n = O(\log n)$ and therefore the weatherperson equipped with a universal predictor does not make too many more mistakes than she would have made with prior knowledge of θ.

Remark 8 (Why not use the MLE?)**.** *A natural alternative choice for $\hat{y}_t$ is to make a decision based on the maximum likelihood estimate (MLE), i.e. (in the weather prediction problem) choose $\hat{y}_t = \mathbb{1}\left\{\frac{\sum_{i=1}^{t-1} y_t}{t-1} \geq 1/2\right\}$. Indeed, for the particular case of weather prediction with indicator loss, this strategy does work. However, for several other loss functions such as the log-loss, used in applications like compression and gambling, this would be a catastrophic strategy—if one assigns probability $q(y_t|y^{t-1}) = \frac{\sum_{i=1}^{t-1} y_t}{t-1}$ and $y^{t-1} = 0^{t-1}$ (this might happen, for instance, if θ is quite small) and $y_t = 1$, then the loss suffered* $\log \frac{p_\theta(y_t)}{q(y_t|y^{t-1})} = \infty$.

Bibliography

[1] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[2] F. Orabona, "A modern introduction to online learning," *arXiv preprint arXiv:1912.13213*, 2019.

[3] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press, 2018.

[4] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press, 2019.

[5] I. Diakonikolas and D. M. Kane, "Recent advances in algorithmic high-dimensional robust statistics," *arXiv preprint arXiv:1911.05911*, 2019.

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

[7] T. Begley, *Bayesian Billiards, available online at* *https://tcbegley.com/blog/bayesian-billiards*. 2020.

[8] D. Spiegelhalter, *The Art of Statistics: How to Learn from Data*. Penguin, 2019.

[9] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[10] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[11] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[12] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[13] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[14] A. D. Wyner and J. Ziv, “The sliding-window lempel-ziv algorithm is asymptotically optimal,” *Proceedings of the IEEE*, vol. 82, no. 6, pp. 872–877, 1994.

[15] S. A. Savari, “Redundancy of the lempel-ziv incremental parsing rule,” *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 9–21, 1997.

[16] M. Feder, N. Merhav, and M. Gutman, “Universal prediction of individual sequences,” *IEEE Trans. Inf. Theory*, vol. 38, no. 4, pp. 1258–1270, 1992.

[17] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2006.

[18] T. M. Cover, “Universal portfolios,” *Math. Financ.*, vol. 1, no. 1, pp. 1–29, 1991.

[19] T. M. Cover and E. Ordentlich, “Universal portfolios with side information,” *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 348–363, 1996.

[20] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, “Universal estimation of directed information,” *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6220–6242, 2013.

[21] F. Orabona and D. Pál, “Optimal non-asymptotic lower bound on the minimax regret of learning with expert advice,” *arXiv preprint arXiv:1511.02176*, 2015.

[22] J. J. Ryu, A. Bhatt, and Y.-H. Kim, “Parameter-free online linear optimization with side information via universal coin betting,” *arXiv preprint arXiv:2202.02406*, 2022.

[23] F. Orabona and K.-S. Jun, “Tight concentrations and confidence sequences from the regret of universal portfolio,” *arXiv preprint arXiv:2110.14099*, 2021.

[24] J. Ziv and A. Lempel, “Compression of individual sequences via variable-rate coding,” *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[25] Q. Xie and A. R. Barron, “Minimax redundancy for the class of memoryless sources,” *IEEE Transactions on Information Theory*, vol. 43, no. 2, pp. 646–657, 1997.

[26] Q. Xie and A. R. Barron, “Minimax redundancy for the class of memoryless sources,” *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 646–657, 1997.

[27] Q. Xie and A. R. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction,” *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[28] A. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” *IEEE transactions on information theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[29] P. D. Grünwald, *The minimum description length principle*. MIT press, 2007.

[30] T. M. Cover, “Behavior of sequential predictors of binary sequences,” in *Proc. 4th Prague Conf. Inf. Theory, Stat. Decis. Funct. Random Process.*, pp. 263–272, Prague: Publishing House of the Czechoslovak Academy of Sciences, 1967.

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

[32] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger, "Universal discrete denoising: Known channel," *IEEE Transactions on Information Theory*, vol. 51, no. 1, pp. 5–28, 2005.

[33] J. Ryu and Y.-H. Kim, "Conditional distribution learning with neural networks and its application to universal image denoising," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3214–3218, IEEE, 2018.

[34] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[35] M. Hernaez, D. Pavlichin, T. Weissman, and I. Ochoa, "Genomic data compression," *Annual Review of Biomedical Data Science*, vol. 2, pp. 19–37, 2019.

[36] S. Kwon, G. Kim, B. Lee, J. Chun, S. Yoon, and Y.-H. Kim, "Nascup: Nucleic acid sequence classification by universal probability," *IEEE Access*, vol. 9, pp. 162779–162791, 2021.

[37] R. Rossi and R. Zhou, "GraphZIP: a clique-based sparse graph compression method," *Journal of Big Data*, vol. 5, no. 10, 2018.

[38] Y. Lim, U. Kang, and C. Faloutsos, "Slashburn: Graph compression and mining beyond caveman communities," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 3077–3089, 2014.

[39] P. Boldi and S. Vigna, "The webgraph framework i: Compression techniques," in *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, (New York, NY, USA), pp. 595–602, Association for Computing Machinery, 2004.

[40] T. C. Conway and A. J. Bromage, "Succinct data structures for assembling large genomes," *Bioinformatics*, vol. 27, pp. 479–486, 01 2011.

[41] M. Hayashida and T. Akutsu, "Comparing biological networks via graph compression," *BMC systems biology*, vol. 4 Suppl 2, no. Suppl 2, 2010.

[42] F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, and P. Raghavan, "On compressing social networks," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, (New York, NY, USA), pp. 219–228, Association for Computing Machinery, 2009.

[43] G. Navarro, "Compressing web graphs like texts," tech. rep., Dept. of Computer Science, University of Chile, 2007.

[44] K. Sadakane, "New text indexing functionalities of the compressed suffix arrays," *Journal of Algorithms*, vol. 48, no. 2, pp. 294 – 313, 2003.

[45] N. R. Brisaboa, S. Ladra, and G. Navarro, "K2-trees for compact web graph representation," in *Proceedings of the 16th International Symposium on String Processing and Information Retrieval*, SPIRE '09, (Berlin, Heidelberg), pp. 18–30, Springer-Verlag, 2009.

[46] A. Farzan and J. I. Munro, "Succinct encoding of arbitrary graphs," *Theoretical Computer Science*, vol. 513, pp. 38 – 52, 2013.

[47] G. Turán, "On the succinct representation of graphs," *Discrete Applied Mathematics*, vol. 8, no. 3, pp. 289 – 294, 1984.

[48] M. Naor, "Succinct representation of general unlabeled graphs," *Discrete Applied Mathematics*, vol. 28, no. 3, pp. 303 – 307, 1990.

[49] Y. Choi and W. Szpankowski, "Compression of graphical structures: Fundamental limits, algorithms, and experiments," *IEEE Trans. Inf. Theory*, vol. 58, pp. 620–638, Feb 2012.

[50] P. Delgosha and V. Anantharam, "Universal lossless compression of graphical data," in *Proc. IEEE Internat. Symp. Inf. Theory*, June 2017.

[51] P. Delgosha and V. Anantharam, "Universal lossless compression of graphical data," 2019.

[52] P. Delgosha and V. Anantharam, "A universal low complexity compression algorithm for sparse marked graphs," in *Proc. IEEE Internat. Symp. Inf. Theory*, June 2020.

[53] E. Abbe, "Graph compression: The effect of clusters," in *Proc. 54th Ann. Allerton Conf. Commun. Control Comput.*, pp. 1–8, 2016.

[54] A. Asadi, E. Abbe, and S. Verdú, "Compressing data on graphs with clusters," in *Proc. IEEE Internat. Symp. Inf. Theory*, pp. 1583–1587, August 2017.

[55] M. Besta and T. Hoefler, "Survey and taxonomy of lossless graph compression and space-efficient graph representations," 2018.

[56] M. Effros, K. Visweswariah, S. R. Kulkarni, and S. Verdú, "Universal lossless source coding with the burrows wheeler transform," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1061–1081, 2002.

[57] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[58] P. Delgosha and V. Anantharam, "Universal lossless compression of graphical data," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 6962–6976, 2020.

[59] C. Bordenave and P. Caputo, "Large deviations of empirical neighborhood distribution in sparse random graphs," *Probability Theory and Related Fields*, vol. 163, p. 149–222, Nov 2014.

[60] Y. Polyanskiy and Y. Wu, “Lecture notes on information theory,” 2014.

[61] A. Frieze and M. Karoński, *Introduction to Random Graphs*. Cambridge University Press, 2015.

[62] D. Marpe, H. Schwarz, and T. Wiegand, “Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, 2003.

[63] E. Abbe, A. S. Bandeira, and G. Hall, “Exact recovery in the stochastic block model,” *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 471–487, 2015.

[64] E. Abbe and C. Sandon, “Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery,” in *IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 670–688, 2015.

[65] E. Mossel, J. Neeman, and A. Sly, “Reconstruction and estimation in the planted partition model,” *Probability Theory and Related Fields*, vol. 162, no. 3-4, pp. 431–461, 2015.

[66] S. Lauritzen, A. Rinaldo, and K. Sadeghi, “Random networks, graphical models, and exchangeability,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, 01 2017.

[67] T. Courtade, *Properties of the binary entropy function, available online*. 2012.

[68] P. Billingsley, *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics, New York: John Wiley & Sons Inc., second ed., 1999. A Wiley-Interscience Publication.

[69] C. Bordenave, “Lecture notes on random graphs and probabilistic combinatorial optimization.” `https://www.math.univ-toulouse.fr/∼bordenave/coursRG.pdf`, 2016.

[70] A. Grover and J. Leskovec, “Node2vec: Scalable feature learning for networks,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, (New York, NY, USA), pp. 855–864, Association for Computing Machinery, 2016.

[71] L. Tang and H. Liu, “Relational learning via latent social dimensions,” in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’09, (New York, NY, USA), pp. 817–826, Association for Computing Machinery, 2009.

[72] S. Nandanwar and M. N. Murty, “Structural neighborhood based classification of nodes in a network,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, (New York, NY, USA), pp. 1085–1094, Association for Computing Machinery, 2016.

[73] J. Shun and G. E. Blelloch, “Ligra: A lightweight graph processing framework for shared memory,” in *Proceedings of the 18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP ’13, (New York, NY, USA), pp. 135–146, Association for Computing Machinery, 2013.

[74] J. Shun, L. Dhulipala, and G. E. Blelloch, “Smaller and faster: Parallel processing of compressed graphs with ligra+,” in *2015 Data Compression Conference*, pp. 403–412.

[75] A. Lempel and J. Ziv, “Compression of two-dimensional data,” *IEEE Trans. Inf. Theory*, vol. 32, no. 1, pp. 2–8, 1986.

[76] Y. Fogel and M. Feder, “On the problem of on-line learning with log-loss,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2995–2999, IEEE, 2017.

[77] N. Merhav and M. Feder, “Universal prediction,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[78] J. Rissanen, “A universal data compression system,” *IEEE Transactions on information theory*, vol. 29, no. 5, pp. 656–664, 1983.

[79] J. Rissanen, “A universal prior for integers and estimation by minimum description length,” *The Annals of statistics*, pp. 416–431, 1983.

[80] J. Rissanen, “Universal coding, information, prediction, and estimation,” *IEEE Transactions on Information theory*, vol. 30, no. 4, pp. 629–636, 1984.

[81] Q. Xie and A. R. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction,” *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[82] Y. Shkel, M. Raginsky, and S. Verdú, “Sequential prediction with coded side information under logarithmic loss,” in *Algorithmic Learning Theory*, pp. 753–769, 2018.

[83] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[84] A. Lazaric and R. Munos, “Learning with stochastic inputs and adversarial outputs,” *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1516–1537, 2012.

[85] A. Rakhlin, K. Sridharan, and A. Tewari, “Sequential complexities and uniform martingale laws of large numbers,” *Probability Theory and Related Fields*, vol. 161, no. 1-2, pp. 111–153, 2015.

[86] A. Rakhlin, K. Sridharan, and A. Tewari, “Online learning via sequential complexities.,” *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 155–186, 2015.

[87] A. Rakhlin and K. Sridharan, “Sequential probability assignment with binary alphabets and large classes of experts,” *arXiv preprint arXiv:1501.07340*, 2015.

[88] B. Bilodeau, D. Foster, and D. Roy, “Tight bounds on minimax regret under logarithmic loss via self-concordance,” in *International Conference on Machine Learning*, pp. 919–929, PMLR, 2020.

[89] R. Krichevsky and V. Trofimov, “The performance of universal encoding,” *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[90] Y. Yang and A. Barron, “Information-theoretic determination of minimax rates of convergence,” *Annals of Statistics*, pp. 1564–1599, 1999.

[91] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[92] N. Merhav and M. Feder, “A strong version of the redundancy-capacity theorem of universal coding,” *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 714–722, 1995.

[93] B. Bilodeau, D. J. Foster, and D. M. Roy, “Minimax rates for conditional density estimation via empirical entropy,” *arXiv preprint arXiv:2109.10461*, 2021.

[94] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

[95] S. Li, “Concise formulas for the area and volume of a hyperspherical cap,” *Asian Journal of Mathematics and Statistics*, vol. 4, no. 1, pp. 66–70, 2011.

[96] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

[97] A. Bhatt and Y.-H. Kim, “Sequential prediction under log-loss with side information,” in *Algo. Learn. Theory*, pp. 340–344, PMLR, 2021.

[98] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[99] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[100] B. Yu, “Rates of convergence for empirical processes of stationary mixing sequences,” *Ann. Probab.*, pp. 94–116, 1994.

[101] J. J. Rissanen, “Fisher information and stochastic complexity,” *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[102] Q. Xie and A. R. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction,” *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[103] A. Blum and A. Kalai, “Universal portfolios with and without transaction costs,” *Mach. Learn.*, vol. 35, no. 3, pp. 193–205, 1999.

[104] G. Uziel and R. El-Yaniv, "Long-and short-term forecasting for portfolio selection with transaction costs," in *Int. Conf. Artif. Int. Statist.*, pp. 100–110, PMLR, 2020.

[105] S. S. Kozat, A. C. Singer, and A. J. Bean, "Universal portfolios via context trees," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2093–2096, IEEE, 2008.

[106] A. Tavory and M. Feder, "Universal portfolio algorithms in realistic-outcome markets," in *Proc. IEEE Inf. Theory Workshop*, pp. 1–5, IEEE, 2010.

[107] A. Tavory and M. Feder, "Finite memory universal portfolios," in *Proc. IEEE Internat. Symp. Inf. Theory*, pp. 1408–1412, IEEE, 2008.

[108] J. E. Cross and A. R. Barron, "Efficient universal portfolios for past-dependent target classes," *Math. Financ.*, vol. 13, no. 2, pp. 245–276, 2003.

[109] L. Györfi, G. Lugosi, and F. Udina, "Nonparametric kernel-based sequential investment strategies," *Math. Financ.*, vol. 16, no. 2, pp. 337–357, 2006.

[110] A. Rakhlin and K. Sridharan, "Sequential probability assignment with binary alphabets and large classes of experts," *arXiv preprint arXiv:1501.07340*, 2015.

[111] B. Bilodeau, D. Foster, and D. Roy, "Tight bounds on minimax regret under logarithmic loss via self-concordance," in *Proc. Int. Conf. Mach. Learn.*, pp. 919–929, PMLR, 2020.

[112] Y. Fogel and M. Feder, "On the problem of on-line learning with log-loss," in *Proc. IEEE Internat. Symp. Inf. Theory*, pp. 2995–2999, IEEE, 2017.

[113] A. Rakhlin, K. Sridharan, and A. Tewari, "Sequential complexities and uniform martingale laws of large numbers," *Probability Theory and Related Fields*, vol. 161, no. 1-2, pp. 111–153, 2015.

[114] T. M. Adams and A. B. Nobel, "Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling," *Ann. Probab.*, pp. 1345–1367, 2010.

[115] O. Bousquet, *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

[116] R. L. Karandikar and M. Vidyasagar, "Rates of uniform convergence of empirical means with mixing processes," *Stat. Probab. Lett.*, vol. 58, no. 3, pp. 297–307, 2002.

[117] S. Hanneke and L. Yang, "Statistical learning under nonstationary mixing processes," in *Int. Conf. Artif. Int. Statist.*, pp. 1678–1686, PMLR, 2019.

[118] P. Bertail and F. Portier, "Rademacher complexity for Markov chains: Applications to kernel smoothing and Metropolis–Hastings," *Bernoulli*, vol. 25, no. 4B, pp. 3912–3938, 2019.

[119] H. Luo, C.-Y. Wei, and K. Zheng, "Efficient online portfolio with logarithmic regret," in *Adv. Neural Inf. Proc. Syst.*, pp. 8245–8255, 2018.

[120] T. van Erven, D. van der Hoeven, W. Kotlowski, and W. M. Koolen, "Open problem: Fast and optimal online portfolio selection," in *Conf. Learn. Theory*, pp. 3864–3869, PMLR, 2020.

[121] A. T. Kalai and S. Vempala, "Efficient algorithms for universal portfolios," *J. Mach. Learn. Res.*, pp. 423–440, 2002.