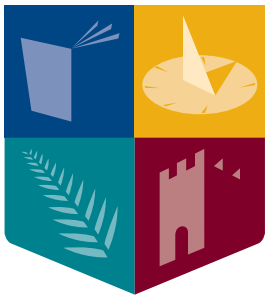

Maynooth University
National University of Ireland Maynooth

Thesis submitted in fulfilment of the requirements of the PhD degree, Maynooth University Hamilton Institute

Guesswork

Mark Mikael Christiansen

June 2015

Advisor: Professor Ken R. Duffy

Department Head: Professor Ken R. Duffy

Acknowledgements

Thank you to my supervisor Prof. Ken Duffy for guiding me through this process for the last 4 years and helping me through every step of the way. I would also like to thank Prof. Muriel Médard and Dr. Flavio du Pin Calmon for their immeasurable help.

A special thank you to friends too many to list who have given me so much joy and wonder over the years.

To my mum, my grandparents and the rest of my family for supporting and encouraging me in everything I have done in life.

Finally to Sinéad for being by my side these past few years.

Contents

1 Introduction

The security of systems is often predicated on a user or application selecting an object, a password or key, from a large list. If an inquisitor who wishes to identify the object in order to gain access to a system can only query each possibility, one at a time, then the number of guesses they must make in order to identify the selected object is likely to be large. If the object is selected uniformly at random using, for example, a cryptographically secure pseudo-random number generator, then the analysis of the distribution of the number of guesses that the inquisitor must make is trivial.

Since the earliest days of code-breaking, deviations from perfect uniformity have been exploited. For example, it has long since been known that human selected passwords are highly non-uniform, e.g. [36], and this forms the basis of dictionary attacks. In information theoretic security, uniformity of the string source is typically assumed on the basis that the source has been compressed. Recent work has cast some doubt on the appropriateness of that assumption by establishing that fewer queries are required to identify strings chosen from a typical set than one would expect by a naïve application of the asymptotic equipartition property. This arises by exploitation of the mild non-uniformity of the distribution of strings conditioned to be in the typical set [11].

If the string has not been selected perfectly uniformly, but with a distribution that is known to the inquisitor, then the quantification of security is relatively involved. Assume that a string, W, is selected stochastically from a finite list, $\mathbb{A} = \{0, \ldots, m-1\}$. An inquisitor who knows the selection probabilities, $P(W = w)$ for all $w \in \mathbb{A}$, is equipped with a method to test one string at a time and develops a strategy, $G : \mathbb{A} \mapsto \{1, \ldots, m\}$, that defines the order in which strings are guessed. As the string is stochastically selected, the number of queries, $G(W)$, that must be made before it is identified correctly is also a random variable, dubbed guesswork. Analysis of the distribution of guesswork serves as a natural a measure of computational security in brute force determination. Guesswork is the subject of this thesis, both in the original setting described above as well as in generalized scenarios.

Motivated by both lossless compression and brute force searching, in a brief paper in 1994 it was Massey [39] who first framed this question of guesswork. If W is uniformly distributed, all guesswork orders G result in the same stochastic properties of $G(W)$ and no more than elementary algebra is required to study the guesswork distribution. If W is not uniformly distributed, however, the choice of G does matter. Massey introduced the natural guesswork ordering, G, of inquiring about characters from most likely to

least likely, breaking ties arbitrarily, and analysed it. To put this in more mathematical terms, G is such such that $G(w) < G(w')$ implies that $P(W = w) \geq P(W = w')$ for all w and $w' \in \mathbb{A}$. Later in the thesis, we place a formal meaning on the optimality of this guesswork strategy in terms of stochastic dominance.

The Shannon entropy of a random variable, e.g. [14],

$$H(W) := -\sum_{w \in \mathbb{A}} P(W = w) \log P(W = w),$$

is a commonly appearing measure of variability. Massey asked if the average guesswork $E(G(W))$ could be characterized in terms of the Shannon entropy of W, $H(W)$, and demonstrated that this was not the case. He established the following lower bound on the expected guesswork

$$E(G(W)) = \sum_{i=1}^{m} iP(G(W) = i) \geq \frac{1}{4} 2^{H(W)} + 1,$$

but found is there is no similar upper bound. This discrepancy is most readily understood by the following example.

Consider the distribution of W with a single likely element and the rest of the probability distributed uniformly amongst the remaining letters,

$$P(W = i) = \begin{cases} \dfrac{m-2}{m} & \text{if } i = 0 \\ \dfrac{2}{(m-1)m} & \text{if } i \in \{1, \ldots, m-1\} \end{cases}$$

For this distribution, shown in Figure 1.1 with $|\mathbb{A}| = m = 5$, $E(G(W)) = 2$ for all m, but the Shannon entropy of W tends to 0 as m becomes large. The Shannon entropy of W is dominated by $P(W = 0)$, while the average guesswork depends heavily on the fact that if $W \neq 0$, then a large number of guesses will, on average, be required to identify it. A comparison of the average guesswork, $E(G(W))$, and the Shannon entropy of W is shown in Figure 1.2 for a range of alphabet sizes, m.

As Shannon entropy is not a good measure of average guesswork, what is the appropriate measure? In 1996, this is the question that Arikan [1] addressed, introducing regularity into the analysis by considered a sequence of guesswork problems with increasing string lengths. He considered a sequence of string distributions $\{W_k\}$, where W_k maps to $\mathbb{A}^k$,

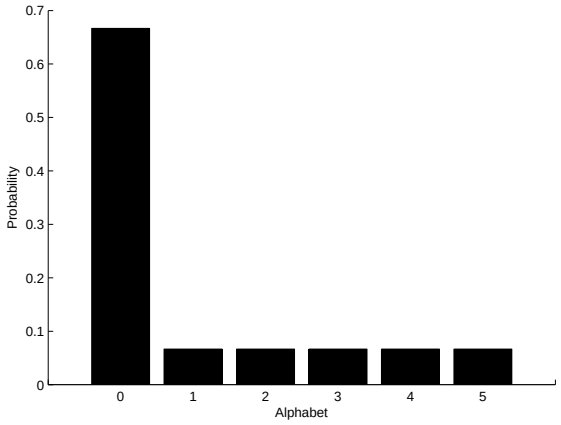

Figure 1.1: Probability mass function for $P(W = 0) = (m-2)/m$ and $P(W = i) = 2/(m^2 - m)$ for $i \in \{1, \ldots, m-1\}$ with $m = 5$.

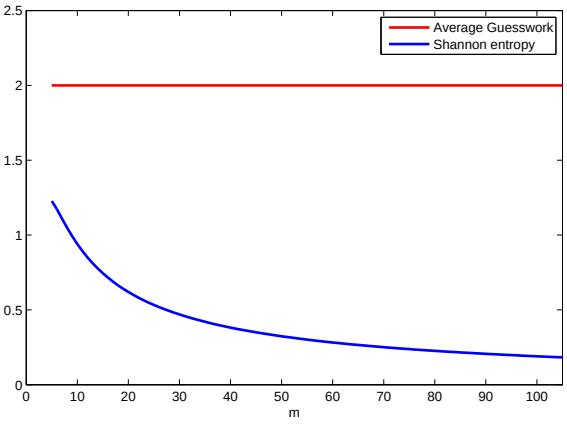

Figure 1.2: Comparison between the Shannon entropy, $H(W)$, and the average guesswork, $E(G(W))$, for the distribution of $P(W = 0) = (m-2)/m$ and $P(W = i) = 2/(m^2 - m)$ for $i \in \{1, \ldots, m-1\}$, shown as a function of m.

consisting of independent and identically distributed characters (i.i.d.), and analysed the moments of the guesswork distribution in the long string-length limit, i.e. as k becomes large. As k increases, the number of strings $|\mathbb{A}|^k = m^k$ grows exponentially and so the appropriate scaling to capture the dominant behaviour of the moments of guesswork is

$$\frac{1}{k} \log E(G(W_k)^\alpha), \text{ for } \alpha > 0.$$

In this i.i.d. character setting, Arikan established that

$$\lim_{k\to\infty} \frac{1}{k} \log E(G(W_k)^\alpha) = \alpha R_1 \left(\frac{1}{1+\alpha} \right), \text{ for } \alpha > 0, \tag{1.1}$$

where $R_1(\beta)$ is the Rényi entropy, e.g. [14], of a single character W_1 given by

$$R_1(\beta) = \frac{1}{1-\beta} \log \sum_{w\in\mathbb{A}} P(W_1 = w)^\beta, \text{ for } \beta > 0.$$

In particular, for $\alpha = 1$, $E(G(W_k)) \approx \exp(kR_1(1/2))$ and the expected guesswork grows with exponent $R_1(1/2)$, the Rényi entropy of a character with a parameter $1/2$, a value that is necessarily no smaller than the Shannon entropy of W_1.

The i.i.d. character assumption made by Arikan was subsequently significantly relaxed, proving the robustness of this result, with analogous deductions made replacing the Rényi entropy of a single character with the Rényi rate, also known as specific Rényi entropy,

$$R(\beta) = \lim_{k\to\infty} \frac{1}{k} R_k(\beta), \text{ where } R_k(\beta) = \frac{1}{1-\beta} \log \sum_{w\in\mathbb{A}^k} P(W_k = w)^\beta.$$

The ultimate result being that

$$\lim_{k\to\infty} \frac{1}{k} \log E(G(W_k)^\alpha) = \alpha R\left(\frac{1}{1+\alpha}\right), \text{ for } \alpha > 0.$$

in greater generality than strings made of i.i.d. characters.

This work began with a paper by Malone and Sullivan [37] in 2004 which extended Arikan's result to the case where the characters of each W_k are formed a stationary, irreducible Markov chain, e.g. Billingsley [5]. Employing the tools of Ergodic Theory [54], in the same year, Pfister and Sullivan [46] relaxed Arikan's i.i.d. assumption significantly further still. In that paper, the process $\{W_k\}$ is constructed via an ergodic measure v on $\mathbb{A}^\mathbb{N}$ where $\mathbb{A}$ is equipped with the discrete topology and its Borel $\sigma-$algebra. Let v_k represent v's restriction to $\mathbb{A}^k$, $v_k(w_k) = v(A(w,k))$ where $A(w,k) := \{w' \in \mathbb{A}^\mathbb{N} : [w']_k = w_k\}$ and $[w]_k$ denotes the first k characters of w. Then this setup relates to previous work by defining the string distributions as $P(W_k = w_k) = v_k(w_k)$ for all $w_k \in \mathbb{A}^k$. The Shannon entropy of any shift invariant probability measure ρ on $\mathbb{A}^\mathbb{N}$ is defined to be

$$h(\rho) := -\lim_{k\to\infty} \frac{1}{k} \sum_{w_k\in\mathbb{A}^k} \rho_k(w_k) \log \rho_k(w_k).$$

Two conditions are imposed on the string source v in [46]. The first is that for all shift invariant probability measures ρ on $\Sigma^v = \cap_k \{w \in \mathbb{A}^\mathbb{N} : v([w]_k) > 0\}$, for any open neighbourhood U of ρ and given $\epsilon > 0$, there exists an ergodic shift invariant probability measure $\rho^* \in U$ such that $h(\rho^*) \geq h(\rho) - \epsilon$. The second is that there exists a continuous nonnegative function $e_v : \mathbb{A}^\mathbb{N} \to \mathbb{R}$ that satisfies

$$\lim_{k\to\infty} \sup_{\{w\in\mathbb{A}^k : v_k(w) > 0\}} \frac{1}{k} |\log v_k(w) + e_v(w)| = 0.$$

As well as generalizing the i.i.d. and Markovian character assumptions, they also showed that equation (1.1) holds for $\alpha > -1$. Pfister and Sullivan (private communication)

suggested that this latter extension, which appears unusual in considering the scaling limits of $E(G(W_k)^\alpha)$, was done solely for mathematical generality as their arguments did not require α to be greater than 0. This extension will, however, prove crucial to the developments in this thesis.

Finally, we mention one last generalization of Arikan's source assumptions. Hanawal and Sundaresan [25] showed that

$$\lim_{k\to\infty} \frac{1}{k} \log E(G(W_k)^\alpha)$$

exists for $\alpha > 0$ if and only if the Rényi rate $R(\beta)$ exists for $\beta > 0$, but that it is unknown if the former corresponds to $\alpha R(1/(1+\alpha))$ in all such cases.

All of the work described so far relates to results on the moments of guesswork, but it does not provide a direct approximation to the guesswork distribution, $\{P(G(W_k) = n) : n \in \{1, \ldots, m^k\}\}$, which is the initial aim in this thesis. In doing so, we will develop a new set of tools for studying guesswork that allow us to substanially extend its remit.

The work presented in this thesis begins with the simplest of observations: the scaling of the moments in equation (1.1) can be rewritten as

$$\lim_{k\to\infty} \frac{1}{k} \log E(G(W_k)^\alpha) = \lim_{k\to\infty} \frac{1}{k} \log E\left(e^{\alpha \log G(W_k)}\right) = \Lambda(\alpha)$$

demonstrating that these earlier results can, in fact, be considered as identifying the scaled Cumulant Generating Function (sCGF), Λ, [16] of the process $\{k^{-1} \log G(W_k)\}$ for $\alpha > -1$. This suggests leveraging the results in [1, 37, 46, 25] to prove a Large Deviation Principle (LDP) for the process $\{k^{-1} \log G(W_k)\}$ from which estimates on the guesswork distribution can be developed. Due to the covariance of the LDP, this provides us with a new tool that we use to extend the guesswork remit to guessing over noisy channels as well as multi-user systems. In the latter case, the resulting rate functions are typically non-convex, which explains why approaches via the sCGF would not suffice to obtain answers.

The key contributions to the study of guesswork that can be found in this thesis are:

- Chapter 3 extends equation (1.1) to $\alpha \leq -1$ in order to fully characterize the sCGF

and proves, with little more than the assumptions of [46], that $\{k^{-1}\log G(W_k)\}$ satisfies a large deviations principle. We identify the rate function, which need not be strictly convex, in terms of the Legendre-Fenchel transform of the sCGF. This is then used to get direct estimates on the probability mass function of guesswork. Returning to Massey's original observations, we show that the expectation of the logarithm converges to the specific Shannon entropy of the string selection process. This work, performed in collaboration with K. Duffy, was published in IEEE Transactions on Information Theory in 2013, [10] and provides insights into both brute force searching and lossless coding.

- Chapter 4 uses guesswork to show that a commonly used approximation in Information Theoretic Security [6], which is suggested by the Asymptotic Equipartition Property and source compression, that every string inside a typical set is uniformly distributed is ill advised. Most importantly we prove that the expected guesswork for a source conditioned to create strings in the typical set grows at a lower exponential rate than that of the uniform approximation. The case of independent and identically distributed characters is published in the proceedings of the IEEE International Symposium on Information Theory in 2013 [11], based on work performed in collaboration with K. Duffy as well as F. du Pin Calmon and M. Médard (MIT). A more general version is presented here.

- Chapter 5 examines the case of guessing the missing characters of a string sent across a binary erasure channel, again establishing a LDP for the resulting guesswork. This provides an unusual result in the world of wiretap channels as it can be easier, on average, to guess a string over a channel that is, on average, noisier. This work appears in the proceedings from the 2013 Asilomar Conference on Signals, Systems, and Computers [12] and was work performed in collaboration with K. Duffy as well as F. du Pin Calmon and M. Médard (MIT).

- Chapter 6 significantly extends the guesswork question to multi-user systems. In particular, it studies the case of strings being selected independently by V users, with the inquisitor wishing to identify U of them. It is assumed that the inquisitor can guess (user, word) combinations one at a time. We show that, unlike the single-user setting, an optimal strategy does not always exist, but that there is a strategy that is optimal in an asymptotic sense. We prove a LDP for the guesswork process in this setting, establishing that the rate function may, in general, be non-

convex which explains why this would not have been possible prior to the work in Chapter 3. In the restricted setting where user's string statistics are the same, the rate function is necessarily convex and we find that the average guesswork growth-rate is $R((V-U+1)/(V-U+2))$, generalizing the single user case. The results contained in this chapter informs the security of multi-user and cloud-based systems from both a system designer and hacker's point of view. This work performed in collaboration with K. Duffy as well as F. du Pin Calmon and M. Médard (MIT), and has been accepted at IEEE Transactions on Information Theory. The submitted version can be found on the ArXiV [13].

- The thesis comes to a close in Chapter 7 with some speculative, partial results and a conjecture. On studying the process $\{k^{-1}\log G(W_k)\}$, it becomes apparent that the scaling provides a good approximation to the guesswork distribution $P(G(W_k)=n)$ for small n, but becomes increasingly approximate for larger n. This suggests considering reversing the order of G, which we denote G^R, guessing from the least likely string to the most likely string, to get better approximations at the other end of the guesswork distribution. Establishing results for the LDP of $\{k^{-1}\log G^R(W_k)\}$ appear fraught, but we establish the LDP for the simplest case, strings made of i.i.d. binary characters, establishing that the resulting large deviations rate function is strictly concave. This estimate is then combined with the estimate from [10] to create a conjectured, more accurate approximation to the guesswork distribution. This work has not yet been submitted for publication.

Before providing, in Chapter 2, a review of the tools in Large Deviations that are relevant in this thesis, we end this introduction by summarizing other developments in the study of guesswork that are related, but somewhat tangential, to work developed here.

Arikan and Merhav [3] altered the framework of Arikan [1] by saying that the inquisitor stops guessing once they identify the chosen string within a distance $D \in [0,\infty)$ of W, based on a metric $d : \mathbb{A}^k \times \mathbb{A}^k \to [0,\infty)$. Assuming that strings are created from i.i.d. characters, the scaling of guesswork moments is identified for all $\alpha > 0$. Those results are further expanded upon by Merhav, Roth and Arikan [43] by considering a successive round of guessing. In the first round, as above, the inquisitor is informed when W is found within some distance D. In the next round, the resolution is increased and the inquisitor is alerted when the string is identified within a distance D', possibly using a different metric $d' : \mathbb{A}^k \times \mathbb{A}^k \to [0,\infty)$. Assuming that strings are created from

i.i.d. characters, the scaling of guesswork moments is lower bounded for all $\alpha > 0$, but achievability of this lower bound was left as an open problem. Ghazaryan and van der Meulen [22] subsequently showed that the minimal exponent for the average guesswork can be achieved if $d = d'$.

The closest piece of work to our multi-user analysis is that of Merhav and Arikan [42]. They consider a string picked with i.i.d. characters from a finite alphabet and encrypted using a key of the same length chosen perfectly uniformly, potentially using a different sized alphabet. It is assumed that the inquisitor knows everything needed to decrypt the message except the message and the key. Thus the inquisitor has the choice of either guessing the string directly or the key and using it to decrypt the string. The authors identify how the moments of guesswork scale in this case. We demonstrate, however, that this is one of the situations where the rate function for the associated LDP is non-convex and so the LDP could not be deduced by their methodology.

The results in [42] have been extended in several ways. In [28] Hayashi and Yamamoto examined the case where there is an additional i.i.d. source correlated to the first used for coding purposes. Harountunian and Ghazaryan [27] operate in the setting of Arikan and Merhav [3] so that an inquisitor need only identify the string within a certain distance, but allow the second string to not necessarily be uniform and consider only expected guesswork. Hanawal and Sundaresan [26] returned to the bounds of [1] showing that they are tight for Markovian and unifilar sources.

Sundaresan [49] studies the case where the inquisitor does not know the distribution of W, but believes they do. Therefore they might not be asking in decreasing order of string-likelihood. In this case all that can be found are bounds on the amount of guesses required from the inquisitor. Altering that model slightly, what if the inquisitor knows that they don't know the process used to pick the string? In [49] Sundaresan shows that, from an asymptotic point of view, if the inquisitor knows the strings were created in an i.i.d. fashion, there still exists an ordering for universal guessing that minimises the exponent of the average guesswork. To achieve that ordering, the set of all possible strings are divided into types. Two strings are in the same type if they contain the same number of characters of each type. The ordering is defined by guessing from the type containing the least amount of strings to the type containing the most, breaking any ties arbitrarily.

Sundaresan [50] used length functions to show the link between guesswork and compres-

sion. A length function is a function $L : \mathbb{A} \to \mathbb{N}$ that satisfies the Kraft inequality

$$\sum_{w \in \mathbb{A}} e^{-L(w)} \leq 1.$$

Length functions allow Sundaresan connect guesswork and Campbell's coding problem [9]. This comparison is extended by Hanawal and Sundaresan [24] to compare guesswork and compression with a countably infinite alphabet.

Malone and Sullivan [38] provide an experimental study related to Massey's original work. Taking passwords from leaked datasets and treating them as single characters, they compare the average guesswork to the Shannon entropy, providing a practical illustration of Massey's results.

Boztas [7] and Dragomir [18] established tighter bounds in the finite string length case than Arikan's [1]. Boztas [8] examines strategies where the inquisitor has no memory and so guesses in a randomized order, potentially repeating queries, and identifies a strategy that minimizes the expected guesswork exponent.

Lundin and Lindskog, [34] and [35], examined the average guesswork required as an entropy. By considering a two-character word, they showed that it satisfies entropy-like properties, but it does not satisfy the natural generalization of the entropy chain rule.

In [2], Arikan and Boztas changed the original guesswork framework by allowing some uncertainty in whether or not the inquisitor has found W. To model this, if an inquisitor guesses any string that is not W, they are told they have not guessed W. If the inquisitor correctly guesses W, then the inquisitor is told that they have found W with some non-zero probability, but are told they have not have guessed W the rest of the time. To finish guessing the inquisitor must guess W and be told that they have done so. This implies that an inquisitor may have to repeat guesses to finish guessing. They study the case where the string length is fixed and find bounds for the expected guesswork as well as providing an order of guessing that minimizes the expected guesswork.

2 Large Deviations

2.1 Introduction

Large Deviation Theory is the study of the asymptotic likelihood of rare events. This is relevant in the study of Guesswork as the likelihood of any individual string being selected in a large collection decays as the string length becomes long. In this chapter we recap some of the results from Large Deviation Theory that will be used in later chapters. It is a well established subject, not all of which will be repeated here and, instead, this chapter focuses on the parts that are relevant to the rest of this thesis. This chapter serves as a brief recap of work that has already been done and so no proofs are provided. This chapter is based on material available in Dembo and Zeitouni's book [16].

2.2 Large Deviations

We restrict most of our discussion to the following setting: Let $\{Z_k\}$, $k \in \mathbb{N}$, be a stochastic process whose random variables take values in the real line, $\mathbb{R}$. All of the results recounted here have, however, versions where the random variables take values in general topological spaces.

Definition 2.1 (Cumulant Generating Function) *The Cumulant Generating Function, $\Lambda_k : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, of Z_k is defined by*

$$\Lambda_k(\alpha) = \log E\left(\exp\left(\alpha k Z_k\right)\right),$$

with the scaled Cumulant Generating Function (sCGF) of the process $\{Z_k\}$ defined by

$$\Lambda(\alpha) := \lim_{k\to\infty} \frac{1}{k}\Lambda_k(\alpha), \tag{2.1}$$

if it exists as an extended real number, i.e. in $\mathbb{R} \cup \{-\infty, \infty\}$.

The sCGF can be defined in greater generality, but the definition given above is sufficient for our purposes. In this chapter we adopt the standard notation that if Γ is a set in $\mathbb{R}$ then $\bar{\Gamma}$ is its closure, Γ^o is its interior and Γ^c is its complement.

Definition 2.2 (Level sets) *The level sets of a function $f : \mathcal{X} \to \mathbb{R}$ are defined for $\lambda \in \mathbb{R}$ by*

$$\Psi_f(\lambda) := \{x \in \mathcal{X} : f(x) \leq \lambda\}.$$

Definition 2.3 (Semi-continuity) *A function $f : \mathbb{R} \to \mathbb{R}$ is upper semicontinuous at $x_0 \in \mathbb{R}$ if*

$$\limsup_{x \to x_0} f(x) \leq f(x_0)$$

and f is lower semicontinuous at x_0 if

$$\liminf_{x \to x_0} f(x) \geq f(x_0).$$

A function is lower semicontinuous if it is lower semicontinuous at all $x \in \mathbb{R}$, with a similar definition for upper semicontinuous.

Definition 2.4 (Rate function) *A function $I : \mathcal{X} \to [0, \infty]$ is called a rate function if it is lower semicontinuous such that the level sets $\Psi_I(\lambda)$ are closed subsets of $\mathcal{X}$ and are compact for all $\lambda \in [0, \infty)$.*

The Large Deviations Principle (LDP) associates a rate function to a process $\{Z_k\}$ and governs the exponential decay rate of the probability of Z_k being in a given set as k increases.

Definition 2.5 (Large Deviations Principle) *The process $\{Z_k\}$ satisfies a Large Deviations Principle with a rate function I if for all $\Gamma \subset \mathbb{R}$,*

$$-\inf_{x \in \Gamma^o} I(x) \leq \liminf_{k \to \infty} \frac{1}{k} \log P(Z_k \in \Gamma) \leq \limsup_{k \to \infty} \frac{1}{k} \log P(Z_k \in \Gamma) \leq -\inf_{x \in \bar{\Gamma}} I(x). \quad (2.2)$$

This definition shows us the importance of a LDP as it bounds how the probabilities decay as the sequence progresses. In the case of Guesswork this happens as the strings become longer. In Chapter 3 this will be used to gain direct estimates on the n^{th} most likely word.

A common candidate for the rate function is the Legendre Fenchel transform of the sCGF, $\Lambda(\cdot)$, given in equation (2.1).

Definition 2.6 (Legendre Fenchel) *Define the Legendre Fenchel transform, Λ^* of Λ by*

$$\Lambda^*(x) := \sup_{\alpha \in \mathbb{R}} \{\alpha x - \Lambda(\alpha)\}.$$

One property of the Legendre Fenchel transform of any function that is important for us is that it is convex (see e.g. [16][Lemma 2.2.5]).

If it is the case that the Legendre Fenchel transform, Λ^* of Λ is to be the rate function governing an LDP, there are two significant definitions: exposed points; and essentially smooth. The definition of exposed points can be thought of heuristically as points whose derivative lies tangent to the function.

Definition 2.7 (Exposed point) *Let $x, y \in \mathbb{R}$, then y is an exposed point of Λ^* if for some $\alpha \in \mathbb{R}$ and all $x \neq y$,*

$$\alpha y - \Lambda^*(y) > \alpha x - \Lambda^*(x),$$

and α is called an exposed hyperplane.

Definition 2.8 (Effective domain) *The effective domain of Λ is $D_\Lambda := \{\alpha \in \mathbb{R} : \Lambda(\alpha) < \infty\}$ with its interior denoted by D_Λ^o.*

Definition 2.9 (Essentially Smooth) *A function $\Lambda : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is essentially smooth if*

- *$\Lambda(\cdot)$ is convex*
- *D_Λ^o is non-empty*
- *$\Lambda(\cdot)$ is differentiable throughout D_Λ^o*

- *the absolute value of the derivative of $\Lambda(\cdot)$ converges to ∞ over any sequence of points that converge to the boundary of D_Λ^o.*

Next we state the Gärtner-Ellis theorem, which provides sufficient, though not necessary, conditions for $\{Z_k\}$ to satisfy a LDP. This is important as we will see later that the sCGF of $\{k^{-1}\log G(W_k)\}$ does not satisfy all of the conditions of the Gärtner-Ellis theorem. We state it here as it is the basis so that we can use a more general version known as Baldi's Theorem as part of the proof that $\{k^{-1}\log G(W_k)\}$ satisfies a LDP and to show and illustrate why other proof techniques are necessary.

Theorem 2.1 (Gärtner-Ellis Theorem [16](Theorem 2.3.6)) *Let $\{Z_k\}$ be a stochastic process whose random variables take values in $\mathbb{R}$. Assume the origin belongs to D_Λ^o. For any set $\Gamma \subset \mathbb{R}$,*

$$\limsup_{k\to\infty} \frac{1}{k}\log P\left(Z_k \in \bar{\Gamma}\right) \leq -\inf_{x\in\bar{\Gamma}} \Lambda^*(x)$$

and

$$\liminf_{k\to\infty} \frac{1}{k}\log P\left(Z_k \in \Gamma^o\right) \geq -\inf_{x\in\Gamma^o\cap\mathcal{F}} \Lambda^*(x),$$

where $\mathcal{F}$ is the set of exposed points of $\Lambda^*(x)$ whose exposing hyperplanes belong to D_Λ^o. If $\Lambda(\cdot)$ is an essentially smooth, lower semicontinuous function, then the LDP holds with the rate function $\Lambda^*(\cdot)$.

Definition 2.10 (Exponentially tight) *A stochastic process $\{Z_k\}$ is exponentially tight if for every $\alpha < \infty$, there exists a compact set $K_\alpha \subset \mathbb{R}$ such that*

$$\limsup_{k\to\infty} \frac{1}{k}\log P\left(Z_k \in K_\alpha^c\right) < -\alpha.$$

A more general version of the Gärtner-Ellis Theorem is Baldi's Theorem.

Theorem 2.2 (Baldi's Theorem [16](Theorem 4.5.20),[4]) *Assume that $\{Z_k\}$ is*

a stochastic process of exponentially tight random variables on $\mathbb{R}$*. For every set* $\Gamma \subset \mathbb{R}$*,*

$$\limsup_{k\to\infty} \frac{1}{k} \log P\left(Z_k \in \bar{\Gamma}\right) \leq -\inf_{x\in\bar{\Gamma}} \Lambda^*(x).$$

Let $\mathcal{F}$ *be the set of exposed points of* $\Lambda^*(x)$ *with* $\Lambda(\xi\alpha) < \infty$ *for some* $\xi > 1$*. Then, for every set* $\Gamma \subset \mathbb{R}$*,*

$$\liminf_{k\to\infty} \frac{1}{k} \log P\left(Z_k \in \Gamma^o\right) \geq -\inf_{x\in\Gamma^o\cap\mathcal{F}} \Lambda^*(x).$$

If for every set Γ^o*,*

$$\inf_{x\in\Gamma^o\cap\mathcal{F}} \Lambda^*(x) = \inf_{x\in\Gamma^o} \Lambda^*(x),$$

then $\{Z_k\}$ *satisfies a LDP with the rate function* $\Lambda^*(x)$*.*

The sCGF approach is not the only method available to prove a LDP. One that does not rely on exposed points appears in both Lewis and Pfister [31], which attributed it to Ruelle and Lanford, and Dembo and Zeitouni [16]. Here we give a restricted version of that result, suitable for the needs of later chapters.

Theorem 2.3 ([31], [16](Theorem 4.1.11)) *Let* $\{Z_k\}$ *be a stochastic process and assume that* Z_k *takes values in* $\mathcal{X}$*, a compact subset of* $\mathbb{R}$*. If for all* $x \in \mathcal{X}$ *and some rate function* I*,*

$$\begin{aligned} -I(x) &\leq \lim_{\epsilon\to 0} \liminf_{k\to\infty} \frac{1}{k} \log P\left(Z_k \in (x-\epsilon, x+\epsilon)\right) \\ &\leq \lim_{\epsilon\to 0} \limsup_{k\to\infty} \frac{1}{k} \log P\left(Z_k \in (x-\epsilon, x+\epsilon)\right) \leq -I(x) \end{aligned}$$

then $\{Z_k\}$ *satisfies a LDP with the rate function* I*.*

The Contraction Principle establishes that the Large Deviation Principle is a covariant notion. If $\{Z_k\}$ satisfies a LDP and f is a continuous function, then it states that $\{f(Z_k)\}$ also satisfies a LDP.

Theorem 2.4 (Contraction Principle [16](Theorem 4.2.1)) *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be Hausdorff topological spaces and* $f : \mathcal{X} \to \mathcal{Y}$ *be a continuous function. Assume the*

process $\{Z_k\}$ *on* $\mathcal{X}$ *satisfies a LDP with the rate function* $I : \mathcal{X} \to [0, \infty]$*, then* $\{f(Z_k)\}$ *satisfies the LDP in* $\mathcal{Y}$ *with the rate function*

$$J(y) := \inf\{I(x) : x \in \mathcal{X}, y = f(x)\}.$$

One final well-known theorem that we shall employ is Varadhan's lemma. In our setting, the Z_k will all take values in a compact subset of $\mathbb{R}$ and so the required condition will be automatically satisfied.

Theorem 2.5 (Varadhan's Lemma [16](4.3.1)) *Let* $\{Z_k\}$ *satisfy a LDP with a rate function* $I : \mathcal{X} \to [0, \infty]$*, and let* $\phi : \mathcal{X} \to \mathbb{R}$ *be any continuous function. Assume either the tail condition*

$$\lim_{n\to\infty} \limsup_{k\to\infty} \frac{1}{k} \log E\left(\exp(k\phi(Z_k)) 1_{\{\phi(Z_k)\geq n\}}\right) = -\infty$$

or the following moment condition for some $\xi > 1$*,*

$$\limsup_{k\to\infty} \frac{1}{k} \log E(\exp(\xi k\phi(Z_k))) < \infty.$$

Then

$$\lim_{k\to\infty} \frac{1}{k} \log E(\exp(k\phi(Z_k))) = \sup_{x\in\mathcal{X}}\{\phi(x) - I(x)\}.$$

3 Single User Guesswork

3.1 Introduction

Let $\mathbb{A} = \{0, \ldots, m-1\}$ be a finite alphabet and for each $k \geq 1$ let $W_k : \Omega \mapsto \mathbb{A}^k$ be a random string of length k. For each k, an inquisitor knows the probability mass function $\{P(W_k = w) : w \in \mathbb{A}^k\}$ and wishes to identify the random variable W_k. The inquisitor can make one guess of the sort "Is $W_k = w$?" at a time, and keeps guessing until W_k is correctly identified. We assume that the inquisitor guesses in such a way as to minimise the expected number of guesses required to identify W_k, which means guessing from the most likely string to the least likely string as in Massey [39] and Arikan [1]. This guesswork ordering is codified by a function $G : \mathbb{A}^k \to \{1, \ldots, m^k\}$ such that $P(W_k = w) > P(W_k = w')$ implies that $G(w) < G(w')$.

The main content of this chapter is the proof that $\{k^{-1} \log G(W_k)\}$ satisfies a Large Deviations Principle (LDP), the determination of the accompanying rate function, and an examination of the properties of the accompanying rate function. The LDP gives us a handle on how the probability of $k^{-1} \log G(W_k)$ being in a given set decays as the string length increases, from which we get an estimate on the probability mass function of the guesswork distribution $\{P(G(W_k) = n) : n \in \{1, \ldots, m^k\}\}$ itself.

As a corollary, the LDP is used to prove a conjecture by Arikan and Merhav [3] and Sundaresan [49] that $\lim_{k\to\infty} E(k^{-1} \log G(W_k))$ is the specific Shannon entropy of the source. Aside from the results in this chapter being a significant development in their own right, proving that $\{k^{-1} \log G(W_k)\}$ satisfies a LDP with a certain rate function is a necessary stepping stone for work in later chapters.

3.2 A Large Deviation Principle

We introduce the assumption on the process creating strings to give us some regularity.

Assumption 3.1 *For $\alpha > -1$, the scaled cumulant generating function, $\Lambda(\alpha)$, for*

$\{k^{-1}\log G(W_k)\}$ *exists, has a continuous derivative and*

$$\lim_{k\to\infty}\frac{1}{k}\log E(G(W_k)^\alpha) = \alpha R\left(\frac{1}{1+\alpha}\right)$$
$$= \lim_{k\to\infty}\frac{1}{k}(1+\alpha)\log\sum_{i=1}^{m^k} P(G(W_k)=i)^{1/(1+\alpha)}. \qquad (3.1)$$

Note that in Assumption 3.1 the limits are also assumed to exist. Assumption 3.1 is satisfied, for example, by the processes considered by Pfister and Sullivan [46].

Consider the sequence of random variables $\{k^{-1}\log G(W_k)\}$. Our starting point is the observation that the left hand side of (3.1) is equal to the sCGF of this sequence:

$$\Lambda(\alpha) = \lim_{k\to\infty}\frac{1}{k}\log E\left(\exp(\alpha\log G(W_k))\right) \text{ for } \alpha\in\mathbb{R}.$$

A reasonable supposition is that should $\{k^{-1}\log G(W_k)\}$ satisfy a LDP, the rate function will be the Legendre-Fenchel transform of Λ,

$$\Lambda^*(x) := \sup_{\alpha\in\mathbb{R}}\{\alpha x - \Lambda(\alpha)\}.$$

Thus we first need to determine $\Lambda(\alpha)$ for $\alpha\le -1$.

Lemma 3.1 (Existence of the sCGF) *Under assumption 3.1, for all* $\alpha\le -1$

$$\Lambda(\alpha) = \lim_{k\to\infty}\frac{1}{k}\log P\left(\frac{1}{k}\log G(W_k)=0\right) = \lim_{\beta\downarrow -1}\Lambda(\beta).$$

and that the above limits exist.

Proof: Let $\alpha\le -1$ and note that

$$\log P\left(\frac{1}{k}\log G(W_k)=0\right) \le \log\sum_{i=1}^{m^k} P(G(W_k)=i)i^\alpha$$
$$= \log E\left(\exp(\alpha\log G(W_k))\right) \le \log P\left(\frac{1}{k}\log G(W_k)=0\right) + \log\sum_{i=1}^{m^k} i^\alpha.$$

We know from Assumption 3.1 that $\lim_{\alpha\downarrow -1}\Lambda(\alpha)$ exists. So by the above equa-

tion $\limsup_{k\to\infty} k^{-1}\log P(k^{-1}\log G(W_k) = 0) \leq \Lambda(-1)$ and similarly we know that $\liminf_{k\to\infty} k^{-1}\log P(k^{-1}\log G(W_k) = 0) + \log\sum_{i=1}^{m^k} i^{\alpha} \geq \Lambda(-1)$ Taking limits while using the Principle of the Largest Term [16, Lemma 1.2.15] in conjunction with usual estimate that the harmonic series, $\sum_{i=1}^{n} n^{-1}$, is approximately $\log n$ (e.g. [29], Chapter 7, Theorem 10) if $\alpha = -1$ and boundedness of the sum if $\alpha < -1$, we have that the limit $\lim_{k\to\infty} k^{-1}\log P(k^{-1}\log G(W_k) = 0)$ exists and

$$\lim_{k\to\infty}\frac{1}{k}\log E(\exp(\alpha\log G(W_k))) = \lim_{k\to\infty}\frac{1}{k}\log P\left(\frac{1}{k}\log G(W_k) = 0\right)$$

for all $\alpha \leq -1$.

As Λ is the limit of a sequence of convex functions and is finite everywhere, it is continuous and therefore $\lim_{\beta\downarrow -1}\Lambda(\beta) = \Lambda(-1)$.

■

Thus the sCGF Λ exists and is finite for all α, with a potential discontinuity in its derivative at $\alpha = -1$. This discontinuity, when it exists, will have a bearing on the nature of the rate function governing the LDP for $\{k^{-1}\log G(W_k)\}$. Indeed, the following quantity will play a significant role in our results:

$$\gamma := \lim_{\alpha\downarrow -1}\frac{d}{d\alpha}\Lambda(\alpha). \tag{3.2}$$

The derivative on the right hand side of equation (3.2) has the interpretation of a tilted measure. As $\alpha \downarrow -1$ this measure will, in an appropriate sense, converge to the uniform measure on the set of strings with asymptotically maximal probability. In particular, we will prove that the number of strings with approximately equally highest probability is close to $\exp(k\gamma)$. In the special case where the $\{W_k\}$ are constructed of i.i.d. characters this is exactly true and the veracity of the following Lemma can be verified directly.

Lemma 3.2 (The number of most likely strings) *If $\{W_k\}$ are constructed of i.i.d. characters, then*

$$\gamma = \lim_{\alpha\downarrow -1}\frac{d}{d\alpha}\alpha R_1((1+\alpha)^{-1}) = \log|\{w : P(W_1 = w) = P(G(W_1) = 1)\}|,$$

where $|\cdot|$ indicates the number of elements in the set.

PROOF: This follows as

$$
\begin{aligned}
&\frac{d}{d\alpha}\alpha R_1\left(\frac{1}{1+\alpha}\right)\\
&=\log\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}-\frac{\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}\log P(W_1=w)}{(1+\alpha)\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}}\\
&=\log\max_{w'\in\mathbb{A}}P(W_1=w')^{1/(1+\alpha)}\sum_{w\in\mathbb{A}}\frac{P(W_1=w)^{(1/(1+\alpha))}}{\max_{w'\in\mathbb{A}}P(W_1=w')^{1/(1+\alpha)}}\\
&\quad-\frac{\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}\log P(W_1=w)}{(1+\alpha)\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}}\\
&=\log\sum_{w\in\mathbb{A}}\frac{P(W_1=w)^{(1/(1+\alpha))}}{\max_{w'\in\mathbb{A}}P(W_1=w')^{1/(1+\alpha)}}\\
&+\frac{1}{1+\alpha}\log\max_{w'\in\mathbb{A}}P(W_1=w')-\frac{\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}\log P(W_1=w)}{(1+\alpha)\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}}.
\end{aligned}
$$

Examining $P(W_1=w)/(\max_{w'\in\mathbb{A}}P(W_1=w'))$ we see that for $|\{w:P(W_1=w)=P(G(W_1)=1)\}|$ elements it is 1 and for every other element of $\mathbb{A}$ it is <1 and tends to 0 as $\alpha\downarrow-1$. This implies that

$$
\begin{aligned}
&\lim_{\alpha\downarrow-1}\log\max_{w'\in\mathbb{A}}P(W_1=w')^{1/(1+\alpha)}\sum_{w\in\mathbb{A}}\frac{P(W_k=w)^{(1/(1+\alpha))}}{\max_{w'\in\mathbb{A}}P(W_1=w')^{1/(1+\alpha)}}\\
&=\log|\{w:P(W_1=w)=P(G(W_1)=1)\}|.
\end{aligned}
$$

Then taking

$$
\frac{\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}(\log\max_{w'\in\mathbb{A}}P(W_1=w')-\log P(W_1=w))}{(1+\alpha)\sum_{w\in\mathbb{A}}P(W_1=w)^{(1/(1+\alpha))}}
$$

as α decreases to -1 achieves the desired result.

■

This i.i.d. result doesn't extend directly to the non-i.i.d. case and in general Lemma 3.2 can only be used to establish a lower bound on γ defined in equation (3.2):

$$
\gamma\geq\limsup_{k\to\infty}\frac{1}{k}\lim_{\alpha\downarrow-1}\frac{d}{d\alpha}\alpha R_k\left(\frac{1}{1+\alpha}\right), \tag{3.3}
$$

(e.g [48, Theorem 24.5]). This lower bound can be loose, as can be seen with the following example. Consider the sequence of distributions for some $\epsilon > 0$

$$P(W_k = i) = \begin{cases} m^{-k}(1+\epsilon) & \text{if } i = 1 \\ m^{-k}(1-\epsilon(m^k-1)^{-1})) & \text{otherwise.} \end{cases}$$

For each fixed k there is one most likely string and we have $\log(1) = 0$ on the right hand side of equation (3.3) by Lemma 3.2. The left hand side, however, gives $\log m$. Regardless, this intuition guides our understanding of γ, but the formal statement of it approximately capturing the number of most likely strings will transpire to be

$$\lim_{\alpha \downarrow -1} R(1/(1+\alpha)) = \lim_{k\to\infty} \frac{1}{k} \log \inf_{\{w: G(w) < \exp(k\gamma)\}} P(W_k = w).$$

The candidate rate function is the Legendre-Fenchel [48, Chapter 26] transform of the sCGF

$$\begin{aligned} \Lambda^*(x) &= \sup_{\alpha\in\mathbb{R}} \{x\alpha - \Lambda(\alpha)\} \\ &= \begin{cases} -x - \Lambda(-1) & \text{if } x \in [0, \gamma] \\ \sup_{\alpha\in\mathbb{R}}\{x\alpha - \Lambda(\alpha)\} & \text{if } x \in (\gamma, \log m], \\ +\infty & \text{if } x \notin [0, \log m]. \end{cases} \end{aligned} \tag{3.4}$$

The graphical illustrations of examples of $\Lambda^*(x)$ is shown in Figure 3.1 showing the three possible shapes of linear, linear then strictly convex or strictly convex, in each case $\Lambda^*(x) = \infty$ if $x < \log(1)$ or $x > \log m$. The LDP cannot be proved directly by Baldi's version of the Gärtner-Ellis theorem, [4], Theorem 2.2 or [16, Theorem 4.5.20], as Λ^* does not have exposing hyper-planes for $x \in [0, \gamma]$. Instead we use a combination of Baldi's theorem with the methodology described in detail in [31] where, as our random variables are bounded $0 \le k^{-1} \log G(W_k) \le \log m$, in order to prove the LDP it suffices to show that the following exist in $[0, \infty]$ for all $x \in [0, \log m]$ and equals $-\Lambda^*(x)$:

$$\begin{aligned} &\lim_{\epsilon\downarrow 0} \liminf_{k\to\infty} \frac{1}{k} \log P\left(\frac{1}{k}\log(G(W_k)) \in B_\epsilon(x)\right) \\ &= \lim_{\epsilon\downarrow 0} \limsup_{k\to\infty} \frac{1}{k} \log P\left(\frac{1}{k}\log(G(W_k)) \in B_\epsilon(x)\right), \end{aligned} \tag{3.5}$$

where $B_\epsilon(x) = (x - \epsilon, x + \epsilon)$.

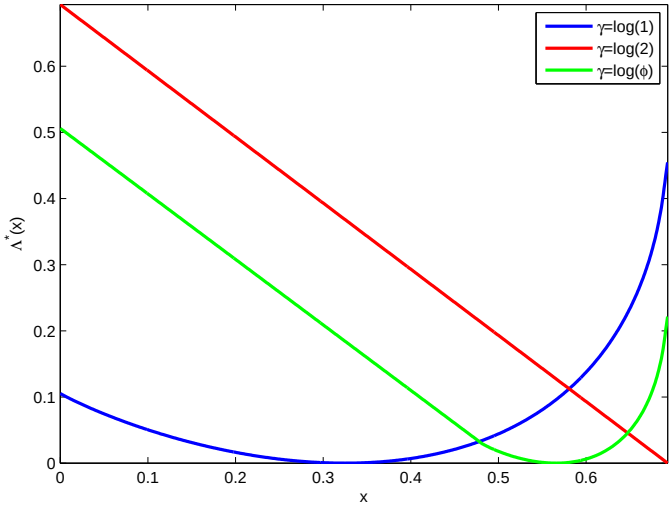

Figure 3.1: Illustration of equation (3.4). Shows the different possible shapes for $\Lambda^*(x)$ with $\gamma = \log(1)$, $\gamma \in (\log(1), \log m)$, in this case $\gamma = \log(\phi)$ with $\phi = (1+\sqrt{5})/2$, and $\gamma = \log m$, in this example $\gamma = \log(2)$.

Theorem 3.3 (The large deviations of guesswork) *Under assumption 3.1, the sequence* $\{k^{-1}\log G(W_k)\}$ *satisfies a LDP with rate function* Λ^*.

Proof: To establish (3.5) we have separate arguments depending on x. We divide $[0, \log m]$ into two parts: $[0, \gamma]$ and $(\gamma, \log m]$. Baldi's upper bound holds for any $x \in [0, \log m]$. Baldi's lower bound applies for any $x \in (\gamma, \log m]$ as Λ^* is continuous and, as $\Lambda(\alpha)$ has a continuous derivative for $\alpha > -1$, it only has a finite number of points without exposing hyper-planes in that region. For $x \in [0, \gamma]$, however, we need an alternate lower bound.

Consider $x \in [0, \gamma]$ and define the sets

$$K_k(x, \epsilon) := \left\{ w \in \mathbb{A}^k : k^{-1}\log G(w) \in B_\epsilon(x) \right\},$$

letting $|K_k(x, \epsilon)|$ denote the number of elements in each set. We have the bound

$$|K_k(x,\epsilon)| \inf_{w \in K_k(x,\epsilon)} P(W_k = w) \leq P\left(\frac{1}{k}\log G(W_k) \in B_\epsilon(x) \right). \tag{3.6}$$

As $\lfloor \exp(k(x-\epsilon)) \rfloor \leq |K_k(x,\epsilon)| \leq \lceil \exp(k(x+\epsilon)) \rceil$, we have that

$$x = \lim_{\epsilon \to 0} \lim_{k \to \infty} \frac{1}{k} \log |K_k(x,\epsilon)|. \tag{3.7}$$

By either the complementary upper bound to equation (3.6) or by Baldi's upper bound, we have that

$$\begin{aligned} &\lim_{\epsilon \downarrow 0} \limsup_{k \to \infty} \frac{1}{k} \log P\left(\frac{1}{k} \log G(W_k) \in B_\epsilon(x) \right) \leq |K_k(x,\epsilon)| \sup_{w \in K_k(x,\epsilon)} P(W_k = w) \\ &\leq x + \lim_{\alpha \downarrow -1} R(1/(1+\alpha)) \end{aligned}$$

as

$$P\left(\frac{1}{k} \log G(W_k) \in B_\epsilon(x) \right) \leq |K_k(x,\epsilon)| \sup_{w \in K_k(x,\epsilon)} P(W_k = w) \leq |K_k(x,\epsilon)| P(G(W_k) = 1).$$

Thus to complete the argument, for the complementary lower bound it suffices to show that for any $x \in [0, \gamma]$

$$\lim_{\epsilon \downarrow 0} \liminf_{k \to \infty} \inf_{w \in K_k(x,\epsilon)} \frac{1}{k} \log P(W_k = w) \geq \lim_{\alpha \downarrow -1} R(1/(1+\alpha)).$$

If $\Lambda^*(x) < \infty$ for some $x > \gamma$, then for $\epsilon > 0$ sufficiently small let x^* be such that $\Lambda^*(x^*) < \infty$ and $x^* - \epsilon > \max(\gamma, x + \epsilon)$. Then by Baldi's lower bound, which applies as $x^* \in (\gamma, \log m]$, we have

$$-\inf_{y \in B_\epsilon(x^*)} \Lambda^*(y) \leq \liminf_{k \to \infty} \frac{1}{k} \log P\left(\frac{1}{k} \log G(W_k) \in B_\epsilon(x^*) \right).$$

Now

$$\begin{aligned} P\left(\frac{1}{k} \log G(W_k) \in B_\epsilon(x^*) \right) &\leq |K_k(x^*,\epsilon)| \sup_{w \in K_k(x^*,\epsilon)} P(W_k = w) \\ &\leq |K_k(x^*,\epsilon)| \inf_{w \in K_k(x,\epsilon)} P(W_k = w), \end{aligned}$$

where in the last line we have used the monotonicity of guesswork and the fact that $x^* - \epsilon > x + \epsilon$. Taking lower limits and using equation (3.7) with $|K_k(x^*,\epsilon)|$, we have that

$$-\inf_{y \in B_\epsilon(x^*)} \Lambda^*(y) \leq x^* + \liminf_{k \to \infty} \inf_{w \in K_k(x,\epsilon)} \frac{1}{k} \log P(W_k = w)$$

for all such x^*, x. Taking limits as $\epsilon \downarrow 0$ and then limits as $x^* \downarrow \gamma$ we have

$$-\lim_{x^*\downarrow\gamma} \Lambda^*(x^*) \leq \gamma + \lim_{\epsilon\downarrow 0} \liminf_{k\to\infty} \inf_{w\in K_k(x,\epsilon)} \frac{1}{k} \log P(W_k = w),$$

but $\lim_{x^*\downarrow\gamma} \Lambda^*(x^*) = -\gamma - \lim_{\alpha\downarrow -1} R(1/(1+\alpha))$ so that

$$\lim_{\epsilon\downarrow 0} \liminf_{k\to\infty} \inf_{w\in K_k(x,\epsilon)} \frac{1}{k} \log P(W_k = w) = \lim_{\alpha\downarrow -1} R(1/(1+\alpha)),$$

as required.

Only one case remains. If $\Lambda^*(x) = \infty$ for all $x > \gamma$, then we require an alternative argument to ensure that

$$\liminf_{k\to\infty} \inf_{w\in K_k(x,\epsilon)} \frac{1}{k} \log P(W_k = w) \geq \lim_{\alpha\downarrow -1} R(1/(1+\alpha)).$$

Note that in this case $\gamma = \log m$ and as $\Lambda'(\alpha) \leq \log m$ for all α it implies that $\Lambda'(\alpha) = \gamma$ for all $\alpha > -1$. Then as $\Lambda(0) = 0$ and $\Lambda'(\alpha) = \gamma$ for all α, using equation (3.4) we have that $\lim_{\alpha\downarrow -1} R(1/(1+\alpha)) = -\gamma$. Let $x < \gamma$. This situation happens if, in the limit, the distribution of strings is near uniform on the set of all strings with positive probability. To see this note that $H(W) = \Lambda'(0) = \log m$. Consider

$$l = \limsup_{k\to\infty} \sup_{w\in K_k(x+2\epsilon,\epsilon)} \frac{1}{k} \log P(W_k = w) \leq \liminf_{k\to\infty} \inf_{w\in K_k(x,\epsilon)} \frac{1}{k} \log P(W_k = w).$$

We shall assume that $l < \lim_{\alpha\downarrow -1} R(1/(1+\alpha))$ and show this results in a contradiction. Let $\epsilon > 0$, then there exists N_ϵ such that for all $k \geq N_\epsilon$, $P(G(W_k) = i) \leq \exp(k(\lim_{\alpha\downarrow -1} R(1/(1+\alpha)) + \epsilon))$, for all $i \in \{1, \ldots, m^k\}$,

$$\begin{aligned} &P(G(W_k) = i) \leq \exp(k(l+\epsilon)), \text{ for all } i \in \{\exp(k(x+\epsilon)), \ldots, m^k\} \\ &\text{and } P(G(W_k) \geq \exp(k(\gamma+\epsilon))) \leq \exp\left(\frac{-k}{\epsilon}\right). \end{aligned}$$

Let $0 < \epsilon < \min(\lim_{\alpha\downarrow -1} R(1/(1+\alpha)) - l, \gamma - x)/2$ be given, then, using a potentially

gross overestimate that suffices for our purposes, we have that

$$\begin{aligned}\sum_{w\in\mathbb{A}^k} P(W_k = w) &= \sum_{i=1}^{m^k} P(G(W_k) = i) \\ &\leq \exp(k(x+\epsilon))\exp\left(k(\lim_{\alpha\downarrow -1} R(1/(1+\alpha)) + \epsilon)\right) \\ &\quad + \exp\left(k(\gamma+\epsilon)\right)\exp(k(l+\epsilon)) + \exp\left(\frac{-k}{\epsilon}\right)\end{aligned}$$

for all $k > N_\epsilon$, but as $l < \lim_{\alpha\downarrow -1} R(1/(1+\alpha)) = -\gamma$ this is strictly less than 1 for k sufficiently large and thus $l = \lim_{\alpha\downarrow -1} R(1/(1+\alpha))$. Finally, for $x = \gamma$, and $\epsilon > 0$, note that we can decompose $[0, \log m]$ into three parts, $[0, \gamma-\epsilon] \cup (\gamma-\epsilon, \gamma+\epsilon) \cup [\gamma+\epsilon, \log m]$, where the scaled probability of the guesswork being in either the first or last set is decaying, but

$$0 = \lim_{k\to\infty} \frac{1}{k} \log P\left(\frac{1}{k}\log G(W_k) \in [0, \log m]\right)$$

and so the result follows from an application of the principle of the largest term.

Thus for any $x \in [0, \log m]$,

$$\begin{aligned}&\lim_{\epsilon\downarrow 0} \liminf_{k\to\infty} \frac{1}{k} \log P\left(\frac{1}{k}\log(G(W_k)) \in B_\epsilon(x)\right) \\ &= \lim_{\epsilon\downarrow 0} \limsup_{k\to\infty} \frac{1}{k} \log P\left(\frac{1}{k}\log(G(W_k)) \in B_\epsilon(x)\right) = -\Lambda^*(x)\end{aligned}$$

and the LDP is proved.

■

In establishing the LDP, we have shown that the rate function in equation (3.4) must have the form of a straight line in $[0, \gamma]$ with a slope of -1 followed by a strictly convex section. The initial straight line comes from all strings that are, in an asymptotic sense, of greatest likelihood.

Theorem 3.3 uses Assumption 3.1 in its proof, however the proof of Theorem 3.3 does not show that Assumption 3.1 is required for the theorem to hold. The following lemma

shows that assuming $\alpha R(1/(1+\alpha)) = \Lambda(\alpha)$ for $\alpha > -1$ is not sufficient to prove Theorem 3.3.

Lemma 3.4 *Theorem 3.3 no longer holds if $\Lambda'(\alpha)$ is not continuous for $\alpha > -1$.*

PROOF: For this we just require a counter example, the one we use here is a four letter alphabet where 2^k of the strings have probability $0.5/2^k$ and the remainder of the probability is divided equally across the remaining strings so they each have $0.5/(4^k - 2^k)$ probability of being the chosen string.

To show that this example is a counter example we have two conditions that need to be satisfied, first we need that the continuity of $\Lambda'(\alpha)$ is the only condition that is broken and second that Theorem 3.3 is not true for this example. Here we check the rest of Assumption 3.1 to hold so that $\alpha R(1/(1+\alpha)) = \Lambda(\alpha)$ for $\alpha > -1$. The Rényi entropy for our example is,

$$\begin{aligned}\alpha R\left(\frac{1}{1+\alpha}\right) &= (1+\alpha)\lim_{k\to\infty}\frac{1}{k}\log\left(2^k\left(\frac{0.5}{2^k}\right)^{(1/(1+\alpha))} + (4^k - 2^k)\left(\frac{0.5}{4^k-2^k}\right)^{(1/(1+\alpha))}\right)\\ &= (1+\alpha)\lim_{k\to\infty}\frac{1}{k}\log((2^k)^{(\alpha/(1+\alpha))} + (4^k-2^k)^{(\alpha/(1+\alpha))}),\end{aligned}$$

which by the principle of largest term, [16, Lemma 1.2.15],

$$\alpha R\left(\frac{1}{1+\alpha}\right) = \begin{cases}\alpha\log 2 \text{ if } \alpha \in (-1,0)\\ \alpha\log 4\alpha > 0.\end{cases}$$

For the sCGF

$$\begin{aligned}
\Lambda(\alpha) &= \lim_{k\to\infty} \frac{1}{k} \log \left(\frac{1}{2^k} \sum_{i=1}^{2^k} i^\alpha + \frac{1}{4^k - 2^k} \sum_{i=2^k+1}^{4^k} i^\alpha \right) \\
&\geq \lim_{k\to\infty} \frac{1}{k} \log \left(\frac{1}{2^k} \int_0^{2^k} x^\alpha dx + \frac{1}{4^k - 2^k} \int_0^{4^k - 2^k} x^\alpha dx \right) \\
&= \lim_{k\to\infty} \frac{1}{k} \log \left(\frac{1}{2^k} \frac{(2^k)^{1+\alpha}}{1+\alpha} + \frac{1}{4^k - 2^k} \frac{(4^k - 2^k)^{1+\alpha}}{1+\alpha} \right) \\
&= \begin{cases} \alpha \log 2 \text{ if } \alpha \in (-1, 0) \\ \alpha \log 4\alpha > 0. \end{cases}
\end{aligned}$$

This is using the Principle of Largest Term again for the last line. For the lower bound

$$\begin{aligned}
&\lim_{k\to\infty} \frac{1}{k} \log \left(\frac{1}{2^k} \sum_{i=1}^{2^k} i^\alpha + \frac{1}{4^k - 2^k} \sum_{i=2^k+1}^{4^k} i^\alpha \right) \\
&\leq \lim_{k\to\infty} \frac{1}{k} \log \left(\frac{1}{2^k} (2^k)^\alpha + \frac{1}{4^k - 2^k} (4^k)^\alpha \right) = \begin{cases} \alpha \log 2 \text{ if } \alpha \in (-1, 0) \\ \alpha \log 4\alpha > 0. \end{cases}
\end{aligned}$$

simply by taking the largest term for each entry in the sums and the Principle of Largest Term.

This shows that the sCGF is described by Rényi entropy for $\alpha > -1$. Figure 3.2 shows the shape of the sCGF for this specific example. While $\Lambda^*(x)$ may not be the rate function in this example, a LDP can be established using other means and we label the rate function $I(x)$ instead. To calculate the rate function, $I(x)$, if $x \in (\log 2, \log 4)$ we need to find

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0} \lim_{k\to\infty} \frac{1}{k} \log P \left(\frac{1}{k} \log G(W_k) \in (x - \epsilon, x + \epsilon) \right) \\
&= \lim_{\epsilon\downarrow 0} \lim_{k\to\infty} \frac{1}{k} \log |K_k(x, \epsilon)| \inf_{w \in K_k(x,\epsilon)} P(W_k = w) \\
&= x + \lim_{k\to\infty} \frac{1}{k} \log \frac{1}{4^k - 2^k} \\
&= -x + \log 4.
\end{aligned}$$

The rate function for any other x can be worked out, using Baldi's theorem [4][16, Theorem 4.5.20] to be $I(x) = \infty$ if $x < 0, x > \log 4$ and $I(x) = \log 2 - x$ if $x \in [0, \log 2]$.

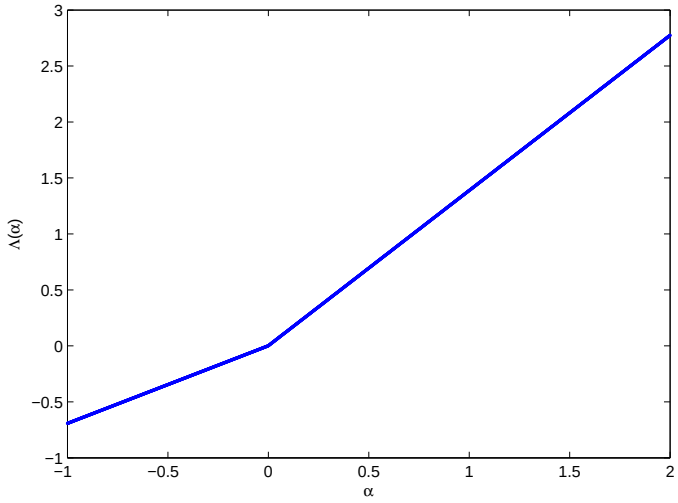

Figure 3.2: Illustration of Lemma 3.4. Shows the sCGF, $\Lambda(\alpha)$, if characters are chosen from $\mathbb{A} = \{0, 1, 2, 3\}$ and 2^k of the possible strings have probability $0.5(2^{-k})$ and the remaining $4^k - 2^k$ have probability $0.5(4^k - 2^k)^{-1}$. The discontinuity in the derivative at $\alpha = 0$ can clearly be seen.

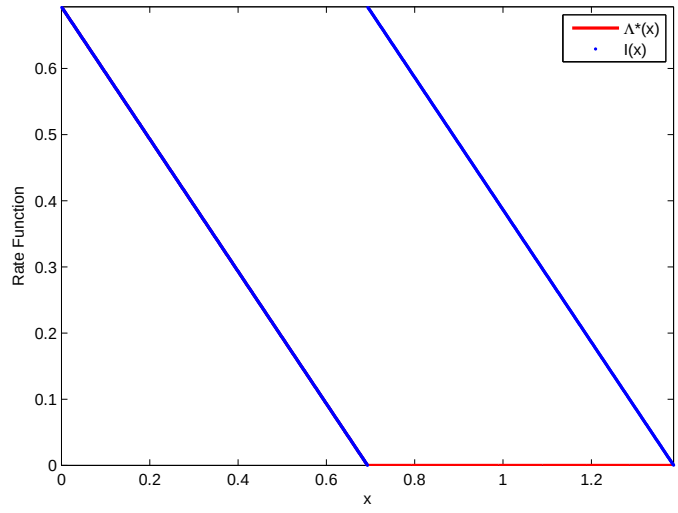

Figure 3.3: Illustration of Lemma 3.4. Shows the rate if characters are chosen from $\mathbb{A} = \{0, 1, 2, 3\}$ and 2^k of the possible strings have probability $0.5(2^{-k})$ and the remaining $4^k - 2^k$ have probability $0.5(4^k - 2^k)^{-1}$. It compares this rate function, $I(x)$, to the Legendre Fenchel transform, $\Lambda^*(x)$, of the sCGF of the same process showing that they agree on $[0, \log 2]$ but not on $(\log 2, \log 4]$.

We compare $I(x)$ to the Legendre Fenchel transform of the sCGF, $\Lambda^*(x)$, in Figure 3.3. Obviously these can be seen to be different showing that the assumption that $\Lambda'(\alpha)$ be continuous for $\alpha > -1$ is needed for Theorem 3.3 .

■

While the LDP is for the sequence $\{k^{-1}\log G(W_k)\}$, it can be used to develop the direct estimate of the distribution of $G(W_k)$ given by

$$P(G(W_k) = n) \asymp \frac{1}{n}\exp(-k\Lambda^*(k^{-1}\log n)). \tag{3.8}$$

which can't be derived from previous results. The next corollary provides a rigorous statement, but an intuitive, non-rigorous argument for understanding the result therein is that from the LDP we have the approximation that for large k

$$dP\left(\frac{1}{k}\log G(W_k) = x\right) \approx \exp(-k\Lambda^*(x))dx.$$

As for large k the distribution of $k^{-1}\log G(W_k)$ and $G(W_k)/k$ are ever closer to having densities, using the change of variables formula gives

$$dP\left(\frac{1}{k}G(W_k) = x\right) = \frac{1}{kx}dP\left(\frac{1}{k}\log G(W_k) = x\right) \approx \frac{1}{kx}\exp\left(-k\Lambda^*\left(\frac{1}{k}\log(kx)\right)\right)dx.$$

Finally, the substitution $kx = n$ gives the approximation in equation (3.8). To make this heuristic precise requires distinct means, explained in the following corollary.

Corollary 3.1 (Direct estimates on guesswork) *Recall the definition*

$$K_k(x,\epsilon) := \left\{w \in \mathbb{A}^k : k^{-1}\log G(w) \in B_\epsilon(x)\right\}.$$

Under assumption 3.1, for any $x \in [0, \log m)$ we have

$$\begin{aligned}\lim_{\epsilon\downarrow 0}\liminf_{k\to\infty}\frac{1}{k}\log\inf_{w\in K_k(x,\epsilon)}P(W_k = w) &= \lim_{\epsilon\downarrow 0}\limsup_{k\to\infty}\frac{1}{k}\log\sup_{w\in K_k(x,\epsilon)}P(W_k = w)\\ &= -\left(x + \Lambda^*(x)\right).\end{aligned}$$

Proof: We show how to prove the upper bound as the lower bound follows using anal-

ogous arguments, as do the edge cases. Let $x \in (0, \log m)$ and $\epsilon > 0$ be given. Using the monotonicity of guesswork

$$\limsup_{k\to\infty} \frac{1}{k} \log \sup_{w\in K_k(x,\epsilon)} P(W_k = w) \leq \liminf_{k\to\infty} \frac{1}{k} \log \inf_{w\in K_k(x-2\epsilon,\epsilon)} P(W_k = w).$$

Using the estimate found in Theorem 3.3 and the LDP provides an upper bound on the latter:

$$\begin{aligned}
&(x - 3\epsilon) + \liminf_{k\to\infty} \frac{1}{k} \log \inf_{w\in K_k(x-2\epsilon,\epsilon)} P(W_k = w) \\
&\leq \liminf_{k\to\infty} \frac{1}{k} \log P\left(\frac{1}{k}\log(G(W_k)) \in B_\epsilon(x - 2\epsilon)\right) \\
&\leq \limsup_{k\to\infty} \frac{1}{k} \log P\left(\frac{1}{k}\log(G(W_k)) \in [x - 3\epsilon, x - \epsilon]\right) \\
&\leq - \inf_{x\in[x-3\epsilon,x-\epsilon]} \Lambda^*(x).
\end{aligned}$$

Thus

$$\limsup_{k\to\infty} \frac{1}{k} \log \sup_{w\in K_k(x,\epsilon)} P(W_k = w) \leq -x + 3\epsilon - \inf_{x\in[x-3\epsilon,x-\epsilon]} \Lambda^*(x).$$

As Λ^* is convex, it is continuous where finite, and thus the upper-bound follows taking $\epsilon \downarrow 0$.

■

Unpeeling limits, this corollary shows that when k is large the probability of the n^{th} guess being correct is approximately $1/n \exp(-k\Lambda^*(k^{-1} \log n))$, without the need to identify the string itself. This justifies the approximation (3.8), whose complexity of evaluation does not depend on k. We demonstrate its merit by example in Section 3.3.

Before that, as a corollary to the LDP we find the following role for the specific Shannon entropy. Thus, although Massey established that for a given string length the Shannon entropy is only a lower bound on the guesswork, for growing string length the specific Shannon entropy determines the linear growth rate of the expectation of the logarithm of guesswork (c.f [3] and [49]).

Corollary 3.2 (Shannon entropy and guesswork) *Under assumption 3.1,*

$$\lim_{k\to\infty} \frac{1}{k} E(\log G(W_k)) = \lim_{k\to\infty} \frac{1}{k} H(W),$$

the specific Shannon entropy.

PROOF: As both $\Lambda(\alpha)$ and $\alpha R_k((1+\alpha)^{-1})$ are finite and differentiable in a neighborhood of 0, by [48, Theorem 25.7]

$$\Lambda'(0) = \lim_{k\to\infty} \frac{1}{k} \frac{d}{d\alpha} \alpha R_k((1+\alpha)^{-1})|_{\alpha=0} = \lim_{k\to\infty} \frac{1}{k} H(W).$$

Note that $\Lambda^*(x) = 0$ if and only if $x = \Lambda'(0) = \lim k^{-1} H(W)$. Thus the weak law then follows by concentration of measure (e.g. [32] Theorem 2.1 taking $f(x)$ as the identity function on $\log G(W_k)$ and B_k as the entire set in conjunction with the fact that $\lim_k k^{-1} \log P(k^{-1} \log B_\epsilon(H(W))) = 0$ for all $\epsilon > 0$).

■

This also provides proof analogous to that of Massey's result that the Shannon entropy only provides a lower bound on the average guesswork. By Jensen's inequality [14, Lemma 2.6.2],

$$\begin{aligned} \lim k^{-1} \log E(G(W_k)) &\geq \lim \frac{1}{k} E(\log G(W_k)) \\ \Lambda(1) &\geq \lim \frac{1}{k} H(W). \end{aligned}$$

3.3 Examples

These examples will allow us to clarify some of the properties of the rate function for the guesswork. They also allow us to explore properties of specific instances such as the appearance of the golden ratio as the number of asymptotically most likely strings.

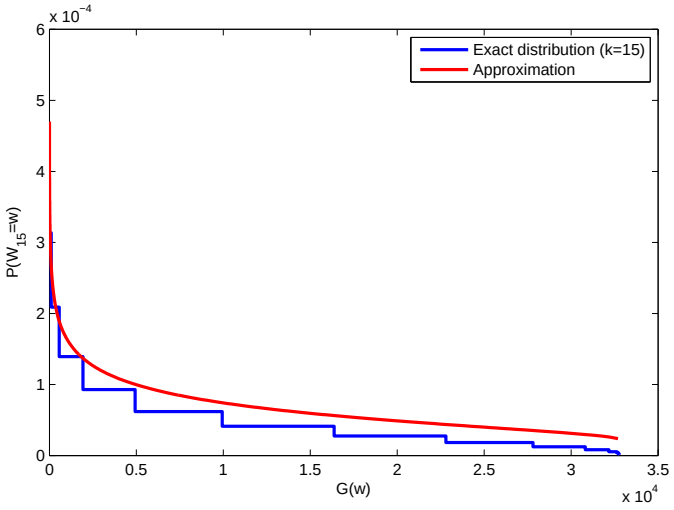

Figure 3.4: Illustration of Corollary 3.1. Strings constructed from i.i.d characters with $P(W_1 = 0) = 0.6, P(W_1 = 1) = 0.4$. For $k = 15$ comparison of the probability of n^{th} most likely string and the approximation $1/n \exp(-k\Lambda^*(k^{-1} \log n))$ versus $n \in \{1, \ldots, 3^{15}\}$.

3.3.1 I.i.d characters

.

Assume strings are constructed of i.i.d. characters. Let W_1 take values in $\mathbb{A} = \{1, \ldots, m\}$ and assume $P(W_1 = i) \geq P(W_1 = j)$ if $i \leq j$. Then from [1, 46] and Lemma 3.1 we have that

$$\Lambda(\alpha) = \begin{cases} (1+\alpha) \log \sum_{w \in \mathbb{A}} P(W_1 = w)^{1/(1+\alpha)} & \text{if } \alpha > -1 \\ \log P(W_1 = 1) & \text{if } \alpha \leq -1. \end{cases}$$

From Lemma 3.2 we have that

$$\gamma = \lim_{\alpha \downarrow -1} \Lambda'(\alpha) \in \{0, \log(2), \ldots, \log m\}$$

and no other values are possible. Unless the distribution of W_1 is uniform, $\Lambda^*(x)$ does not have a closed form for all x, but is readily calculated numerically. With $|\mathbb{A}| = 3$ and $k = 15$, Figure 3.4 compares the exact distribution $P(W_k = w)$ versus $G(w)$ with the approximation found in equation (3.8). As there are $3^{15} \approx 1.4$ million strings, the

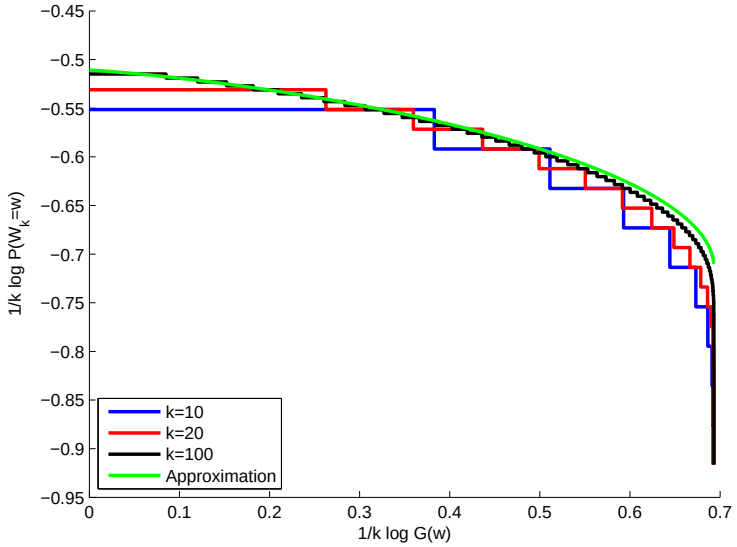

Figure 3.5: Illustration of Corollary 3.1. Strings constructed from i.i.d characters with $P(W_1 = 0) = 0.6, P(W_1 = 1) = 0.4$. For $k = 10, 20$ and 100, comparison of k^{-1} times the logarithm of the probability of n^{th} most likely string versus k^{-1} times the logarithm of n, as well as the approximation $-x - \Lambda^*(x)$ versus x.

likelihood of any one string is tiny, but the quality of the approximation can clearly be seen. Rescaling the guesswork and probabilities to make them comparable for distinct k, Figure 3.5 illustrates the quality of the approximation as k grows. By $k = 100$ there are $3^{100} \approx 5.1$ times 10^{47} strings and the underlying combinatorial complexities of the explicit calculation become immense, yet the complexity of calculating the approximation has not increased.

3.3.2 The Golden Ratio

The golden ratio arises in many different areas of mathematics. For guesswork, it enters via the quantity γ, defined in equation (3.2), and binary Markov sources.

As an example of strings constructed of correlated characters, consider $\{W_k\}$ where the characters are chosen via a process a Markov chain with transition matrix P and some initial distribution on $|\mathbb{A}| = 2$. Define the matrix P_α by $(P_\alpha)_{i,j} = p_{i,j}^{1/(1+\alpha)}$, then by

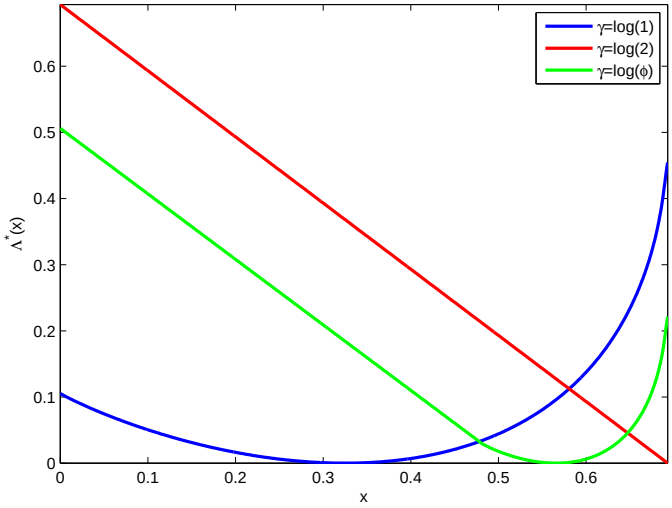

Figure 3.6: Illustration of rate functions in Theorem 3.3. Strings constructed from Markov characters on $|\mathbb{A}| = 2$. Three rate functions illustrating only values of γ possible, $\log(1)$, $\log(\phi) \approx 0.48$ and $\log(2)$, from Lemma 3.5.

[37, 46] and Lemma 3.1 we have that

$$\Lambda(\alpha) = \begin{cases} (1+\alpha)\log\rho(P_\alpha) & \text{if } \alpha > -1 \\ \log\max(p_{0,0}, p_{1,1}, \sqrt{p_{0,1}\,p_{1,0}}) & \text{if } \alpha \leq -1, \end{cases}$$

where ρ is the spectral radius operator. In the two letter alphabet case, with $\beta = 1/(1+\alpha)$ we have that $\rho(P_{(1-\beta)/\beta})$ equals

$$\frac{p_{0,0}^\beta + p_{1,1}^\beta}{2} + \frac{\sqrt{(p_{0,0}^\beta - p_{1,1}^\beta)^2 + 4(1-p_{1,1})^\beta(1-p_{0,0})^\beta}}{2}.$$

As with the i.i.d. characters example, apart from in special cases the rate function $\Lambda^*(\cdot)$ cannot be calculated in closed form, but is readily evaluated numerically. Regardless, we have the following, perhaps surprising, result on the exponential rate of growth of the size of the set of almost most likely strings.

Lemma 3.5 (The Golden Ratio and Markovian characters) *For $\{W_k\}$ constructed of Markovian characters with $|\mathbb{A}| = 2$,*

$$\gamma = \lim_{\alpha \downarrow -1} \Lambda'(\alpha) \in \{0, \log(\phi), \log(2)\},$$

where $\phi = (1+\sqrt{5})/2$ *is the Golden Ratio, and no other values are possible.*

This lemma can be proved by directly evaluating the derivative of $\Lambda(\alpha)$ with respect to α. Note that here $\exp(k\gamma)$ definitely only describes the number of strings of equal highest likelihood when k is large as the initial distribution of the Markov chain plays no rôle in γ's evaluation.

The case where $\gamma = \log(2)$ occurs when $p_{0,0} = p_{1,1} = 1/2$. The most interesting case is when there are approximately ϕ^k approximately equally most likely strings. This occurs if $p_{0,0} = \sqrt{p_{0,1}p_{1,0}} > p_{1,1}$. For large k, strings of near-maximal probability have the form of a sequence of 0s, where a 1 can be inserted anywhere so long as there is a 0 between it and any other 1s. A further sub-exponential number of aberrations are allowed in any given sequence and the starting distribution is ultimately irrelevant. For example, with an equiprobable initial distribution and $k = 4$ there are 8 most likely strings (0000, 0001, 0010, 0100, 0101, 1000, 1010, 1001) and $\phi^4 \approx 6.86$. Note that the golden ratio also appears in the analysis of the trapdoor channel [45], but there it is directly as a result of the appearance of the Fibonacci sequence. The case of $\gamma = \log(1)$ occurs if we only have one or two most likely strings. So that one of $p_{0,0} \neq \sqrt{p_{0,1}p_{1,0}}$ and $p_{1,1} \neq \sqrt{p_{0,1}p_{1,0}}$.

Figure 3.6 gives plots of $\Lambda^*(x)$ versus x illustrating the full range of possible shapes that rate functions can take: linear, linear then strictly convex, or strictly convex, based on the transition matrices

$$\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \begin{pmatrix} 0.6 & 0.4 \\ 0.9 & 0.1 \end{pmatrix} \text{ and } \begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{pmatrix}$$

respectively.

4 Guesswork, the Asymptotic Equipartition Property and Typical Sets

4.1 Introduction

Consider the problem of identifying the value of a discrete random variable by only asking questions of the sort: is its value X? That this is a time-consuming task is a cornerstone of computationally secure ciphers [41]. In computational security it is tempting to appeal to the Asymptotic Equipartition Property (AEP) [14], and the resulting assignment of code words only to elements of the Typical Set of the source, to justify restriction to consideration of a uniform source, e.g. [47, 19, 51]. This assumed uniformity has many desirable properties, including maximum obfustication and difficulty for the inquisitor, e.g. [20].

In Typical Set coding it is necessary to generate codes for strings whose logarithmic probability is within a small distance of the string length times the specific Shannon entropy. As a result, while all these strings have near-equal likelihood, the distribution is not precisely uniform. It is the consequence of this lack of perfect uniformity that we investigate here by proving that results on Guesswork mentioned in earlier chapters extend to this setting. The results in this chapter establish that for a variety of sources, as a function of string length, it is exponentially easier to guess strings conditioned to be in the source's Typical Set in comparison to the corresponding equipartition approximation. This suggests that appealing to the AEP to justify sole consideration of the uniform distributions for cryptanalysis is ill-advised and provides alternate results in their place.

4.2 The Typical Set and Guesswork

Let $\mathbb{A} = \{0, \ldots, m-1\}$ be a finite alphabet and consider a stochastic sequence of words, $\{W_k\}$, where W_k is a word of length k taking values in $\mathbb{A}^k$. The process $\{W_k\}$ has specific Shannon entropy

$$H_W := -\lim_{k\to\infty} \frac{1}{k} \sum_{w\in\mathbb{A}^k} P(W_k = w) \log P(W_k = w),$$

and we shall take all logs to base e. For $\epsilon > 0$, the Typical Set of strings of length k is

$$T_k^\epsilon := \left\{ w \in \mathbb{A}^k : e^{-k(H(W)+\epsilon)} \leq P(W_k = w) \leq e^{-k(H(W)-\epsilon)} \right\}. \tag{4.1}$$

For most reasonable sources [14], $P(W_k \in T_k^\epsilon) > 0$ for all k sufficiently large and Typical Set encoding results in a new source of strings of length k, W_k^ϵ, with statistics

$$P(W_k^\epsilon = w) = \begin{cases} \dfrac{P(W_k = w)}{P(W_k \in T_k^\epsilon)} & \text{if } w \in T_k^\epsilon, \\ 0 & \text{if } w \notin T_k^\epsilon. \end{cases} \tag{4.2}$$

Appealing to the AEP, these distributions are often substituted for their more readily manipulated uniformly random counterpart, U_k^ϵ,

$$P(U_k^\epsilon = w) := \begin{cases} \dfrac{1}{|T_k^\epsilon|} & \text{if } w \in T_k^\epsilon, \\ 0 & \text{if } w \notin T_k^\epsilon, \end{cases} \tag{4.3}$$

where $|T_k^\epsilon|$ is the number of elements in T_k^ϵ. While the distribution of W_k^ϵ is near-uniform for large k, it is not perfectly uniform unless the original W_k was uniformly distributed on a subset of $\mathbb{A}^k$. Is a string selected using the distribution of W_k^ϵ easier to guess than if the string was selected uniformly, U_k^ϵ?

For fixed k it is shown in [39] that the Shannon entropy of the underlying distribution bears little relation to the expected guesswork, $E(G(W_k))$, the average number of guesses required to guess a word chosen with distribution W_k using the optimal strategy. In a series of subsequent papers [1, 37, 46, 25], under ever less restrictive stochastic assumptions from words made up of i.i.d. letters to Markovian letters to sofic shifts, an asymptotic relationship as word length grows between scaled moments of the guesswork and specific Rényi entropy was identified:

$$\lim_{k\to\infty} \frac{1}{k} \log E(G(W_k)^\alpha) = \alpha R_W\left(\frac{1}{1+\alpha}\right), \tag{4.4}$$

for $\alpha > -1$, where $R_W(\beta)$ is the specific Rényi entropy for the process $\{W_k\}$ with parameter $\beta > 0$,

$$R_W(\beta) := \lim_{k\to\infty} \frac{1}{k} \frac{1}{1-\beta} \log\left(\sum_{w\in\mathbb{A}^k} P(W_k = w)^\beta\right).$$

In Chapter 3 we build built on those results to prove that $\{k^{-1}\log G(W_k)\}$ satisfies a Large Deviation Principle (LDP), e.g [16]. Define the scaled Cumulant Generating

Function (sCGF) of $\{k^{-1}\log G(W_k)\}$ by

$$\Lambda_W(\alpha) := \lim_{k\to\infty} \frac{1}{k}\log E\left(e^{\alpha \log G(W_k)}\right) \text{ for } \alpha \in \mathbb{R}$$

and make the following assumption.

Assumption 4.1 *For $\alpha > -1$, the sCGF $\Lambda_W(\alpha)$ exists, is equal to $\alpha R_W\left(1/(1+\alpha)\right)$ and has a continuous derivative in that range.*

Should Assumption 4.1 hold, Theorem 3.3 establishes that $\Lambda_W(\alpha) = \lim_{\alpha\downarrow -1} R_W(\alpha/(1+\alpha))$ for all $\alpha \leq -1$ and that the sequence $\{k^{-1}\log G(W_k)\}$ satisfies a LDP with a rate function given by the Legendre Fenchel transform of the sCGF, $\Lambda_W^*(x) := \sup_{\alpha\in\mathbb{R}}\{x\alpha - \Lambda_W(\alpha)\}$. Assumption 1 is motivated by equation (4.4). With

$$\gamma_W := \lim_{\alpha\downarrow -1} \frac{d}{d\alpha}\Lambda_W(\alpha), \tag{4.5}$$

where the order of the size of the set of maximum probability words of W_k is $\exp(k\gamma_W)$ [10], $\Lambda_W^*(x)$ can be identified as

$$= \begin{cases} -x - \lim_{\alpha\downarrow -1} R_W(\alpha/(1+\alpha)) & \text{if } x \in [0, \gamma_W] \\ \sup_{\alpha\in\mathbb{R}}\{x\alpha - \Lambda_W(\alpha)\} & \text{if } x \in (\gamma_W, \log(m)], \\ +\infty & \text{if } x \notin [0, \log(m)]. \end{cases}$$

Corollary 3.2 of Chapter 3 uses this LDP to prove a result suggested in [3, 49], that

$$\lim_{k\to\infty} \frac{1}{k} E(\log(G(W_k))) = H_W, \tag{4.6}$$

making clear that the specific Shannon entropy determines the expectation of the logarithm of the number of guesses to guess the word W_k. The growth rate of the expected guesswork is a distinct quantity whose scaling rules can be determined directly from the sCGF in equation (4.4),

$$\lim_{k\to\infty} \frac{1}{k}\log E(G(W_k)) = \Lambda_W(1).$$

From these expressions and Jensen's inequality, it is clear that the growth rate of the expected guesswork is more than H_W. Finally, as a corollary to the LDP, Chapter 3

provides the following approximation to the guesswork distribution for large k:

$$P(G(W_k) = n) \approx \frac{1}{n} \exp\left(-k\Lambda_W^*(k^{-1}\log n)\right)$$

for $n \in \{1, \ldots, m^k\}$. Thus to approximate the guesswork distribution, it is sufficient to know the specific Rényi entropy of the source and the decay-rate of the likelihood of the sequence of most likely words.

The rest of this chapter is split into three parts.

- The first establishes that if $\{W_k\}$ is constructed from i.i.d. letters, then both of the processes $\{U_k^\epsilon\}$ and $\{W_k^\epsilon\}$ also satisfy Assumption 4.1 so that, with the appropriate rate functions, the approximation in equation (3.8) can be used with U_k^ϵ or W_k^ϵ in lieu of W_k. This enables us to compare the guesswork distribution for Typical Set encoded words with their assumed uniform counterpart. Even in the simple binary alphabet case we establish that, apart from edge cases, a word chosen via W_k^ϵ is exponential easier in k to guess on average than one chosen via U_k^ϵ.

- The second part, found in Section 4.4, provides an example to illustrate those results.

- The final part, beginning in Section 4.5, generalises the source assumptions to the setting of Pfister and Sullivan [46], establishing the robustness of the i.i.d. deductions.

4.3 Statement of main i.i.d. results

Assume that the strings $\{W_k\}$ are made of i.i.d. characters, defining $p = (p_0, \ldots, p_{m-1})$ by $p_a = P(W_1 = a)$. We shall employ the following short-hand: $h(l) := -\sum_a l_a \log l_a$ for $l = (l_0, \ldots, l_{m-1}) \in [0,1]^m$, $l_a \geq 0$, $\sum_a l_a = 1$, so that $H(W) = h(p)$, and $D(l\|p) := -\sum_a l_a \log(p_a/l_a)$, the KL-Divergence of the source. Furthermore, define $l^- \in [0,1]^m$ and $l^+ \in [0,1]^m$ by

$$l^- \in \arg\max_l \{h(l) : h(l) + D(l\|p) - \epsilon = h(p)\}, \tag{4.7}$$

$$l^+ \in \arg\max_l \{h(l) : h(l) + D(l\|p) + \epsilon = h(p)\}, \tag{4.8}$$

should they exist. For $\alpha > -1$, define $l_W(\alpha)$ as $(l_{W0}(\alpha), \ldots, l_{Wm-1}(\alpha))$, with $l_{Wa}(\alpha)$ and $\eta(\alpha)$ are defined by

$$l_{Wa}(\alpha) := \frac{p_a^{(1/(1+\alpha))}}{\sum_{b\in\mathbb{A}} p_b^{(1/(1+\alpha))}} \text{ for all } a \in \mathbb{A} \text{ and} \tag{4.9}$$

$$\eta(\alpha) := -\sum_a l_{Wa}(\alpha) \log p_a = -\frac{\sum_{a\in\mathbb{A}} p_a^{1/(1+\alpha)} \log p_a}{\sum_{b\in\mathbb{A}} p_b^{1/(1+\alpha)}}. \tag{4.10}$$

Assume that $h(p) + \epsilon \leq \log(m)$. If this is not the case, $\log(m)$ should be substituted in place of $h(l^-)$ for the $\{U_k^\epsilon\}$ results.

Lemma 4.1 *Assumption 4.1 holds for* $\{U_k^\epsilon\}$ *and* $\{W_k^\epsilon\}$ *with*

$$\Lambda_{U^\epsilon}(\alpha) := \alpha h(l^-)$$

and

$$\Lambda_{W^\epsilon}(\alpha) = \alpha h(l^*(\alpha)) - D(l^*(\alpha)\|p),$$

where

$$l^*(\alpha) = \begin{cases} l^+ & \text{if } \alpha > -1,\ \eta(\alpha) < h(p) - \epsilon, \\ l_W(\alpha) & \text{if } \eta(\alpha) \in [h(p) - \epsilon, h(p) + \epsilon], \\ l^- & \text{if } \eta(\alpha) > h(p) + \epsilon. \end{cases} \tag{4.11}$$

Thus by direct evaluation of the sCGFs at $\alpha = 1$,

$$\lim_{k\to\infty} \frac{1}{k} \log E(G(U_k^\epsilon)) = h(l^-) \text{ and } \lim_{k\to\infty} \frac{1}{k} \log E(G(W_k^\epsilon)) = \Lambda_{W^\epsilon}(1).$$

As the conditions of Theorem 3.3 are satisfied

$$\lim_{k\to\infty} \frac{1}{k} E(\log(G(U_k^\epsilon)) = \Lambda'_{U^\epsilon}(0) = h(l^-) \text{ and } \lim_{k\to\infty} \frac{1}{k} E(\log(G(W_k^\epsilon)) \quad = \Lambda'_{W^\epsilon}(0) = h(p),$$

and we have the approximations

$$P(G(U_k^\epsilon) = n) \approx \frac{1}{n} \exp\left(-k\Lambda^*_{U^\epsilon}(k^{-1}\log n)\right) \text{ and}$$
$$P(G(W_k^\epsilon) = n) \approx \frac{1}{n} \exp\left(-k\Lambda^*_{W^\epsilon}(k^{-1}\log n)\right).$$

The proof of Lemma 4.1 is deferred until after some preliminary results. Note that by the definition of T_k^ϵ as a Typical Set, $P(W_k \in T_k^\epsilon) > 1 - \epsilon$ for all k sufficiently large and thus

$$\lim_{k\to\infty} \frac{1}{k} \log P(W_k \in T_k^\epsilon) = 0.$$

The proportion of the character $a \in \mathbb{A}$ in a string $w = (w_1, \ldots, w_k) \in \mathbb{A}^k$ is given by

$$n_k(w, a) := \frac{|\{1 \le i \le k : w_i = a\}|}{k}.$$

The number of strings in a type l, where $l_a \in [0, 1]$ for all $a \in \mathbb{A}$ and $\sum_{a\in\mathbb{A}} l_a = 1$, is given by

$$N_k(l) := |\{w \in \mathbb{A}^k \text{ such that } n_k(w, a) = l_a \ \forall a \in \mathbb{A}\}|.$$

The set of all types, those just in the Typical Set and smooth approximations to those in the Typical Set are denoted

$$L_k := \{l : \exists w \in \mathbb{A}^k \text{ such that } n_k(w, a) = l_a \ \forall a \in \mathbb{A}\},$$
$$L_{\epsilon,k} := \{l : \exists w \in T_k^\epsilon \text{ such that } n_k(w, a) = l_a \ \forall a \in \mathbb{A}\},$$
$$L_\epsilon := \left\{l : \sum_a l_a \log p_a \in [-h(p) - \epsilon, -h(p) + \epsilon]\right\},$$

where it can readily seen that $L_{\epsilon,k} \subset L_\epsilon$ for all k.

For $\{U_k^\epsilon\}$ we need the following Lemma.

Lemma 4.2 *The exponential growth rate of the size of the Typical Set is*

$$\lim_{k\to\infty}\frac{1}{k}\log|T_k^\epsilon| = \begin{cases} \log m \text{ if } \log m \le h(p)+\epsilon \\ h(l^-) \text{ otherwise.} \end{cases}$$

where l^- is defined in equation (4.7).

Proof: For fixed k, by the union bound

$$\max_{l\in L_{\epsilon,k}}\frac{k!}{\prod_{a\in\mathbb{A}}(kl_a)!} \le |T_k^\epsilon| \le (k+1)^m \max_{l\in L_{\epsilon,k}}\frac{k!}{\prod_{a\in\mathbb{A}}(kl_a)!}.$$

For the logarithmic limit, these two bounds coincide so consider the concave optimization

$$\max_{l\in L_{\epsilon,k}}\frac{k!}{\prod_{a\in\mathbb{A}}(kl_a)!}.$$

We can upper bound this optimization by replacing $L_{\epsilon,k}$ with the smoother version, its superset L_ϵ. Using Stirling's bound we have that

$$\begin{aligned}&\limsup_{k\to\infty}\frac{1}{k}\log\sup_{l\in L_\epsilon}\frac{k!}{\prod_{a\in\mathbb{A}}(kl_a)!} \\ &\qquad\le \sup_{l\in L_\epsilon} h(l) = \begin{cases}\log(m) & \text{if } h(p)+\epsilon \ge \log(m) \\ h(l^-) & \text{if } h(p)+\epsilon < \log(m).\end{cases}\end{aligned}$$

For the lower bound, we need to construct a sequence $\{l^{(k)}\}$ such that $l^{(k)} \in L_{\epsilon,k}$ for all k sufficiently large and $h(l^{(k)})$ converges to either $\log(m)$ or $h(l^-)$, as appropriate. Let $l^* = (1/m,\dots,1/m)$ or l^- respectively, letting $c \in \arg\max p_a$ and define

$$l_a^{(k)} = \begin{cases} k^{-1}\lfloor kl^*{}_a\rfloor + 1 - \sum_{b\in\mathbb{A}}\frac{1}{k}\lfloor kl^*{}_b\rfloor & \text{if } a = c, \\ k^{-1}\lfloor kl^*{}_a\rfloor & \text{if } a \ne c. \end{cases}$$

Then $l^{(k)} \in L_{\epsilon,k}$ for all $k > -m\log(p_c)/(2\epsilon)$ and $h(l^{(k)}) \to h(l^*)$, as required.

■

PROOF: Proof of Lemma 4.1. Considering $\{U_k^\epsilon\}$ first,

$$\alpha R_{U^\epsilon}\left(\frac{1}{1+\alpha}\right) = \alpha \lim_{k\to\infty} \frac{1}{k} \log |T_k^\epsilon| = \alpha h(l^-),$$

by Lemma 4.2. To evaluate $\Lambda_{U^\epsilon}(\alpha)$, using that for $n \in \mathbb{N}$ and $\alpha > 0$

$$\sum_{i=1}^{n} i^\alpha \geq \int_0^n x^\alpha dx,$$

we use Lemma 4.2 again and we have

$$\begin{aligned}\alpha h(l^-) &= \lim_{k\to\infty} \frac{1}{k} \log \frac{1}{1+\alpha} |T_k^\epsilon|^\alpha \leq \lim_{k\to\infty} \frac{1}{k} \log E(e^{\alpha \log G(U_k^\epsilon)}) \\ &= \lim_{k\to\infty} \frac{1}{k} \log \frac{1}{|T_k^\epsilon|} \sum_{i=1}^{|T_k^\epsilon|} i^\alpha \leq \lim_{k\to\infty} \frac{1}{k} \log |T_k^\epsilon|^\alpha = \alpha h(l^-).\end{aligned}$$

The reverse of these bounds holds for $\alpha \in (-1, 0]$, giving the result.

We break the argument for $\{W_k^\epsilon\}$ into three steps. Step 1 is to show the equivalence of the existence of $\Lambda_{W^\epsilon}(\alpha)$ and $\alpha R_{W^\epsilon}(1/(1+\alpha))$ for $\alpha > -1$ with the existence of the following limit

$$\lim_{k\to\infty} \frac{1}{k} \log \max_{l \in L_{\epsilon,k}} \left\{ N_k(l)^{1+\alpha} \prod_{a\in\mathbb{A}} p_a^{kl_a} \right\}. \tag{4.12}$$

Step 2 then establishes this limit and identifies it. Step 3 shows that $\Lambda'_{W^\epsilon}(\alpha)$ is continuous for $\alpha > -1$. To achieve steps 1 and 2, we adopt and adapt the method of types argument employed in the elongated web-version of [37].

Step 1: Two changes from the bounds of [37, Lemma 5.5] are necessary: the consideration of non-i.i.d. sources by restriction to T_k^ϵ; and the extension of the α range to include $\alpha \in (-1, 0]$ from that for $\alpha \geq 0$ given in that document. Adjusted for conditioning on

the Typical Set we get

$$\frac{1}{1+\alpha}\max_{l\in L_{\epsilon,k}}\left\{N_k(l)^{1+\alpha}\frac{\prod_{a\in\mathbb{A}}p_a^{kl_a}}{\sum_{w\in T_k^\epsilon}P(W_k=w)}\right\}\leq E(e^{\alpha\log G(W_k^\epsilon)})\leq \tag{4.13}$$
$$(k+1)^{m(1+\alpha)}\max_{l\in L_{\epsilon,k}}\left\{N_k(l)^{1+\alpha}\frac{\prod_{a\in\mathbb{A}}p_a^{kl_a}}{\sum_{w\in T_k^\epsilon}P(W_k=w)}\right\}.$$

The necessary modification of these inequalities for $\alpha\in(-1,0]$ gives

$$\max_{l\in L_{\epsilon,k}}\left\{N_k(l)^{1+\alpha}\frac{\prod_{a\in\mathbb{A}}p_a^{kl_a}}{\sum_{w\in T_k^\epsilon}P(W_k=w)}\right\}\leq E(e^{\alpha\log G(W_k^\epsilon)})\leq \tag{4.14}$$
$$\frac{(k+1)^m}{1+\alpha}\max_{l\in L_{\epsilon,k}}\left\{N_k(l)^{1+\alpha}\frac{\prod_{a\in\mathbb{A}}p_a^{kl_a}}{\sum_{w\in T_k^\epsilon}P(W_k=w)}\right\}.$$

To show the lower bound holds if $\alpha\in(-1,0]$ let

$$l^*\in\arg\max_{l\in L_{\epsilon,k}}\left\{N_k(l)^{1+\alpha}\frac{\prod_{a\in\mathbb{A}}p_a^{kl_a}}{\sum_{w\in T_k^\epsilon}P(W_k=w)}\right\}.$$

Taking $\liminf_{k\to\infty}k^{-1}\log$ and $\limsup_{k\to\infty}k^{-1}\log$ of equations (4.13) and (4.14) establishes that if the limit (4.12) exists, $\Lambda_{W^\epsilon}(\alpha)$ exists and equals it. For the Rényi entropy see that

$$\begin{aligned}
&\lim_{k\to\infty}\frac{1+\alpha}{k}\log\max_{l\in L_{\epsilon,k}}\left\{N_k(l)\left(\frac{\prod_{a\in\mathbb{A}}p_a^{kl_a}}{\sum_{w\in T_k^\epsilon}P(W_k=w)}\right)^{1/(1+\alpha)}\right\}\\
&\leq\lim_{k\to\infty}\frac{(1+\alpha)}{k}\log\sum_{w\in\mathbb{A}^k}P(W_k=w)^{1/(1+\alpha)}\\
&\leq\lim_{k\to\infty}\frac{1+\alpha}{k}\log(k+1)^m\max_{l\in L_{\epsilon,k}}\left\{N_k(l)\left(\frac{\prod_{a\in\mathbb{A}}p_a^{kl_a})}{\sum_{w\in T_k^\epsilon}P(W_k=w)}\right)^{1/(1+\alpha)}\right\}.
\end{aligned}$$

Here the first inequality follows by taking only the maximal type from the sum and the second by taking each type to have the same value as the maximal type the function above. Then take limits to obtain the desired result.

Step 2: The problem has been reduced to establishing the existence of

$$\lim_{k\to\infty} \frac{1}{k} \log \max_{l\in L_{\epsilon,k}} \left\{ N_k(l)^{1+\alpha} \prod_{a\in\mathbb{A}} p_a^{kl_a} \right\}$$

and identifying it. The method of proof is similar to that employed in Lemma 4.1: we provide an upper bound for the limsup and then establish a corresponding lower bound.

If $l^{(k)} \to l$ with $l^{(k)} \in L_k$, then using Stirling's bounds we have that

$$\lim_{k\to\infty} \frac{1}{k} \log N_k(l^{(k)}) = h(l).$$

This convergence occurs uniformly in l and so, as $L_{\epsilon,k} \subset L_\epsilon$ for all k,

$$\begin{aligned} \limsup_{k\to\infty} \frac{1}{k} \log \max_{l\in L_{\epsilon,k}} \left\{ N_k(l)^{1+\alpha} \prod_{a\in\mathbb{A}} p_a^{kl_a} \right\} &\le \sup_{l\in L_\epsilon} \left((1+\alpha)h(l) + \sum_a l_a \log p_a \right) \\ &= \sup_{l\in L_\epsilon} \left(\alpha h(l) - D(l\|p) \right). \end{aligned} \tag{4.15}$$

This is a concave optimization problem in l with convex constraints. Not requiring $l \in L_\epsilon$, the unconstrained optimizer over all l is attained at $l_W(\alpha)$ defined in equation (4.9), which determines $\eta(\alpha)$ in equation (4.10). Thus the optimizer of the constrained problem (4.15) can be identified as that given in equation (4.11). Thus we have that

$$\limsup_{k\to\infty} \frac{1}{k} \log \max_{l\in L_{\epsilon,k}} \left\{ N_k(l)^{1+\alpha} \prod_{a\in\mathbb{A}} p_a^{kl_a} \right\} \le \alpha h(l^*(\alpha)) + D(l^*(\alpha)\|p),$$

where $l^*(\alpha)$ is defined in equation (4.11).

We complete the proof by generating a matching lower bound. To do so, for given $l^*(\alpha)$ we need only create a sequence such that $l^{(k)} \to l^*(\alpha)$ and $l^{(k)} \in L_{\epsilon,k}$ for all k. If $l^*(\alpha) = l^-$, then the sequence used in the proof of Lemma 4.2 suffices. For $l^*(\alpha) = l^+$, we use the same sequence but with floors in lieu of ceilings and the surplus probability distributed to a least likely character instead of a most likely character. For $l^*(\alpha) = l_W(\alpha)$, either of these sequences can be used.

Step 3: As $\Lambda_{W^\epsilon}(\alpha) = \alpha h(l^*(\alpha)) - D(l^*(\alpha)\|p)$, with $l^*(\alpha)$ defined in equation (4.11),

$$\frac{d}{d\alpha}\Lambda_{W^\epsilon}(\alpha) = h(l^*(\alpha)) + \Lambda_{W^\epsilon}(\alpha)\frac{d}{d\alpha}l^*(\alpha).$$

Thus to establish continuity it suffices to establish continuity of $l^*(\alpha)$ and its derivative, which can be done readily by calculus.

■

4.4 Example

Consider a binary alphabet $\mathbb{A} = \{0, 1\}$ and strings $\{W_k\}$ constructed of i.i.d. characters with $P(W_1 = 0) = p_0 > 1/2$. In this case there are unique l^- and l^+ satisfying equations (4.7) and (4.8) determined by:

$$\begin{aligned} l_0^- &= p_0 - \frac{\epsilon}{\log(p_0) - \log(1-p_0)}, \\ l_0^+ &= p_0 + \frac{\epsilon}{\log(p_0) - \log(1-p_0)}. \end{aligned}$$

Selecting $0 < \epsilon < (\log(p_0) - \log(1-p_0)) \min(p_0 - 1/2, 1 - p_0)$ ensures that the Typical Set is growing more slowly than 2^k and that $1/2 < l_0^- < p_0 < l_0^+ < 1$.

With $l_W(\alpha)$ defined in equation (4.9), we have that

$$\begin{aligned} \Lambda_W(\alpha) &= \begin{cases} \log(p_0) & \text{if } \alpha < -1, \\ \alpha h(l_W(\alpha)) - D(l_W(\alpha)\|p), & \text{if } \alpha \geq -1. \end{cases} \\ &= \begin{cases} \log(p_0) & \text{if } \alpha < -1, \\ (1+\alpha)\log\left(p_0^{\frac{1}{1+\alpha}} + (1-p_0)^{\frac{1}{1+\alpha}}\right) & \text{if } \alpha \geq -1, \end{cases} \end{aligned}$$

From Lemma 4.1 we obtain

$$\Lambda_{U^\epsilon}(\alpha) = \begin{cases} -h(l^-) & \text{if } \alpha < -1, \\ \alpha h(l^-) & \text{if } \alpha \geq -1, \end{cases}$$

and

$$\Lambda_{W^\epsilon}(\alpha) = \alpha h(l^*(\alpha)) - D(l^*(\alpha)\|p),$$

where $l^*(\alpha)$ is defined in equation (4.11) and $\eta(\alpha)$ defined in equation (4.10).

With γ defined in equation (4.5), we have $\gamma_W = 0$, $\gamma_{U^\epsilon} = h(l^-)$ and $\gamma_{W^\epsilon} = h(l^+)$ so that, as $h(l^-) > h(l^+)$, the ordering of the growth rates with string length of the set of most likely strings from smallest to largest is: unconditioned source, conditioned source and uniform approximation.

From these sCGF equations, we can determine the average growth rates and estimates on the guesswork distribution. In particular, we have that

$$\begin{aligned}
\lim_{k\to\infty} \frac{1}{k} E(\log(G(W_k))) &= \Lambda'_W(0) = h(p), \\
\lim_{k\to\infty} \frac{1}{k} E(\log(G(W_k^\epsilon))) &= \Lambda'_{W^\epsilon}(0) = h(p), \\
\lim_{k\to\infty} \frac{1}{k} E(\log(G(U_k^\epsilon))) &= \Lambda'_{U^\epsilon}(0) = h(l^-).
\end{aligned}$$

As $h(x)$ is monotonically decreasing for $x > 1/2$ and $1/2 < l_0^- < p_0$, the expectation of the logarithm of the guesswork is growing faster for the uniform approximation than for either the unconditioned or conditioned string source. The growth rate of the expected guesswork reveals more features. In particular, with $A = \eta(1) - (h(p) + \epsilon)$,

$$\begin{aligned}
\lim_{k\to\infty} \frac{1}{k} \log E(G(W_k)) &= 2\log(p_0^{\frac{1}{2}} + (1-p_0)^{\frac{1}{2}}), \\
\lim_{k\to\infty} \frac{1}{k} \log E(G(W_k^\epsilon)) &= \begin{cases} 2\log(p_0^{\frac{1}{2}} + (1-p_0)^{\frac{1}{2}}), A \le 0 \\ h(l^-) - D(l^-\|p), A > 0 \end{cases} \\
\lim_{k\to\infty} \frac{1}{k} \log E(G(U_k^\epsilon)) &= h(l^-).
\end{aligned}$$

For the growth rate of the expected guesswork, from these it can be shown that there is no strict order between the unconditioned and uniform source, but there is a strict ordering between the the uniform approximation and the true conditioned distribution, with the former being strictly larger.

With $\epsilon = 1/10$ and for a range of p_0, these formulae are illustrated in Figure 4.1. The

top line plots

$$\lim_{k\to\infty} \frac{1}{k} E(\log(G(U_k^\epsilon)) - \log(G(W_k)))$$
$$= \lim_{k\to\infty} \frac{1}{k} E(\log(G(U_k^\epsilon)) - \log(G(W_k^\epsilon))) = h(l^-) - h(p),$$

showing that the expected growth rate in the logarithm of the guesswork is always higher for the uniform approximation than both the conditioned and unconditioned sources. The second highest line in Figure 4.1 plots the difference in growth rates of the expected guesswork of the uniform approximation and the true conditioned source

$$\lim_{k\to\infty} \frac{1}{k} \log \frac{E(G(U_k^\epsilon))}{E(G(W_k^\epsilon))} = \begin{cases} h(l^-) - 2\log(p_0^{\frac{1}{2}} + (1-p_0)^{\frac{1}{2}}) & \text{if } \eta(1) \le h(p) + \epsilon \\ D(l^-\|p) & \text{if } \eta(1) > h(p) + \epsilon. \end{cases}$$

That this difference is always positive, which can be readily established analytically, shows that the expected guesswork of the true conditioned source is growing at a slower exponential rate than the uniform approximation. The second line in Figure 4.1 and the lowest in Figure 4.1 line, the growth rates of the uniform and unconditioned expected guesswork

$$\lim_{k\to\infty} \frac{1}{k} \log \frac{E(G(U_k^\epsilon))}{E(G(W_k))} = h(l^-) - 2\log(p_0^{\frac{1}{2}} + (1-p_0)^{\frac{1}{2}}),$$

initially agree. It can, depending on p_0 and ϵ, be either positive or negative. It is negative if the Typical Set is particularly small in comparison to the number of unconditioned strings.

For $p_0 = 8/10$, the Typical Set is growing sufficiently quickly that a string selected from the uniform approximation is easier to guess than for unconditioned source. For this value of p, we illustrate the difference in guesswork distributions between the unconditioned, $\{W_k\}$, conditioned, $\{W_k^\epsilon\}$, and uniform, $\{U_k^\epsilon\}$, string sources. If we used the approximation in Chapter 3, (3.8) directly, the graph would not be informative as the range of the unconditioned source is growing exponentially faster than the other two. Instead Figure 4.2 plots $-x - \Lambda^*(x)$ for each of the three processes. That is, using equation (3.8) and its equivalents for the other two processes, it plots

$$\frac{1}{k}\log G(w), \text{ where } G(w) \in \{1, \ldots, 2^k\},$$

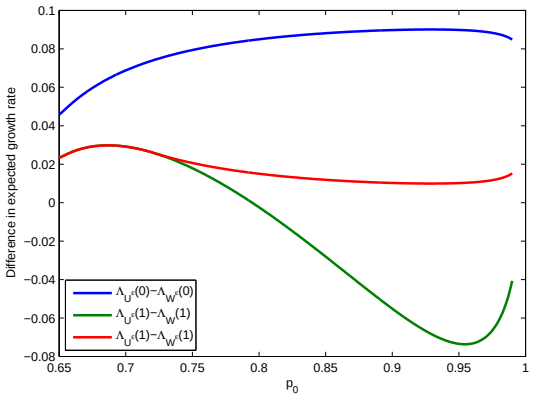

Figure 4.1: Bernoulli$(p_0, 1-p_0)$ source. Difference in exponential growth rates of guesswork between uniform approximation, unconditioned and conditioned distribution with $\epsilon = 0.1$. Top curve is the difference in expected logarithms between the uniform approximation and both the conditioned and unconditioned string sources. Bottom curve is the log-ratio of the expected guesswork of the uniform and unconditioned string sources, with the latter harder to guess for large p_0. Middle curve is the log-ratio of the uniform and conditioned string sources, which initially follows the lower line, before separating and staying positive, showing that the conditioned source is always easier to guess than the typically used uniform approximation.

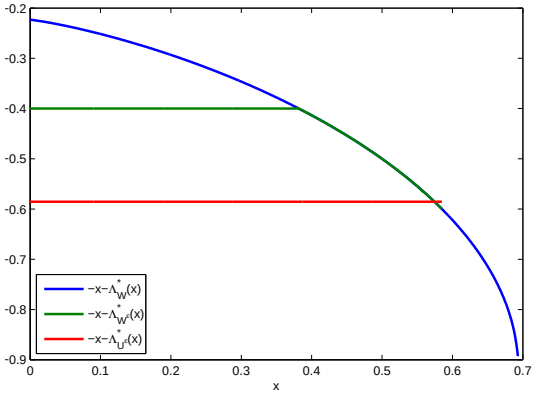

Figure 4.2: Bernoulli$(8/10, 2/10)$ source, $\epsilon = 0.1$. Guesswork distribution approximations. For large k, x-axis is $x = 1/k \log G(w)$ for $G(w) \in \{1, \ldots, 2^k\}$ and the y-axis is the large deviation approximation $1/k \log P(X = w) \approx -x - \Lambda_X^*(x)$ for $X = W_k, W_k^\epsilon$ and $X = U_k^\epsilon$.

against the large deviation approximations to

$$\frac{1}{k} \log P(W_k = w), \frac{1}{k} \log P(W_k^\epsilon = w) \text{ and } \frac{1}{k} \log P(U_k^\epsilon = w),$$

as the resulting plot is unchanging in k. The source of the discrepancy in expected guesswork is apparent, with the unconditioned source having substantially more strings to cover (due to the log x-scale). Both the unconditioned source and the true conditioned sources having higher probability strings that skew their guesswork. The first plateau for the conditioned and uniform distributions correspond to those strings with maximum highest probability (slowest exponential decay-rate).

4.5 Generalisation

In the third and final part of this chapter we show that the result is not confined by the i.i.d. assumption. This section closely follows the proof of Pfister and Sullivan [46].

Let $\mathbb{A} := \{0, \ldots, m-1\}$ equiped with the discrete topology and $\Omega = \mathbb{A}^{\mathbb{N}}$. Define $X_k : \mathbb{A}^{\mathbb{N}} \to \mathbb{A}^k$ to be the projection $w \in \mathbb{A}^{\mathbb{N}} \to (w_1, \ldots, w_k) \in \mathbb{A}^k$. We let M denote the space of Borel probability measures on Ω and define $S : \Omega \to \Omega$ to be the shift operator $(S(w))_j := w_{j+1}$ for each $j \in \mathbb{N}$. Let $M_S \subset M$ denote the the shift invariant probability

measures, so $\nu \in M_S$ implies $\nu(w) = \nu(Sw)$ for all $w \in \Omega$. Define

$$\Sigma_k^\nu := \{w_k \in \mathbb{A}^k : \nu_k(w_k) > 0\}, \quad \Sigma^\nu := \bigcap_k X_k^{-1}(\Sigma_k^\nu).$$

Then M^ν denotes the set of Borel probability measures on Σ^ν, M_S^ν the shift invariant probability measures on Σ^ν. Assume that for $\nu \in M_S$, a string is chosen with probability $\nu(w)$ and a string w_k of length k is chosen with probability $P(W_k = w_k) = \nu_k(w_k) = \nu(\{w : X_k(w) = w_k\})$. Use $L_k(w)$ to denote the empirical measure

$$L_k(w) := \frac{1}{k}\left(\delta_w + \delta_{S(w)} + \ldots + \delta_{S^{k-1}(w)}\right)$$

where $\delta_{S^j(w)}$ denotes the measure concentrated on the point $S^j(w) = (w_{j+1}, w_{j+2}, \ldots)$ and $\delta_w = \delta_{S^0}(w)$. The number of guesses that has to be made to guess w_k is labeled $G(w_k)$ and $G : \mathbb{A}^k \to \{1, \ldots, m^k\}$ that has the properties that $G(w_k) < G(w_k')$ if and only if $\nu_k(w_k') < \nu_k(w_k)$, and that $G(w_k) = G(w_k')$ implies $w_k = w_k'$. Using the notation of [46], the specific Shannon Entropy of ν is

$$h_{sh}(\nu) := -\lim_{k\to\infty} \frac{1}{k} \sum_{w_k \in \mathbb{A}^k} \nu_k(w_k) \log \nu_k(w_k).$$

We make three assumptions on ν, the first two of which are taken from [46].

Assumption 4.2 *For any neighbourhood U of $\rho \in M_S^\nu$ and for any $\epsilon > 0$, there exists an ergodic $\rho' \in U \cap M_S^\nu$ such that*

$$h_{sh}(\rho') \geq h_{sh}(\rho) - \epsilon.$$

Assumption 4.3 *The given reference probability measure ν is shift invariant. There exists a continuous nonnegative function $e_\nu : \Omega \to \mathbb{R}$ satisfying*

$$\lim_{k\to\infty} \sup_{w : X_k(w) \in \Sigma_k^\nu} \left| \frac{1}{k} \log \nu_k(X_k(w)) + e_\nu(w) \right| = 0.$$

Assumption 4.3 implies that the probability, $\nu_k(X_k(w))$, is approximately determined by the first k characters of w. This gives us some regularity on the system allowing us to break the space into types. There is no assumption that there is only one possible

e_ν that satisfies Assumption 4.3 but any such function will be sufficient as long as the reader is consistent.

For the final assumption we need to define the specific Rényi entropy in terms of our current notation. The specific Rényi entropy is defined for $\beta > 0, \beta \neq 1$,

$$R_W(\beta) := \lim_{k\to\infty} \frac{1}{k(1-\beta)} \log \sum_{w_k \in \mathbb{A}^k} \nu_k(w_k)^\beta, \tag{4.16}$$

with $R_W(1) := h_{sh}(\nu)$.

Assumption 4.4 *The shift invariant measure ν is ergodic and the function $R_W(\beta)$ has a continuous derivative for $\beta \in (0, \infty)$.*

Given $\epsilon > 0$, strings of length k the Typical Set, equation (4.1), can be written as

$$T_k^\epsilon := \left\{ w_k \in \mathbb{A}^k : \frac{1}{k} \log \nu_k(w_k) \in (-h_{sh}(\nu) - \epsilon, -h_{sh}(\nu) + \epsilon) \right\}.$$

Define

$$M^{\nu,\epsilon} := \{\rho \in M^\nu : e_\nu(\rho) \in (h_{sh}(\nu) - \epsilon, h_{sh}(\nu) + \epsilon)\}.$$

Conditioning on the Typical Set provides a new source of strings of length k:

$$\nu_k^\epsilon(w_k) = \begin{cases} \dfrac{\nu_k(w_k)}{\nu_k(T_k^\epsilon)} \text{ if } w_k \in T_k^\epsilon \\ 0 \text{ otherwise.} \end{cases} \tag{4.17}$$

Now we will define some statistics of these processes so we have something concrete to compare what happens to the guesswork when the two different processes, $\{\nu_k\}, \{\nu_k^\epsilon\}$, are used.

The sCGF of the original process, $\{\nu_k\}$, is defined for $\alpha \in \mathbb{R}$,

$$\Lambda_W(\alpha) := \lim_{k\to\infty} \frac{1}{k} \log \sum_{w_k \in A^k} \nu(w_k)(\exp(\alpha \log G(w_k))).$$

The equivalent for the process conditioned on the Typical Set is

$$\Lambda_{W^\epsilon}(\alpha) := \lim_{k\to\infty} \frac{1}{k} \log \sum_{w_k \in T_k^\epsilon} \nu_k^\epsilon(w_k)(\exp(\alpha \log G(w_k))).$$

The specific Rényi entropy of the process conditioned on being in the Typical Set is denoted by $R_{W^\epsilon}(\beta)$ and is obtained by replacing $\nu_k(w_k)$ with $\nu_k^\epsilon(w_k)$ in (4.16).

The condition needed for Theorem 3.3 in Chapter 3 to hold is that $\Lambda_W(\alpha)$ exists, for all $\alpha > -1$ that,

$$\Lambda_W(\alpha) = \alpha R_W\left(\frac{1}{1+\alpha}\right)$$

and that $\Lambda'_W(\alpha)$ is continuous for $\alpha > -1$. The continuity of the derivative is true by Assumption 4.4 as

$$\frac{d}{d\alpha}\left(\alpha R_W\left(\frac{1}{1+\alpha}\right)\right) = R_W\left(\frac{1}{1+\alpha}\right) + \alpha \frac{d}{d\alpha} R_W\left(\frac{1}{1+\alpha}\right)$$

which is continuous as R_W and $(1+\alpha)^{-1}$ are continuous. The rest of the condition is proven by Pfister and Sullivan [46].

4.5.1 Main Theorems

To use Theorem 3.3 in Chapter 3 for the measure as defined in equation (4.17), ν_k^ϵ, we need to establish that $\Lambda_{W^\epsilon}(\alpha)$ exists, for all $\alpha > -1$, that

$$\Lambda_{W^\epsilon}(\alpha) = \alpha R_{W^\epsilon}\left(\frac{1}{1+\alpha}\right)$$

and that $\Lambda'_{W^\epsilon}(\alpha)$ is continuous for $\alpha > -1$.

Establishing that condition holds is achieved by the following three theorems, whose proofs follow later.

Theorem 4.3 *Let $\nu \in M_S$ satisfy Assumptions 4.2, 4.3 and 4.4. If $\beta > 0$, $\beta \neq 1$ then*

$$\lim_{k\to\infty} \frac{1}{k(1-\beta)} \log \sum_{w_k \in T_k^\epsilon} (\nu_k(w_k))^\beta = \frac{1}{1-\beta} \sup_{\rho \in M_S^{\nu,\epsilon}} \left[h_{sh}(\rho) - \beta e_\nu(\rho) \right].$$

With

$$\lim_{\beta\uparrow 1} \lim_{k\to\infty} \frac{1}{k(1-\beta)} \log \sum_{w_k \in T_k^\epsilon} (\nu_k(w_k))^\beta = \lim_{\beta\downarrow 1} \lim_{k\to\infty} \frac{1}{k(1-\beta)} \log \sum_{w_k \in T_k^\epsilon} (\nu_k(w_k))^\beta = h_{sh}(\nu).$$

Theorem 4.4 *Let $\nu \in M_S$ satisfy Assumptions 4.2, 4.3 and 4.4. If $\alpha > -1$, then*

$$\Lambda_{W^\epsilon}(\alpha) = \lim_{k\to\infty} \frac{1}{k} \log \sum_{w_k \in T_k^\epsilon} \nu_k^\epsilon(w_k)(G(w_k)^\alpha) = \sup_{\rho \in M_S^{\nu,\epsilon}} [(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)].$$

Theorem 4.5 *Let $\nu \in M_S$ satisfy Assumptions 4.2, 4.3 and 4.4. Then the sCGF, $\Lambda_{W^\epsilon}(\alpha)$ has a continuous derivative for $\alpha \in (-1, \infty]$.*

It can be quickly seen that for $\alpha \geq 0$

$$\sup_{\rho \in M_S^{\nu,\epsilon}} [(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)] \leq \sup_{\rho \in M_S^{\nu}} [(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)]$$

so that conditioning on the Typical Set will never increase the rate at which the average guesswork increases at.

On the other hand

$$\Lambda_{U^\epsilon}(1) = \log |T_k^\epsilon| = \sup_{\rho \in M_S^{\nu}} h_{sh}(\rho) = \sup_{\alpha \in \mathbb{R}} (\sup_{\rho \in M_S^{\nu}} [(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)])'$$

with equality only if W is uniformly distributed. This in conjunction with $\Lambda_{U^\epsilon}(0) = \Lambda_{W^\epsilon}(0) = 0$ and $\Lambda'_{W^\epsilon}(0) = \sup_{\rho \in M_S^{\nu}} h_{sh}(\rho) \geq h_{sh}(\nu) = \Lambda'_{W^\epsilon}(0)$ means that $\Lambda_{U^\epsilon}(1) > \Lambda_{W^\epsilon}(1)$ if W is not uniformly distributed again showing that assuming that all strings inside the Typical Set are uniform is ill advised.

4.5.2 Proofs

To prove the theorems in the main results section we are going to state two interim propositions that will be proved first.

Proposition 4.1 *Let $\nu \in M_S$ satisfy Assumption 4.3. Let F be a closed subset of $M^{\nu,\epsilon}$. Then*

$$\limsup_{k\to\infty} \frac{1}{k} \log \frac{\nu(\{w \in \mathbb{A}^{\mathbb{N}} : X_k(w) \in T_k^\epsilon, L_k(w) \in F\})}{\nu_k(T_k^\epsilon)} \leq \sup_{\rho \in F \cap M_S^\nu} -h(\rho|\nu).$$

Proposition 4.2 *Let $\nu \in M_S$ satisfy Assumption 4.2 and Assumption 4.3. Let D be a open subset of $M^{\nu,\epsilon}$. Then*

$$\liminf_{k\to\infty} \frac{1}{k} \log \frac{\nu(\{w \in \mathbb{A}^{\mathbb{N}} : X_k(w) \in T_k^\epsilon, L_k(w) \in D\})}{\nu_k(T_k^\epsilon)} \geq \sup_{\rho \in D \cap M_S^\nu} -h(\rho|\nu).$$

These give us a lower bound and an upper bound on the probability the chosen string has L_k in a certain set, this is an abstraction of the notion of types used in [37] and the proof in the first part of this chapter.

For the proof we will approximate e_ν with a local function f which we need to define some notation for next. Let $f : \Omega \to \mathbb{R}$, $\|f\| := \sup_{w\in\Omega} |f(w)|$. Then let $\mathcal{F}_k$ denote the σ-algebra generated by X_k. Write $f \in \mathcal{F}_k$ to mean that the function f is $\mathcal{F}_k$ measurable.

To divide up the possible string space into more manageable blocks we use the following notation.

For $B \subset M$,

$$\hat{\Gamma}_{k,B}^\epsilon = \{w_k \in T_k^\epsilon : \exists w \in \Omega \text{ with } X_k(w) = w_k \text{ and } L_k(w) \in B\}.$$

For $B \subset M$,

$$\tilde{\Gamma}_{k,B}^\epsilon = \{w_k \in T_k^\epsilon : X_k(w) = w_k \Rightarrow L_k(w) \in B\}.$$

We need to prove some smaller results first before we prove Proposition 4.1 and 4.2.

The following lemma allows us to show the effect of conditioning the strings on being in the Typical Set.

Lemma 4.6 *Let ν satisfy Assumptions 4.2, 4.3 and 4.4. Then* $\lim_{k\to\infty} k^{-1} \log \nu_k(T_k^\epsilon) = 0$.

PROOF: By Cover & Thomas, Theorems 3.1.2 and 15.7.1 [14], using the fact that $\mathbb{A}$ is finite, the ergodicity and stationarity of ν.

■

The following lemma is similar to Lemma 4.1 in Pfister and Sullivan [46] but we replace $\hat{\Gamma}$ and $\tilde{\Gamma}$ with $\hat{\Gamma}^\epsilon$ and $\tilde{\Gamma}^\epsilon$. The following Lemmas give us the tools we require for operating with $\hat{\Gamma}^\epsilon$ and $\tilde{\Gamma}^\epsilon$.

Lemma 4.7 *Let $F \subset D \subset M$ with F closed and D open. Then there exists $k' \in \mathbb{N}$ such that for all $k \geq k'$, $\hat{\Gamma}^\epsilon_{k,F} \subset \tilde{\Gamma}^\epsilon_{k,D}$.*

PROOF: If there were no such k' then we could find a sequence $\{(w_{k_n}, w'_{k_n})\} \in \mathbb{A}^{\mathbb{N}^2}$ with $X_{k_n}(w_{k_n}) = X_{k_n}(w'_{k_n})$ and $k^{-1} \log v_k(X_{k_n}(w_{k_n})) \in (h_{sh}(\nu) - \epsilon, h_{sh}(\nu) + \epsilon)$ such that

$$\begin{aligned} L_{k_n}(w_{k_n}) \in F, L_{k_n}(w'_{k_n}) \notin D \\ \lim_{k_n\to\infty} L_{k_n}(w_{k_n}) = \rho^* \in F, \lim_{k_n\to\infty} L_{k_n}(w'_{k_n}) = \rho' \notin D. \end{aligned} \tag{4.18}$$

If $f \in \mathcal{F}_b$ and $X_k(w) = X_k(w')$, then

$$|f(L_k(w)) - f(L_k(w'))| \leq 2\|f\| \frac{b-1}{k}, \tag{4.19}$$

so $f(\rho^*) = f(\rho')$. As equation (4.19) holds for all local f which implies $\rho^* = \rho'$ which contradicts (4.18). The inequality comes from $2\|f\|$ being the largest possible gap between them not taking into account $X_k(w) = X_k(w')$. $(b-1)/k$ is the fraction of the string that $f(L_k(w))$ depends on that where both strings are not exactly equal due to $X_k(w) = X_k(w')$.

■

The following lemma lower bounds the exponential rate at which $|\tilde{\Gamma}^{\epsilon}_{k,D}|$ increases.

Lemma 4.8 *Let D be an open set in M^{ν}. Let $\rho \in D$ be an ergodic probability measure on Σ^{ν}. Then*

$$\liminf_{k\to\infty} \frac{1}{k} \log \left|\tilde{\Gamma}^{\epsilon}_{k,D}\right| \geq h_{sh}(\rho).$$

PROOF: Let $\{f_i\}$ be a sequence of local functions that determines the topology of M. The open set $D \subset M^{\nu,\epsilon}$ can be expressed as $D = D' \cap M^{\nu,\epsilon}$, with D' open in M and has the property that $\rho' \in D' \Rightarrow e_{\nu}(\rho') \in (h_{sh}(\nu) - \epsilon, h_{sh}(\nu) + \epsilon)$ using the continuity of e_{ν}. As we are using the weak topology on $\mathbb{A}^{\mathbb{N}}$, there exists $\{\delta_1 > 0, \ldots, \delta_m > 0\}$ such that

$$N := \{\rho' \in M : |f_i(\rho' - \rho)| \leq \delta_i, i = 1, \ldots, m\} \subset D'. \tag{4.20}$$

By Lemma 4.7, for all sufficiently large k

$$\hat{\Gamma}^{\epsilon}_{k,N} \subset \tilde{\Gamma}^{\epsilon}_{k,D'} = \tilde{\Gamma}^{\epsilon}_{k,D}.$$

As ρ is assumed to be ergodic, there exists a Borel set $B \in \Sigma^{v}$ so that $\rho(B) = 1$ and $w \in B \Rightarrow \lim_{k\to\infty} f_i(L_k(w)) = f_i(\rho)$, $i = 1, \ldots, m$.

Any element $\rho' \in N$, N defined in equation (4.20), will have the property that $e_{\nu}(\rho') \in (h_{sh}(\nu) - \epsilon, h_{sh}(\nu) + \epsilon)$ as $N \subset D'$, which there exists K such that for $k > K$, for $w \in \Omega$, $L_k(w) \in N \Rightarrow X_k(w) \in T^{\epsilon}_k$ using Assumption 4.3. This part is the main difference to the Pfister and Sullivan version as we must make sure that we do not include elements outside of the Typical Set.

It follows that for each $w \in B$ there exists k_w so that $k > k_w$ implies $L_k(w) \in N$, hence

$$\lim_{k\to\infty} \rho_k(\tilde{\Gamma}^{\epsilon}_{k,D}) = 1.$$

The lemma, with the additional ergodicity assumption, follows by noting that $\tilde{\Gamma}^{\epsilon}_{k,D}$ is a supporting set which gives lower bounds on the sizes of supporting sets in terms of the Shannon Entropy as detailed in Lemma 2.1 [33].

■

Lemma 4.9 *Let D be an open set in M^ν. Let $\rho \in D$ be an probability measure on Σ^ν. Then*

$$\liminf_{k\to\infty} \frac{1}{k} \log \left| \tilde{\Gamma}^\epsilon_{k,D} \right| \geq h_{sh}(\rho).$$

PROOF: Let $C = \{\rho \in D \cap M^\nu_S : \rho \text{ is ergodic}\}$. To expand Lemma 4.8 past the assumption of ergodicity we use Assumption 4.2 which implies for open $D \subset M^\nu$,

$$\sup_{\rho \in D \cap M^\nu_S} h_{sh}(\rho) = \sup_{\rho \in C} h_{sh}(\rho).$$

■

The following lemma and corollary are Lemma 4.3 and Corollary 4.1 from Pfister and Sullivan [46] and are reprinted for ease of reference of the user.

Lemma 4.10 *Let $\nu \in M_S$ be a probability measure satisfying Assumption 4.3. Then for each $\delta > 0$ there exist $m_\delta, N_\delta \in \mathbb{N}$ and f_δ which is $\mathcal{F}_{m_\delta}$ measurable, so that $\forall k \geq N_\delta$, $\forall w \in T^\epsilon_k$, $|e_\nu(w) - f_\delta(w)| \leq \delta/2$ and*

$$\left| f_\delta(L_k(w)) + \frac{1}{k} \log \nu_k(X_k(w)) \right| < \delta.$$

Corollary 4.1 *Let $\nu \in M_S$ be a probability measure verifying Assumption 4.3. For $\rho \in M^{\nu,\epsilon}_S$ we have*

$$-\lim_{k\to\infty} \frac{1}{k} \sum_{w_k \in \Sigma^\nu_k} \rho_k(w_k) \log \nu_k(w_k) = e_\nu(\rho).$$

In Pfister and Sullivan [46], Lemma 4.4 they cover M^ν however for our needs it is sufficient to cover $M^{\nu,\epsilon}$.

Lemma 4.11 *For $\delta > 0$ and $f_\delta, m_\delta, N_\delta$ as in Lemma 4.10, there exists an integer, numbers $0 \leq a_0 \ldots < a_{K_\delta}$ with $a_j - a_{j-1} < \delta$, $j = 1, \ldots, K_\delta$ and sets $\{D^\delta_j \subset F^\delta_j \subset M^{\nu,\epsilon}$:*

$j = 1, \ldots, K_\delta\}$ *so that each* D_j^δ *is open and each* F_j^δ *is closed and*

$$\bigcup_{j=0}^{N_\delta} D_j^\delta = M^{\nu,\epsilon}$$
$$\rho \in D_j^\delta \Rightarrow |f_\delta(\rho) - a_j| < \delta$$
$$\rho \in F_j^\delta \Rightarrow |f_\delta(\rho) - a_j| \leq \delta.$$

PROOF: Define K'_δ, a'_j, D_j^δ and F_j^δ by

$$K'_\delta := \left\lceil \frac{1 + \|f_\delta\|}{\delta} \right\rceil, \; a'_j := \frac{j}{K_\delta} \|f_\delta\|.$$

Then select $a_0 = h_{sh}(\nu) - \epsilon + \delta$, then take all the a'_j satisfying $a'_j \in (h_{sh}(\nu) - \epsilon + \delta, h_{sh}(\nu) + \epsilon - \delta)$. Denote the number of such a_j's as $K_\delta - 1$ and set $a_1, \ldots, a_{K_\delta - 1}$ to be these values, then set a_{K_δ} to be $h_{sh}(\nu) + \epsilon - \delta$.

Take the sets

$$D_j^\delta := \{\rho \in M^\nu : |f_\delta(\rho) - a_j| < \delta\}$$
$$F_j^\delta := \{\rho \in M^\nu : |f_\delta(\rho) - a_j| \leq \delta\}.$$

■

Two definitions are needed for the next lemma. For $B \subset M$,

$$\hat{\Gamma}_{k,B} = \{w_k \in \mathbb{A}^k : \exists w \in \Omega \text{ with } X_k(w) = w_k \text{ and } L_k(w) \in B\}.$$

For $B \subset M$,

$$\tilde{\Gamma}^\epsilon_{k,B} = \{w_k \in \mathbb{A}^k : X_k(w) = w_k \Rightarrow L_k(w) \in B\}.$$

Lemma 4.12 *For each closed* $F \subset M^{\nu,\epsilon}$

$$\limsup_{k \to \infty} \frac{1}{k} \log \left| \hat{\Gamma}^\epsilon_{k,F} \right| \leq \sup_{\rho \in F \cap M_S^\nu} h_{sh}(\rho). \tag{4.21}$$

PROOF: Pfister and Sullivan [46] proved equation (4.21) in Theorem $II.2$, [46] for each $F' \subset M^\nu$. As $F \subset M^{\nu,\epsilon} \subset M^\nu$ their proof still holds. If F is closed in $M^{\nu,\epsilon}$ but not in M^ν then take $B = \bar{F}$ and then B is closed in M^ν so that

$$\limsup_{k\to\infty} \frac{1}{k} \log \left|\hat{\Gamma}^\epsilon_{k,F}\right| \leq \limsup_{k\to\infty} \frac{1}{k} \log \left|\hat{\Gamma}_{k,F}\right| \leq \limsup_{k\to\infty} \frac{1}{k} \log \left|\hat{\Gamma}_{k,B}\right| \leq \sup_{\rho\in F\cap M^\nu_S} h_{sh}(\rho).$$

■

The following proposition is merely repeating Proposition 2.1 from Pfister and Sullivan [46] for the reader.

Proposition 4.3 *Let $\nu \in M_S$ satisfy Assumption 4.3. Then for each $\rho \in M^\nu_S$,*

$$h(\rho|\nu) := \lim_{k\to\infty} \frac{1}{k} \sum_{w_k\in\Sigma^\nu_k} \rho_k(w_k) \log \frac{\rho_k(w_k)}{\nu_k(w_k)}$$

exists and equals $e_\nu(\rho) - h_{sh}(\rho)$.

PROOF: Proof of Proposition 4.1. Breaking the F into the sets F^δ_j from Lemma 4.11, and using Lemma 4.12 we have

$$\limsup_{k\to\infty} \frac{1}{k} \log \left|\hat{\Gamma}^\epsilon_{k,F\cap F^\delta_j}\right| \leq \sup_{\rho\in F^\delta_j\cap F\cap M^\nu_S} h_{sh}(\rho).$$

Also

$$\frac{1}{k} \log \frac{\nu_k(X_k(w) : L_k(w) \in F \cap F^\delta_j)}{\nu_k(T^\epsilon_k)} \leq \frac{1}{k} \log \left|\hat{\Gamma}^\epsilon_{k,F\cap F^\delta_j}\right| + \max_{w_k\in\hat{\Gamma}^\epsilon_{F^\delta_j\cap F}} \frac{1}{k} \log \frac{\nu_k(w_k)}{\nu_k(T^\epsilon_k)}.$$

from Lemma 4.10 we deduce that for $k > K_\delta$, $\rho \in F^\delta_j \cap M^\nu_S$ and $w_k \in \hat{\Gamma}^\epsilon_{k,F^\delta_j}$

$$\frac{1}{k} \log \frac{\nu_k(w_k)}{\nu_k(T^\epsilon_k)} \leq -a_j + 2\delta - \frac{1}{k} \log \nu_k(T^\epsilon_k),\ e_\nu(\rho) \leq a_j + 2\delta.$$

Using Proposition 4.3 to get

$$\sup_{\rho\in F\cap F^{\delta}\cap M_S^{\nu}} h_{sh}(\rho) \leq \sup_{\rho\in F\cap F_j^{\delta}\cap M_S^{\nu}} -h(\rho|\nu) + a_j + 2\delta.$$

Then

$$\frac{1}{k}\log\frac{\nu_k(X_k(w) : L_k(w)\in F\cap F_j^{\delta})}{\nu_k(T_k^{\epsilon})}$$

equals the maximum over the corresponding lim sup's with F replaced with $F\cap F_j^{\delta}$. This gives us

$$\begin{aligned}&\limsup_{k\to\infty}\frac{1}{k}\log\frac{\nu_k(\{X_k(w) : L_k(w)\in F\})}{\nu_k(T_k^{\epsilon})}\\ &\leq \max_{0\leq j\leq K_{\delta}}\sup_{\rho\in F\cap F_j^{\delta}\cap M_S^{\nu}} -h(\rho|\nu) + 4\delta - \limsup_{k\to\infty}\frac{1}{k}\log\nu_k(T_k^{\epsilon}).\end{aligned}$$

The proof follows by Lemma 4.6, so $\limsup k^{-1}\log\nu_k(T_k^{\epsilon}) = 0$ and as δ is arbitrarily close to 0.

■

PROOF: Proof of Proposition 4.2. First let $D = \bigcup_j (D_j^{\delta}\cap D)$. If $\rho\in D\cap M^{\nu,\epsilon}$ then there exists j such that $\rho\in D\cap D_j^{\delta}$. By Lemma 4.8 we have for sufficiently large k,

$$\frac{1}{k}\log\left|\tilde{\Gamma}^{\epsilon}_{k,D\cap D_j^{\delta}}\right| \geq h_{sh}(\rho) - \epsilon.$$

Then by an argument similar to Proposition 4.1,

$$-h(\rho|\nu) \leq h_{sh}(\rho) - e_{\nu}(\rho) + 2\delta$$

and

$$\frac{1}{k}\log\frac{\nu_k(w_k)}{\nu_k(T_k^{\epsilon})} \geq -a_j - 2\delta - \frac{1}{k}\log\nu_k(T_k^{\epsilon}).$$

Finally

$$\frac{1}{k}\log\frac{\nu_k\left(\tilde{\Gamma}^{\epsilon}_{k,D\cap D_j^{\delta}}\right)}{\nu_k(T_k^{\epsilon})} \geq h_{sh}(\rho)-\epsilon-a_j-2\delta-\frac{1}{k}\log\nu_k(T_k^{\epsilon})$$
$$\geq -h(\rho|\nu)-\epsilon-4\delta-\frac{1}{k}\log\nu_k(T_k^{\epsilon}).$$

As ϵ and δ are arbitrary, $\nu_k(X_k(L_k^{-1}(D))) \geq \nu_k\left(\tilde{\Gamma}^{\epsilon}_{k,D\cap D_j^{\delta}}\right)$ and by Lemma 4.6, the proposition follows.

■

The following proof is similar to the proof of Lemma 2.3 in Pfister and Sullivan except we only sum over values in the Typical Set instead of all strings in $\mathbb{A}^k$.

Proof: Proof of Theorem 4.3. We use the covers $\{D_j^{\delta}\}$ and $\{F_j^{\delta}\}$ introduced in Lemma 4.11. For $\alpha \geq 0, \rho \in F_j^{\delta} \cap M_S^{\nu}$, and $n \geq N_{\delta}$, arguing as above we deduce that

$$w_k \in \hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}} \Rightarrow \frac{1}{k}\log(\nu_k(w_k))^{\beta} \leq \beta(-e_{\nu}(\rho)+4\delta)$$

and

$$\limsup_{k\to\infty}\frac{1}{k}\log\left|\hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}}\right| \leq \sup_{\rho\in F_j^{\delta}\cap M_S^{\nu}} h_{sh}(\rho)$$

by Proposition 4.1, so

$$\limsup_{k\to\infty}\frac{1}{k}\log\sum_{w\in\hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}}}\left(\frac{\nu_k(w_k)}{\nu_k(T_k^{\epsilon})}\right)^{\beta} \leq \sup_{F_j^{\delta}\cap M_S^{\nu}}[h_{sh}(\rho)-\beta e_{\nu}(\rho)-4\beta\delta].$$

The theorem follows by noting that $\bigcup\tilde{\Gamma}^{\epsilon}_{k,D_j^{\delta}} \subset T_k^{\epsilon} = \bigcup\hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}}$, that the lim inf and lim sup of the total sum is the same as the maximum over all the sets and δ is arbitrary.

■

The following proof follows the same argument as Pfister and Sullivan in Theorem 2.4 the only difference being our notation, all the differences between their case and ours being handled by previous the lemmas and propositions.

PROOF: Proof of Theorem 4.4. Set $h_j := \left|\hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}}\right|, g_0 := 0, g_{j+1} := g_j + h_j$ select the ranking functions $\{\mathrm{rnk}_j\}$ so that

$$\mathrm{rnk}_j : \hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}} \to \{g_j + 1, \ldots, g_j + h_j\}$$

and define the injection $\mathrm{rnk} : T_k^{\epsilon} \to \{1, \ldots, g_{K_\delta+1}\}$

$$\mathrm{rnk}(w) := \min_j \{\mathrm{rnk}_j(w) : w \in \hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}}\}.$$

The properties of rnk imply that if $\alpha \geq 0$

$$\sum_{w_k \in T_k^{\epsilon}} \nu_k^{\epsilon}(w_k) G(w_k)^{\alpha} \leq \sum_{w_k \in T_k^{\epsilon}} \frac{\nu_k(w_k)}{\nu_k(T_k^{\epsilon})} \mathrm{rnk}^{\alpha}(w_k).$$

This inequality reverses if $\alpha \leq 0$.

For $\alpha \geq 0$

$$\sum_{w_k \in \hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}}} \frac{\nu_k(w_k)}{\nu_k(T_k^{\epsilon})} G(w_k)^{\alpha} \leq \sum_{i=1}^{h_j} (g_j + i)^{\alpha} \max_{w_k \in \hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}}} \frac{\nu_k(w_k)}{\nu_k(T_k^{\epsilon})}$$

by the bound

$$(g+h)^{\beta} h \geq \sum_{i=g+1}^{g+h} i^{\alpha} \geq \int_0^h x^{\beta} dx = \frac{h^{1+\beta}}{1+\beta}. \tag{4.22}$$

Equation (4.22) allows us to deduce

$$\begin{aligned} &\limsup_{k\to\infty} \frac{1}{k} \log \sum_{j=0}^{K_\delta} \sum_{i=1}^{h_j} (g_j + i)^{\alpha} \max_{w \in \hat{\Gamma}^{\epsilon}_{k,F_j^{\delta}}} \frac{\nu_k(w_k)}{\nu_k(T_k^{\epsilon})} \\ &\leq \max_{j=0,\ldots,K_\delta} \left[\limsup_{k\to\infty} \frac{1}{k} (\log h_j + \alpha \log(g_{j+1})) - a_j \right] + 2\delta. \end{aligned}$$

Define B_j,

$$B_j := \limsup_{k\to\infty} \frac{1}{k} \log h_j.$$

Choose j^* so that

$$(1+\alpha)B_j - a_j \le (1+\alpha)B_{j^*} - a_{j^*}, \qquad j = 0, \dots, N_\delta. \tag{4.23}$$

Remember $g_{j+1} = \sum_0^j h_j$, we have

$$\begin{aligned}
&\max_{j=0,\dots K_\delta} \limsup_{k\to\infty} \frac{1}{k}(\log h_j + \alpha \log(g_{j+1})) - a_j \\
&\le \max_{j=0,\dots,K_\delta} \left[B_j + \max_{k\le j} \alpha B_k - a_j \right] = (1+\alpha)B_{j^*} - a_{j^*}.
\end{aligned}$$

Using the same techniques as before, we get

$$\limsup_{k\to\infty} \frac{1}{k} \log E(G(w_k)^\alpha) \le \sup_{\rho\in M_S^{\nu,\epsilon}} (1+\alpha)h_{sh}(\rho) - e_\nu(\rho) + 4\delta.$$

Next we use the second inequality in (4.22) to deduce

$$\sum_{w_k \in \hat{\Gamma}^\epsilon_{k,F_j^\delta}} \nu_k(w_k) G(w_k)^\alpha \ge \frac{h_j^{1+\alpha}}{1+\alpha} \min_{w\in\hat{\Gamma}^\epsilon_{k,F_j^\delta}} \frac{\nu_k(w_k)}{\nu_k(T_k^\epsilon)}$$

for each j.

Since $\tilde{\Gamma}^\epsilon_{k,D_j^\delta} \subset \hat{\Gamma}^\epsilon_{k,F_j^\delta}$, we have

$$\liminf_{k\to\infty} \frac{1}{k} \log \sum_{w_k\in T_k^\epsilon} \nu(w_k) G(w_k)^\alpha \ge \sup_{\rho\in M_S^{\nu,\epsilon}} (1+\alpha)h_{sh}(\rho) - e_\nu(\rho) - 4\delta.$$

This covers the $\alpha \ge 0$ case. The inequality in equation 4.22 reverses to

$$(g+h)^\beta h \le \sum_{i=g+1}^{g+h} i^\beta \le \int_0^h x^\beta dx = \frac{h^{1+\beta}}{1+\beta}$$

From this, the upper bound is easy to find. For the lower bound, if $-1 < \alpha < 0$, note

that

$$\sum_{i=g_j+1}^{g_j+h_j} i^\alpha \geq (g_j+h_j)^\alpha h_j$$

We have

$$\sum_{w_k \in T_k^\epsilon} \nu_k^\epsilon(w_k)^\alpha \geq \min_{w_k \in \hat{\Gamma}^\epsilon_{k,F_j^\delta}} \frac{\nu_k(w_k)}{\nu_k(T_k^\epsilon)} \left(\sum_{i=0}^{j} h_i \right)^\alpha h_j.$$

We redefine $B_j := \liminf_{k\to\infty} k^{-1} \log h_j$ and choose j^* so that (4.23) is obtained. Then

$$\liminf_{k\to\infty} \frac{1}{k} \log \sum_{w_k \in T_k^\epsilon} \nu_k^\epsilon G(w_k)^\alpha \geq B_{j^*} + \alpha \max_{i \leq j^*} B_i - a_{j^*} - 2\delta.$$

This means $i < j^* \Rightarrow a_{j^*} - a_i > 0$, so

$$(1+\alpha)B_i \leq (1+\alpha)B_{j^*} - (a_{j^*} - a_i) \Rightarrow B_i < B_{j^*}. \tag{4.24}$$

Equation (4.24) means that

$$\liminf_{k\to\infty} \frac{1}{k} \log \sum_{w_k \in T_k^\epsilon} \nu_k^\epsilon(w_k) G(w_k)^\alpha \geq \max_{j=0,\ldots,K_\delta} [(1+\alpha)B_j - a_j] - 2\delta.$$

As before, using the fact that $\tilde{\Gamma}_{k,D_j^\delta} \subset \hat{\Gamma}^\epsilon_{k,F_j^\delta}$,

$$\max_{j=0,\ldots,K_\delta} [(1+\alpha)B_j - a_j] - 2\delta \geq \sup_{\rho \in M_S^{\nu,\epsilon}} (1+\alpha)h_{sh}(\rho) - e_\nu(\rho) - 4\delta$$

giving us the lower bound.

■

Lemma 4.13 *Let ν satisfies Assumptions 4.2, 4.3 and 4.4 and $I_W(x)$ be the rate function with which $\{k^{-1}\log G(W_k)\}$ (if the strings are chosen using the measure ν_k as opposed to ν_k^ϵ) satisfies a Large Deviations Principle. If x is in the region*

$\{x' : \sup_{x\in\mathbb{R}}[\alpha x - I_W(x)] = \alpha x' - I_W(x') \text{ for some } \alpha > -1\}$ *then there exists an* $\alpha > -1$ *and a sequence* $(\rho_1, \ldots)$*, with* $\rho_j \in M_S^\nu$*, that satisfies* $\lim_{j\to\infty} h_{sh}(\rho_j') = x$ *and*

$$\sup_{\rho\in M_s^\nu}[(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)] = \lim_{j\to\infty}[(1+\alpha)h_{sh}(\rho_j') - e_\nu(\rho_j')].$$

Proof: By Varadhan's Lemma (see e.g. [16] Section 4.3) on the process ν,

$$\sup_{\rho\in M_s^\nu}[(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)] = \Lambda_W(\alpha) = \sup_{x\in\mathbb{R}}[\alpha x - I_W(x)].$$

So if the left and right hand sides are satisfied by $(\rho_1, \ldots)$ and $(x_1, \ldots)$ respectively then

$$\lim_{j\to\infty}((1+\alpha)h_{sh}(\rho_j) - e_\nu(\rho_j))$$

and

$$\lim_{j\to\infty}(\alpha x_j - I_W(x_j))$$

describe tangents to Λ_W at α. As Λ_W is continuous at α by Assumption 4.4, by [48] Theorem 26.1, they must be equal. This implies that there exists a sequence $(\rho_1', \ldots)$ such that $\lim_{j\to\infty} h_{sh}(\rho_j') = x$ and

$$\lim_{j\to\infty}((1+\alpha)h_{sh}(\rho_j) - e_\nu(\rho_j)) = \sup_{\rho\in M_s^\nu}[(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)].$$

■

Proof: Proof of Theorem 4.5. Solving the problem

$$\sup_{\rho\in M_S^{\nu,\epsilon}}[(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)]$$

has similarities to solving

$$\sup_{\rho\in M_S^\nu}[(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)]. \tag{4.25}$$

For α such that equation (4.25) is maximised by $\rho : e_\nu \in (h_{sh}(\nu) - \epsilon, h_{sh}(\nu) + \epsilon)$ the differentiability of Λ_{W^ϵ} at α follows by Assumption 4.4. If $e_\nu(\rho) < h_{sh}(\nu) - \epsilon$ at α' then

we will show

$$\sup_{\rho \in M^{\nu,\epsilon}} [(1+\alpha)h_{sh}(\rho) - e_\nu(\rho)]$$

is linear for $\alpha < \alpha'$. To show this take the sequences $B := (\rho_1, \ldots) \in M_S^{\nu,\epsilon}$ such that

$$\lim_{i\to\infty} h_{sh}(\rho_i) = \sup_{\rho \in M_S^{\nu,\epsilon}} h_{sh}(\rho).$$

Out of these sequences pick $(\rho'_1, \ldots)$ that satisfies

$$\lim_{j\to\infty} -e_{sh}(\rho'_j) = \sup_{(\rho_1,\ldots)\in B} \lim_{j\to\infty} -e_{sh}(\rho_j)$$

By Lemma 4.13 there exists an α^*:

$$\lim_{j\to\infty} ((1+\alpha^*)h_{sh}(\rho'_j) - e_\nu(\rho'_j)) = \sup_{\rho \in M_S^{\nu,\epsilon}} ((1+\alpha^*)h_{sh}(\rho) - e_\nu(\rho)).$$

As $\alpha \downarrow -1$,

$$\sup_{\rho \in M_S^{\nu,\epsilon}} ((1+\alpha)h_{sh}(\rho) - e_\nu(\rho)) \to \sup_{\rho \in M_S^{\nu,\epsilon}} (-e_\nu(\rho)) = \lim_{j\to\infty} (-e_\nu(\rho'_j)).$$

This means that the choice of $(\rho'_1, \ldots)$ in the function $\lim_{j\to\infty}(1+\alpha)h_{sh}(\rho'_j) - e_\nu(\rho'_j)$ gives a straight line between $\alpha \downarrow -1$ and $\alpha = \alpha^*$, achieving the supremum at both. By the convexity of the sCGF, it must therefore achieve the supremum at all points in between. A similar argument holds if equation (4.25) is satisfied by $\rho : e_\nu(\rho) \geq h_{sh}(\nu) + \epsilon$.

■

5 Guesswork for a Wiretap Erasures Channel

5.1 Introduction

A string is sent over a noisy channel that erases some of its characters. Knowing the statistical properties of the string's source and which characters were erased, a listener that is equipped with an ability to test the veracity of a string, one string at a time, wishes to fill in the missing pieces. Here we characterize the influence of the stochastic properties of both the string's source and the noise on the channel on the distribution of the number of attempts required to identify the string, its guesswork. In particular, we establish that the average noise on the channel is not a determining factor for the average guesswork and illustrate simple settings where one recipient with, on average, a better channel than another recipient, has higher average guesswork. These results stand in contrast to those for the capacity of wiretap channels and suggest the use of techniques such as friendly jamming with pseudo-random sequences to exploit this guesswork behavior.

As a concrete example in advance of the mathematical abstraction, consider a proximity card reader where an electronic signature, a password, is wirelessly transmitted when the card is near the reader. An unintended recipient is eavesdropping, but overhears the card's transmission via a noisy channel that erases certain characters. If the eavesdropper knows the string's source statistics and which characters were erased, how many guesses must he make before identifying the one that causes the card reader to notify success?

For i.i.d. character sources and noise that is independent of the string, but possibly correlated, Theorem 5.1 answers this question, providing an asymptotic approximation to the guesswork distribution as the string becomes long. Corollary 5.1 establishes that the mean number of erasures on the channel and the Shannon entropy of the character source determine the growth rate of the expected logarithm of the number of guesses required to identify the erased sub-string. The exponential growth rate of the average number of guesses, however, is determined by the scaling of the asymptotic moment of the number of erasures evaluated at the Rényi entropy, with parameter $1/2$, of the character distribution.

As a consequence of these results, we provide examples illustrating that the average guesswork can be smaller on a channel that is, on average, noisier demonstrating that average noise is not a useful statistic for guesswork. This conclusion may seem counter-intuitive in the context of capacity results for Wyner's wire-tap [55] that, when applied

to an erasure channel, indicate that secrecy capacity is non-zero only if the probability of erasure of the intended party is lower than that of the eavesdropper. Results in which a first receiver, with more erasures (on average) than a second receiver, can better recover a message than the second receiver are few. One recent exception is [15], which also considers the effect of randomness of erasures in message recovery. In contrast to our work, the authors consider secret message capacity in a specific setting that uses feedback to provide causal channel state information for the intended receiver, allowing the sender to transmit in a way that is advantageous to the intended receiver. In the case of two parties with an erasure, their scheme relies on the fact that the secret key agreement by public discussion from common information developed by [40] reduces to requiring only the channel state be shared over a public channel.

5.2 Guesswork and erasure channels

We begin with summarizing material on the mathematical formulation for guesswork and results from Chapter 3 that shall be needed here, followed by a brief overview of the relevance of erasure channels as models of wireless communication, as this material is not encountered elsewhere in this thesis.

Let $\mathbb{A} = \{0, \ldots, m-1\}$ be a finite alphabet and consider a stochastic sequence of words, $\{W_k\}$, where W_k is a string of length k taking values in $\mathbb{A}^k$. Assume that a word is selected and an inquisitor is equipped with a device, such as a one-way hash function, through which a word can be tested one at a time. With no information beyond the string length and the source statistics, their optimal strategy to identify the word is to generate a partial-order of the words from most likely to least likely and guess them in turn. That is, for each k the attacker generates a function $G : \mathbb{A}^k \to \{1, \ldots, m^k\}$ such that $G(w') < G(w)$ if $P(W_k = w') > P(W_k = w)$. For a word w the integer $G(w)$ is the number of guesses until the string w is guessed, its guesswork. The results in Chapter 3 prove that $\{k^{-1} \log G(W_k)\}$ satisfies a Large Deviation Principle (LDP).

In the present chapter we restrict to i.i.d. letter sources, but include noise sources that could potentially be correlated. This enables us to consider the erasures as a subordinating process for the guesswork, as will become clear.

Assumption 5.1 *The string W_k is constituted of independent and identically distributed*

characters with probability mass function $\{P(W_1 = i) : i \in \mathbb{A}\}$.

Under this assumption, if one must guess the entire word W_k, the following result is known.

Proposition 5.1 ([1, 46, 10]) *The scaled Cumulant Generating Function (sCGF) of* $\{k^{-1} \log G(W_k)\}$ *exists*

$$\Lambda_G(\alpha) = \lim_{k\to\infty} \frac{1}{k} \log E(\exp(\alpha \log(G(W_k)))) = \begin{cases} \alpha R\left(\dfrac{1}{1+\alpha}\right) & \text{if } \alpha > -1 \\ -R(\infty) & \text{if } \alpha \leq -1, \end{cases} \tag{5.1}$$

where $R(\alpha)$ *is the Rényi entropy with parameter* α*,*

$$R(\alpha) = \frac{1}{1-\alpha} \log \left(\sum_{i\in\mathbb{A}} P(W_1 = i)^\alpha \right)$$
$$R(\infty) = -\max_{i\in\mathbb{A}} \log P(W_1 = i).$$

Moreover, the process $\{k^{-1} \log G(W_k)\}$ *satisfies a Large Deviation Principle with rate function*

$$\Lambda_G{}^*(x) = \sup_{\alpha\in\mathbb{R}} (x\alpha - \Lambda_G(\alpha)).$$

As in [1], setting $\alpha = 1$ equation (5.1) gives

$$\begin{aligned} \Lambda_G(1) &= \lim_{k\to\infty} \frac{1}{k} \log E(G(W_k)) \\ &= R(1/2) = 2 \log \left(\sum_{i\in\mathbb{A}} P(W_1 = i)^{1/2} \right), \end{aligned}$$

establishing that the exponential growth rate of the average guesswork as the string gets longer is governed by Rényi entropy of the character distribution with parameter $1/2$, which is greater than its Shannon entropy, with equality if and only if the character source is uniformly distributed. The LDP gives the following approximation [10] for

large k and $n \in \{1, \ldots, m^k\}$,

$$P(G(W_k) = n) \approx \frac{1}{n} \exp\left(-k\Lambda_G{}^*\left(\frac{1}{k}\log n\right)\right).$$

Erasure models are common for coded communications. They arise for systems where an underlying error-correcting code can fail to correct the errors, but error-detection mechanisms will lead to detection of the failure to correct. While it is possible for errors to remain uncorrected in such a way that the receiver cannot detect the failure to correct. That traditional algebraic codes with n symbols of redundancy can correct up to n errors but detect up to $2n - 1$ errors justifies the common assumption that failures to detect errors may be neglected, whereas failures to correct may not. Failure to correct errors may be a design goal in certain systems. In wiretap channels, codes are deliberately constructed in such a way that, under channel conditions less favorable than those of the intended receiver, codes fail to decode (e.g. [6]).

Both intended and unintended recipients may observe the transmitted string through parallel channels or through a common channel where there exists a dependence in the proportion of erasures at different receivers. Such scenarios mirror those that have been considered for secrecy capacity, with the latter having been extensively studied as a model for wireless channels in which the unintended recipients are inquisitors (e.g. [6]) and the former considered less often [56, 57, 58].

In wireless erasure channels, there exist several means of achieving differentiated channel side information between intended receivers and inquisitors. Consider, for example, a fading channel, where fades lead to erasures and where fading characteristics permit prediction of future fades from current channel measurements. A receiver that actively sounds the channel, or receives channel side information from a sender, may know, perfectly or imperfectly, which erasures will occur over some future time.

Friendly jamming instantiates different channel side information between intended and unintended recipients by actively modifying the channel. Friendly jamming has been proposed and demonstrated to modify secrecy regions in wiretap-like settings [23, 53]. A notion related to friendly jamming is that of cooperative jamming [52] where multiple users collude in their use of the shared channel in order to reduce an inquisitor's ability.

5.3 Subordinated Guesswork - general results

We wish to consider the guesswork required to identify a string, W_k, sent over a stochastic, noisy channel that erases characters. We assume that a listener is equipped with an ability to test the veracity of each missing sub-string and wishes to fill in the missing piece. As the string W_k is made up of i.i.d. characters, if $N_k \in \{1, \ldots, k\}$ is the number of characters erased by the noise, the listener must effectively guess a word of N_k characters in length. Thus we are interested in properties of the the guesswork of the word subordinated by the erasures process, $G(W_{N_k})$, wishing to understand the influence of the properties of the string source and the noise on the channel on the distribution of the number of attempts required to identify the missing sub-string.

While in this chapter we assume that the string is made up of i.i.d. characters, the erasure process can be correlated and we make the following assumption, which encompasses, for example, Markovian erasure processes.

Assumption 5.2 *The erasure process is such that $\{N_k/k\}$, where N_k is the number of erasures in a string of length k, satisfies a LDP with convex rate function $\Lambda_N{}^* : \mathbb{R} \mapsto [0, \infty]$ such that $\Lambda_N{}^*(y) = \infty$ if $y \notin [0, 1]$.*

That is, the number of erasures satisfies Cramér's Theorem (e.g [16][2.1.24]). Loosely speaking, this implies that $P(N_k \approx yk) \asymp \exp(-k\Lambda_N{}^*(y))$.

The main theorem in this chapter is the following.

Theorem 5.1 *The subordinated guesswork process $\{k^{-1} \log G(W_{N_k})\}$ satisfies a LDP with convex rate function*

$$\Lambda^*_{NG}(x) = \inf_{y \in [0,1]} \left(y\Lambda_G{}^* \left(\frac{x}{y} \right) + \Lambda_N{}^*(y) \right). \tag{5.2}$$

The sCGF for $\{k^{-1} \log G(W_{N_k})\}$, the Legendre-Fenchel transform of Λ^*_{NG}, is given by the composition of the sCGF for the erasures with the sCGF for the non-subordinated guesswork

$$\Lambda_{NG}(\alpha) = \lim_{k \to \infty} \frac{1}{k} \log E \left(\exp(\alpha \log(G(W_{N_k}))) \right) = \Lambda_N(\Lambda_G(\alpha)).$$

PROOF: The method of proof of the LDP is akin to that used in Chapter 3, establishing that the upper and lower deviation functions coincide, followed by an application of the contraction principle. With $B_\epsilon(x) = (x - \epsilon, x + \epsilon)$. We first show that

$$\begin{aligned}
&\lim_{\epsilon \downarrow 0} \liminf_{k \to \infty} \frac{1}{k} \log P\left(\frac{1}{k} \log G(W_{N_k}) \in B_\epsilon(x), \frac{N_k}{k} \in B_\epsilon(y)\right) \\
&= \lim_{\epsilon \downarrow 0} \limsup_{k \to \infty} \frac{1}{k} \log P\left(\frac{1}{k} \log G(W_{N_k}) \in B_\epsilon(x), \frac{N_k}{k} \in B_\epsilon(y)\right) \\
&= y\Lambda_G{}^*\left(\frac{x}{y}\right) + \Lambda_N{}^*(y) \text{ for all } x \geq 0, y \in [0,1].
\end{aligned}$$

For example, for $y \in (0,1]$, consider

$$\begin{aligned}
&\frac{1}{k} \log P\left(\frac{1}{k} \log G(W_{N_k}) \in B_\epsilon(x), \frac{N_k}{k} \in B_\epsilon(y)\right) \\
\geq &\frac{1}{k} \log P\left(\frac{1}{k} \log G(W_{\lfloor k(y-\epsilon) \rfloor}) \in B_\epsilon(x)\right) + \frac{1}{k} \log P\left(\frac{N_k}{k} \in B_\epsilon(y)\right).
\end{aligned}$$

Taking $\liminf_{k \to \infty}$, using the LDPs for $\{k^{-1} \log G(W_k)\}$ and $\{N_k/k\}$ followed by $\lim_{\epsilon \downarrow 0}$ gives an appropriate lower bound. An equivalent upper bound follows similarly.

For $y = 0$, if $x > 0$ we can readily show that the upper deviation function takes the value $-\infty$ as $G(W_{\lfloor \epsilon y \rfloor}) \leq m^{y\epsilon}$. If $x = 0$, then the lim sup bound is achieved by solely considering the erasure term, while for the lim inf consider the ball $G(W_{N_k}) \leq \exp(k\epsilon \log(m))$, which has probability 1 and so the upper and lower deviation functions again coincide.

As the state space is compact, the LDP for $\{(k^{-1} \log G(W_{N_k}), N_k/k)\}$ follows (e.g. [31, 16]) with the rate function $y\Lambda_G{}^*(x/y) + \Lambda_N{}^*(y)$. From this LDP, the LDP for $\{(k^{-1} \log G(W_{N_k})\}$ via the contraction principle [16] by projection onto the first coordinate.

To prove that $\Lambda^*_{NG}(x)$ is convex in x, first note that $y\Lambda_G{}^*(x/y)$ is jointly convex in x and y, with $y > 0$, by the following argument. For $\beta \in (0,1)$, set $\eta = \beta y_1/(\beta y_1 + (1-\beta)y_2) \in [0,1]$ and note that

$$\begin{aligned}
(\beta y_1 + (1-\beta)y_2)\Lambda_G{}^*\left(\frac{\beta x_1 + (1-\beta)x_2}{\beta y_1 + (1-\beta)y_2}\right) &= (\beta y_1 + (1-\beta)y_2)\Lambda_G{}^*\left(\eta \frac{x_1}{y_1} + (1-\eta)x_2 y_2\right) \\
&\leq \beta y_1 \Lambda_G{}^*\left(\frac{x_1}{y_1}\right) + (1-\beta)y_2 \Lambda_G{}^*\left(\frac{x_2}{y_2}\right),
\end{aligned}$$

where we have used the convexity of ${\Lambda_G}^*$. As the sum of convex functions is convex, $y{\Lambda_G}^*(x/y) + {\Lambda_N}^*(y)$ is convex and as the point-wise minimum of a jointly convex function is convex, $\Lambda^*_{NG}(x)$ is convex.

To see that the point-wise minimum of a jointly convex function is convex take a function $f(a,b)$ that satisfies

$$f(\beta a_1 + (1-\beta)a_2, \beta b_1 + (1-\beta)b_2) \leq \beta f(a_1, b_1) + (1-\beta) f(a_2 + b_2)$$

and as such is jointly convex. To show that $\inf_a f(a,b)$ is convex we need to show that

$$\inf_a f(a, \beta b_1 + (1-\beta)b_2) \leq \beta \inf_a f(a, b_1) + (1-\beta) \inf_a f(a, b_2).$$

Let $a^* = \arg\inf_a f(a, b_1)$ and $a' = \arg\inf_a f(a, b_2)$ then

$$\begin{aligned}\inf_a f(a, \beta b_1 + (1-\beta)b_2) \leq & f(\beta a * + (1-\beta)a', \beta b_1 + (1-\beta)b_2) \\ & \leq \beta \inf_a f(a, b_1) + (1-\beta) \inf_a f(a, b_2)\end{aligned}$$

where the first inequality holds by the nature of an infinum and the second by joint convexity.

An application of Varadhan's Lemma (Theorem 4.3.1 [16]) identifies the sCGF for the subordinated process as the Legendre Fenchel transform of Λ^*_{NG}, $\sup_{x\in\mathbb{R}}(\alpha x - \Lambda^*_{NG}(x))$. To convert this into an expression in terms of Λ_N and Λ_G observe that

$$\begin{aligned}&\sup_{x\in\mathbb{R}}(\alpha x - \Lambda^*_{NG}(x)) = \sup_{x\in\mathbb{R}}\sup_{y\in\mathbb{R}}\left(\alpha x - y{\Lambda_G}^*\left(\frac{x}{y}\right) - {\Lambda_N}^*(y)\right) \\ &= \sup_{y\in\mathbb{R}}\left(y\sup_{z\in\mathbb{R}}(\alpha z - {\Lambda_G}^*(z)) - {\Lambda_N}^*(y)\right) = \sup_{y\in\mathbb{R}}\left(y\Lambda_G(\alpha) - {\Lambda_N}^*(y)\right) \\ &= \Lambda_N(\Lambda_G(\alpha)).\end{aligned}$$

■

Theorem 5.1, in particular, identifies the growth rate of the average subordinated guesswork. By the duality of the Legendre Fenchel transform and the convexity of Λ^*_{NG} implies that $\Lambda^*_{NG} = \sup_{\alpha\in\mathbb{R}}(\alpha x - \Lambda_N(\Lambda_G(\alpha)))$.

Corollary 5.1 *The growth rate of the average of the logarithm of the subordinated guesswork is determined by the average proportion of erasures and the Shannon entropy of the character source*

$$\begin{aligned}\lim_{k\to\infty}\frac{1}{k}E\left(\log G(W_{N_k})\right) &= \frac{d}{d\alpha}\Lambda_N(\Lambda_G(\alpha))|_{\alpha=0}\\ &= \mu_N H_G,\end{aligned}$$

where

$$\mu_N = \lim_{k\to\infty}\frac{E(N_k)}{k} \text{ and } H_G = -\sum_{i\in\mathbb{A}} P(W_1=i)\log P(W_1=i)$$

are the long run average proportion of erasures and the Shannon entropy of the characters distribution respectively. The growth rate of the average subordinated guesswork is, however, given by the sCGF of the erasures evaluated at the character Rényi entropy at $1/2$*,*

$$\lim_{k\to\infty}\frac{1}{k}\log E\left(G(W_{N_k})\right) = \Lambda_N(\Lambda_G(1)) = \lim_{k\to\infty}\frac{1}{k}\log E\left(\exp\left(R(1/2)N_k\right)\right).$$

Thus the determining factor in the average guesswork is not the average proportion of erasures, but the scaling of the cumulant of the erasure process determined by the Rényi entropy with parameter $1/2$. This result is further illustrated in the next section.

These results have ramifications for wiretap models where there is an intended recipient with one channel and an unintended recipient receiving over another. On receipt of a noise corrupted string, both need to guess the erased piece. In this setting, the following corollary proves that if both channels have i.i.d. erasures, then the expected result that having more erasures on average implies having a higher average guesswork holds. In the Examples section that follows we establish this is not true in general.

Corollary 5.2 *Assume the erasures processes in both the intended receiver's and the unintended recipient's channels are i.i.d. with the probabilities of any given character being erased as* p *and* q *respectively. Let the number of characters erased by the intended receiver's and the unintended recipient's channel be* N_k^I *and* N_k^U *respectively. If* $p < q$

then

$$\lim_{k\to\infty}\frac{1}{k}\log E(G(W_{N_k^I})^\alpha) < \lim_{k\to\infty}\frac{1}{k}\log E(G(W_{N_k^U})^\alpha)$$

for $\alpha > 0$, the inequality reverses for $\alpha \in (-1, 0)$ with equality at $\alpha = 0$.

PROOF: We assume that $\alpha > 0$, with the proofs for $\alpha \in (-1, 0]$ being similar. By Corollary 5.1,

$$\lim_{k\to\infty}\frac{1}{k}\log E(G(W_{N_k^I})^\alpha) = \Lambda_{N^I}\left(R\left(\frac{1}{1+\alpha}\right)\right) = \log\left(1-p+p\exp\left(\alpha R\left(\frac{1}{1+\alpha}\right)\right)\right)$$

$$\text{and } \lim_{k\to\infty}\frac{1}{k}\log E(G(W_{N_k^U})^\alpha) = \Lambda_{N^U}\left(R\left(\frac{1}{1+\alpha}\right)\right) = \log\left(1-q+q\exp\left(\alpha R\left(\frac{1}{1+\alpha}\right)\right)\right).$$

Thus

$$\exp(\Lambda_{N^I}(\Lambda_G(\alpha))) - \exp(\Lambda_{N^U}(\Lambda_G(\alpha))) = \left(\exp\left(\alpha R\left(\frac{1}{1+\alpha}\right)\right) - 1\right)(p-q).$$

As $(\alpha R\,(1/(1+\alpha))) > 0$ and $p - q \leq 0$, $(\exp(\alpha R\,(1/(1+\alpha))) - 1)(p-q) < 0$ proving the corollary.

■

5.4 Examples

Corollary 5.1 tells us the growth rate of the average guesswork depends on both the distribution of the strings and the distribution of the erasures. We start with the case where the unintended recipient has an i.i.d. channel with an erasure probability of p while the intended receiver has a deterministic channel with a proportion μ of the characters erased. For the unintended recipient this gives

$$\Lambda_N(\beta) = \log(1-p+p\exp(\beta)).$$

Thus his average subordinated guesswork growth rate is

$$\Lambda_N(R(1/2)) = \log\left(1-p+p\exp\left(R\left(\frac{1}{2}\right)\right)\right) \geq pR(1/2),$$

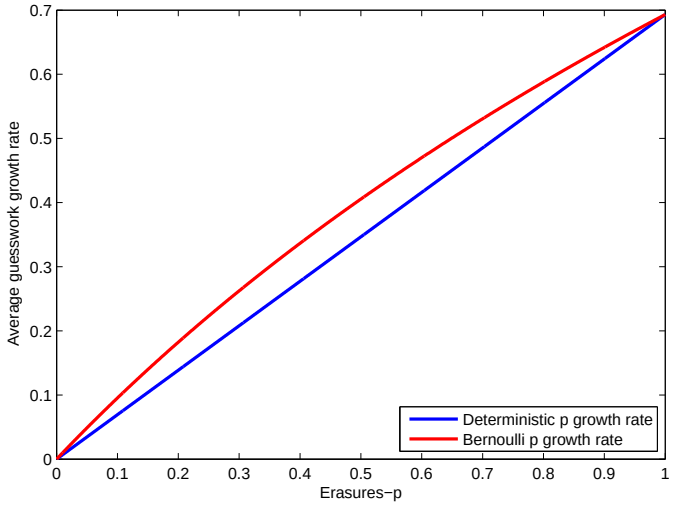

Figure 5.1: Binary source alphabet, $\mathbb{A} = \{0, 1\}$, with $P(W_1 = 0) = 1/2$. Average guesswork growth rate for deterministic channel with proportion p characters erased compared to a memoryless i.i.d. p erasure channel. For a given average number of erasures, the deterministic channel has a lower average guesswork.

where the latter follows by Jensen's inequality with equality if and only if $p = 0$ or 1.

As the intended receiver has a deterministic channel with a proportion μ of characters erased, the growth rate of its average subordinated guesswork is $\mu R(1/2)$. In particular, if $p < \mu < R(1/2)^{-1} \log(1 - p + p \exp(R(1/2)))$ then even though the channel of the unintended recipient is, on average, less noisy than the intended recipient, the average guesswork of the latter is smaller.

This also works in reverse, so if the intended receiver has an i.i.d. channel with erasure probability p and the unintended recipient has a deterministic channel with a proportion μ of characters erased then the average guesswork of the unintended recipient is smaller, for large enough strings, than the average guesswork of the intended receiver even though they may feel safe as their channel is, on average, less noisy.

Figures 5.1 and 5.2 show the potential difference in the asymptotic growth rate of the average guesswork if one channel is i.i.d. and the other is deterministic even if, on average, both channels have the same number of erasures. Both of these graphs assume that that the message is picked from a binary alphabet with $P(W_1 = 0) = P(W_1 = 1) = 0.5$.

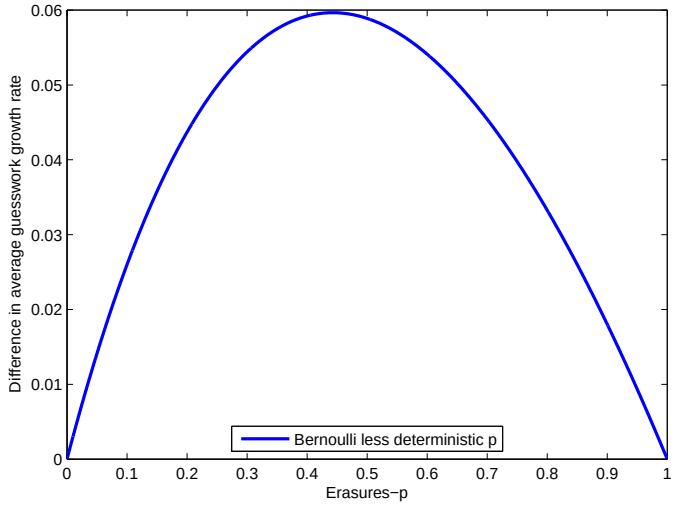

Figure 5.2: Binary source alphabet, $\mathbb{A} = \{0, 1\}$, with $P(W_1 = 0) = 1/2$. Similar to Figure 5.1, but plotting the difference between the i.i.d. p average guesswork growth rate and the deterministic p average guesswork.

For $p \in [0, 1]$, Figure 5.1 plots the average guesswork growth rate for the deterministic channel $pR(1/2)$ and for the i.i.d. channel $\log(1 - p + p\exp(R(1/2)))$. If $p \neq 0$ or 1, the i.i.d. channel has a higher average guesswork. Thus the intended recipient could have, on average, a less noisy channel, yet have a lower average guesswork. For clarity, Figure 5.2 plots the difference between these growth rates.

Figures 5.1 and 5.2 highlight the influence of the channel statistics on the average guesswork growth rate, but Figure 5.3 demonstrates the confounding influence of the source statistics. Here we assume that one channel is deterministic with 12% of characters erased while the other channel is i.i.d. with an average of 10% characters erased. Figure 5.3 plots the difference in average guesswork growth rate between these two channels as the source statistics change. If the source is less variable, the deterministic channel has a higher average guesswork, but as the source statistics become more variable, this reverses and the i.i.d. channel has higher average guesswork growth rate. In other strings, even though the average number of erasures on the deterministic channel is worse, dependent upon the source statistics its average guesswork may be lower than an i.i.d. channel with lower average number of erasures.

Between them, these examples indicate the trade-off in influence of the source and erasure statistics on the guesswork. While we have assumed the simplest erasure channels, these

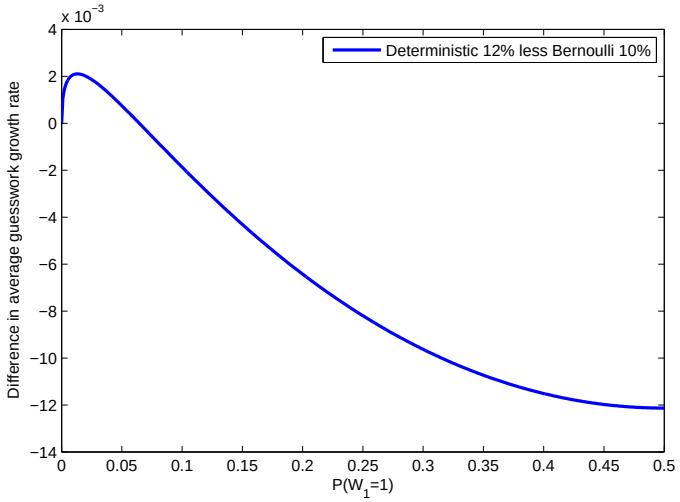

Figure 5.3: Binary source alphabet, $\mathbb{A} = \{0, 1\}$. Difference in average guesswork growth rate, as a function of $P(W_1 = 0)$, between a deterministic channel with 12% characters erased and an i.i.d. channel with 10% chance that each character is deleted. If the character source is less variable, the deterministic channel has a higher growth rate, but as the character source becomes more variable, it has a lower growth rate.

results are characteristic of the system. As a demonstration of that, a more realistic model of bursty type erasures would be that of a Markov chain governed by a transition matrix

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix},$$

where $a, b \in (0, 1)$. The first state corresponds to not erasing the character and the second state corresponds to erasing the character. As we are interested in asymptotic behaviour and our matrix is irreducible the starting distribution plays no role in our result. The stationary distribution is $(b/(a+b), a/(a+b))$ so that the latter is the long run average number of erasures. The second largest eigenvalue, $1-a-b$, is a measure of correlation, with the chain being positively correlated if it is greater than 0, negatively correlated if it is less than 0 and i.i.d. if it is 0.

The sCGF of the number of erasures $\Lambda_N(\theta)$ can be calculated using the techniques

described in [16], with the related rate-function given in [21],

$$\Lambda_N(\theta) = \log\left(\frac{1-a+(1-b)\exp(\theta)+\sqrt{4ab\exp(\theta)+(1-a-(1-b)\exp(\theta))^2)}}{2}\right). \tag{5.3}$$

To simplify matters, assume from here on that $a = b$, so that the long run average number of erasures is $a/(a+b) = 0.5$. Evaluation for any other average proportion of erasures is similar. We will use the notation N_b to signify this dependence on $b = a$. The growth rate of the average guesswork, as determined from Corollary 5.1, is

$$\begin{aligned}
&\Lambda_{N_b}(\Lambda_G(1)) = \\
&\log\left((1-b)\left(1+\exp\left(\alpha R\left(\frac{1}{2}\right)\right)\right)\right) - \log 2 \\
&+(1-b)\left(\sqrt{4b^2\exp\left(\alpha R\left(\frac{1}{2}\right)\right)+\left((1-b)\left(1-\exp\left(\alpha R\left(\frac{1}{2}\right)\right)\right)\right)^2}\right).
\end{aligned}$$

To understand how the average guesswork growth rate changes as b changes, we evaluate the first and second derivative with respect to b. The second derivative is

$$\frac{d^2}{db^2}\Lambda_{N_b}(\Lambda_G(1)) = \frac{4\exp\left(R\left(\frac{1}{2}\right)\right)\left(\exp\left(R\left(\frac{1}{2}\right)\right)-1\right)^2}{(4b^2\exp\left(R\left(\frac{1}{2}\right)\right)+(1-b)^2\exp\left(R\left(\frac{1}{2}\right)\right)-1)^2)^{3/2}},$$

which is positive as $R(1/2) \geq 0$ so $d\Lambda_{N_b}(\Lambda_G(1))/db$ is increasing in b. The first derivative is given by

$$\begin{aligned}
&\frac{d}{db}\Lambda_{N_b}(\Lambda_G(1)) = \\
&\frac{1}{2}\left(\frac{(b-1)\exp\left(2R\left(\frac{1}{2}\right)\right)+2(b+1)\exp\left(R\left(\frac{1}{2}\right)\right)+b-1}{\sqrt{4b^2\exp\left(R\left(\frac{1}{2}\right)\right)+(1-b)^2\exp\left(R\left(\frac{1}{2}\right)\right)-1)^2}}-\exp\left(R\left(\frac{1}{2}\right)\right)-1\right),
\end{aligned}$$

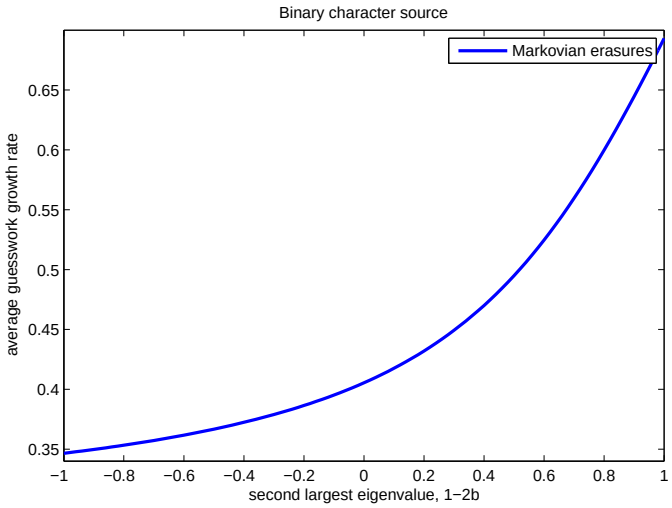

Figure 5.4: Binary source alphabet, $\mathbb{A} = \{0, 1\}$ with $P(W_1 = 0) = P(W_1 = 1) = 0.5$ and Markovian erasures with, on average, 50% erasures. Here we plot the growth rate of the average guesswork against the second largest eigenvalue of the transition matrix, $1 - 2b$, illustrating that two channels can have Markovian noise with the same average proportion of erasures and have different growth rates for the expected guesswork.

and, by taking $b \to 1$, we find

$$\frac{4\exp\left(R\left(\frac{1}{2}\right)\right)}{2\sqrt{\exp\left(R\left(\frac{1}{2}\right)\right)}} - \exp\left(R\left(\frac{1}{2}\right)\right) - 1 = -\exp\left(R\left(\frac{1}{2}\right)\right) + 2\sqrt{\exp\left(R\left(\frac{1}{2}\right)\right)} - 1$$
$$= -\left(\sqrt{\exp\left(R\left(\frac{1}{2}\right)\right)} - 1\right)^2 < 0$$

and so is negative for all b.

Figure (5.4) illustrates the previous example, showing how the average guesswork grows if we have Markovian erasures given by the transition matrix

$$\begin{pmatrix} 1-b & b \\ b & 1-b \end{pmatrix}.$$

This calculation shows that if $b > 0.5$, so erasures are more likely followed by non erasures and vice versa, then the expected guesswork growth rate is reduced in comparison to

that for the i.i.d. channel. If $b < 0.5$ then the expected guesswork is higher than for i.i.d. erasures.

5.5 Conclusions

We have characterized the asymptotic distribution of the guesswork required to reconstitute a string that has been subject to symbol erasure, as occurs on noisy communication channels. The scaled Cumulant Generating Function of the guesswork subordinated by the erasure process has a simple characterization as the composition of the sCGF of the noise with the sCGF of the unsubordinated guesswork. This form is redolent of the well-known result for the moment generating function for a random sun of random summands, but is an asymptotic result for guesswork. These results suggest that methods inspired from the secrecy capacity literature, such as the use of differentiated channel or source side information between the intended receiver and the eavesdropper, can be used in the context of guesswork. Indeed, numerical examples show that deterministic erasures can lead to lower average guesswork than Bernoulli erasures with a lower mean number of erasures. In further work, one may consider the behavior of guesswork in different settings that have been explored in the wiretap and cognate literature.

One may also envisage generalizing this analysis to the case where there are retransmissions of the entire string or of the symbols that have not been received by the intended receiver. Retransmissions are commonly employed in several protocols to enable reliability and, in the case of an erasure channel with perfect feedback, taking the form of acknowledgments, uncoded retransmission is capacity-achieving.

6 Multi-User Guesswork

6.1 Introduction

In the present Chapter we address a natural extension in this investigation of brute force searching: the quantification for multi-user systems. We are motivated by both classical systems, such as the brute force entry to a multi-user computer where the inquisitor need only compromise a single account, as well as modern distributed storage services where coded data is kept at distinct sites in a way where, owing to coding redundancy, several, but not all, servers need to be compromised to access the content [44, 20].

Assume that V users select strings independently from $\mathbb{A}^k$. An inquisitor knows the probabilities with which each user selects their string, is able to query the correctness of each (user, string) pair, and wishes to identify any subset of size U of the V strings. The first question that must be addressed is what is the optimal strategy, the ordering in which (user, string) pairs are guessed, for the inquisitor. For the single user system, since the earliest investigations [39, 1, 42, 47] it has been clear that the strategy of ordering guesses from the most to least likely string, breaking ties arbitrarily, is optimal in any reasonable sense. Here we shall give optimality a specific meaning: that the distribution of the number of guesses required to identify the unknown object is stochastically dominated by all other strategies. Amongst other results, for the multi-user guesswork problem we establish the following:

- If $U < V$, the existence of optimal guessing strategies, those that are stochastically dominated by all other strategies, is no longer assured.

- By construction, there exist asymptotically optimal strategies as the strings become long.

- For asymptotically optimal strategies, we prove a large deviation principle for their guesswork. The resulting large deviations rate function is, in general, not convex and so this result could not have been established by determining how the moment generating function of the guesswork distributions scale.

- The non-convexity of the rate function shows that, if users' string statistics are distinct, there may be no fixed ordering of weakness amongst users. That is, depending on how many guesses are made before the U users' strings are identified, the collection of users whose strings have been identified are likely to be distinct.

- If all V strings are chosen with the same statistics, then the rate function is convex and the exponential growth rate of the average guesswork as string-length increases is the specific Rényi entropy of the string source with parameter

$$\frac{V-U+1}{V-U+2} \in \left\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \dots\right\}.$$

- For homogeneous users, from an inquisitor's point of view, there is a law of diminishing returns for the expected guesswork growth rate in excess number of users $(V-U)$.

- For homogeneous users, from a designer's point of view, coming full circle to Massey's original observation that Shannon entropy has little quantitative relationship to how hard it is to guess a single string, the specific Shannon entropy of the source is a lower bound on the average guesswork growth rate for all V and U.

These results generalize both the original guesswork studies, where $U = V = 1$, as well as some of the results in [42, 26] where, as a wiretap model, the case $U = 1$ and $V = 2$ with one of the strings selected uniformly, is considered and scaling properties of the guesswork moments are established. Interestingly, we shall show that that setting is one where the LDP rate function is typically non-convex, so while results regarding the asymptotic behavior of the guesswork moments can be deduced from the LDP, the reverse is not true. To circumvent the lack of convexity, we prove the main result using the contraction principle, Theorem 4.2.1 [16], and the LDP established in [10], which itself relies on earlier results of work referenced above.

6.2 Optimal strategies

In single-user guesswork, a strategy, $S : \mathbb{A}^k \mapsto \{1, \dots, m^k\}$, is a one-to-one map that determines the order in which guesses are made. That is, for a given strategy S and a given string $w \in \mathbb{A}^k$, $S(w)$ is the number of guesses required before w is queried. In the multi-user setting, this generalises to an ordering of all (user, string) pairs $S : \{1, \cdots, V\} \times (\mathbb{A}^k)^V \to \{1, \cdots, (m^k)^V\}$. Here we allow an inquisitor to adapt their strategy in that they will stop guessing strings for a user once they have guessed that user's chosen string. However the analysis does not rely on this fact and so the results

would remain the same in either case. In all earlier work on single-user guesswork, ordering strings from most likely to least likely was intuitively wise. In order to discuss optimality of strategies when guessing U out of V strings, we must first revisit the question for $U = V = 1$, providing a precise definition. We do this by assigning optimality a precise meaning in terms of stochastic dominance [30, 17], which we show is satisfied by the previously studied single-user guesswork order.

Let $\vec{W}_k = (W_k^{(1)}, \ldots, W_k^{(V)})$ be a random vector taking values in $\mathbb{A}^{kV}$ with independent, not necessarily identically distributed, components. Each component $W_k^{(v)}$ corresponds to the string of length k chosen by user $v \in \{1, \ldots, V\}$ of which the inquisitor wishes to identify U. This means that $G_S(U, \vec{W})$ is the number of guesses until the the inquisitor has guessed U elements of $\vec{W}$ using the strategy S. However we will need more definitions to describe it mathematically.

Definition 6.1 *A strategy S is optimal for $\vec{W}$ if, for all strategies S' the random variable $G_S(U, \vec{W})$ is stochastically dominated by $G_{S'}(U, \vec{W})$. That is, if $P(G_S(U, \vec{W}) \leq n) \geq P(G_{S'}(U, \vec{W}) \leq n)$ for all strategies S' and all $n \in \{1, \ldots, m^k\}$.*

This definition captures the stochastic aspect of guessing, stating that an optimal strategy is one where the stopping time of identification is probabilistically smallest. One consequence of this definition that explains its appropriateness is that for any monotone function $\phi : \{1, \ldots, m^k\} \to \mathbb{R}$, it is the case that $E(\phi(S(W_k))) \leq E(\phi(S'(W_k)))$ for an optimal S and any other S' (e.g. Proposition 3.3.17, [17]), so that $G(W_k)$ has the least moments possible over all strategies.

Lemma 6.1 *If $V = U = 1$, the optimal strategies are those that guess from most likely to least likely.*

Proof: Consider a strategy G of guessing from most likely to least likely, breaking ties arbitrarily, and any other strategy S. By construction, for any $n \in \{1, \ldots, m^k\}$

$$\begin{aligned} P(G(W_k) \leq n) &= \sum_{i=1}^{n} P(G(W_k) = i) \\ &= \max_{w_1, \ldots, w_n} \left(\sum_{i=1}^{n} P(W_k = w_i) \right) \geq \sum_{i=1}^{n} P(S(W_k) = i) = P(S(W_k) \leq n). \end{aligned}$$

■

In the multi-user case, a strategy is now a one-to-one map $S : \{1, \ldots, V\} \times \mathbb{A}^k \mapsto \{1, \ldots, Vm^k\}$ that orders the guesses of (user, string) pairs. The expression for the number of guesses required to identify U strings is a little involved as we must take into account that we stop making queries about a user once their string has been identified.

For a given strategy S, let $N_S : \{1, \ldots, V\} \times \{1, \ldots, Vm^k\} \mapsto \{1, \ldots, m^k\}$ be defined by

$$N_S(v, n) = |\{w \in \mathbb{A}^k : S(v, w) \leq n\}|,$$

which computes the number of queries in the strategy up to query n that correspond to queries regarding user v.

The final query number made if only U strings need be identified is

$$\mathcal{S}(U, \vec{w}) := \text{U-min}\left(S(1, w^{(1)}), \ldots, S(V, w^{(V)})\right),$$

where U-min : $\mathbb{R}^V \to \mathbb{R}$ and U-min($\vec{x}$) gives the U^{th} smallest component of $\vec{x}$. The number of guesses required to identify U components of $\vec{w} = (w^{(1)}, \ldots, w^{(V)})$ is then

$$G_S(U, \vec{w}) = \sum_{v=1}^{V} N_S\left(v, \min\left(S(v, w^{(v)}), \mathcal{S}(U, \vec{w})\right)\right). \tag{6.1}$$

This apparently unwieldy object counts the number of queries made to each user curtailed either when their own string is identified or when U strings of other users are identified.

As an aside, note that if $U = V$, then $\min\left(S(v, w^{(v)}), \mathcal{S}(U, \vec{w})\right) = S(v, w^{(v)})$ for all $v \in \{1, \ldots, V\}$ and so equation (6.1) becomes

$$G_S(V, \vec{w}) = \sum_{v=1}^{V} N_S\left(v, S(v, w^{(v)})\right),$$

the sum of the individual guesswork of each string. Thus, using Lemma 6.1 repeatedly, if $U = V$, again there are optimal strategies, ones stochastically dominated by all others, which is to employ individual optimal strategies in any order.

The formula (6.1) will be largely side-stepped when we consider asymptotically optimal strategies, but is needed to establish that there is, in general, no stochastically dominant strategy if $V > U$. With $\vec{W}_k = (W_k^{(1)}, \ldots, W_k^{(V)})$ being a random vector taking values in $\mathbb{A}^{kV}$ with independent, not necessarily identically distributed, components, we are not guaranteed the existence of an S such that $P(G_S(U, \vec{W}_k) \leq n) \geq P(G_{S'}(U, \vec{W}_k) \leq n)$ for all alternate strategies S'.

Lemma 6.2 *If $V - U > 0$, a stochastically dominant strategy does not always exist.*

Proof: Let $k = 1$, $V = 2$, $U = 1$ and $\mathbb{A} = \{0, 1, 2\}$. Let the distributions of $W_1^{(1)}$ and $W_1^{(2)}$ be

User 1	User 2
$P\left(W_1^{(1)} = 0\right) = 0.6$	$P\left(W_1^{(2)} = 0\right) = 0.5$
$P\left(W_1^{(1)} = 1\right) = 0.25$	$P\left(W_1^{(2)} = 1\right) = 0.4$
$P\left(W_1^{(1)} = 2\right) = 0.15$	$P\left(W_1^{(2)} = 2\right) = 0.1$

If a stochastically dominant strategy exists, its first guess must be user 1, string 0, i.e. $S(1, 0) = 1$, so that $P(G_S(1, \vec{W}_1) = 1) = 0.6$. Given this first guess, if it is right the inquisitor will stop guessing but if it is wrong then $P\left(W_1^{(1)} = 1 | W_1^{(1)} = 1\right) = 5/8$ while $P\left(W_1^{(2)} = 0\right) = 0.5$ so to maximize $P(G_S(1, \vec{W}_1) \leq 2)$, the second guess must be user 1, string 1, $S(1, 1) = 2$, so that $P(G_S(1, \vec{W}_1) \leq 2) = 0.85$.

An alternate strategy with $S(2, 0) = 1$ and $S(2, 1) = 2$, however, gives $P(G_{S'}(1, \vec{W}_1) = 1) = 0.5$ and $P(G_{S'}(1, \vec{W}_1) \leq 2) = 0.9$. While $P(G_S(1, \vec{W}_1) = 1) > P(G_{S'}(1, \vec{W}_1) = 1)$, $P(G_S(1, \vec{W}_1) \leq 2) < P(G_{S'}(1, \vec{W}_1) \leq 1)$ and so there is no strategy stochastically dominated by all others in this case.

■

Despite this lack of optimal strategy, we will prove that as string-length grows, the asymptotic performance of any strategy can be lower bounded and that this lower bound is obtained by an explicit strategy.

6.3 Asymptotically optimal strategies

Let $\{\vec{W}_k\}$ be a sequence of random strings, with $\vec{W}_k$ taking values in $\mathbb{A}^{kV}$, with independent components, $W_k^{(v)}$, corresponding to strings selected by users 1 through V, although each user's string not be constructed from i.i.d. characters. For each individual user, $v \in \{1, \ldots, V\}$ let $G^{(v)}$ denote its optimal strategy. We will show that following is a stochastically dominated lower bound on the guesswork distribution of all strategies:

$$G_{\text{opt}}(U, \vec{W}_k) = \text{U-min}\left(G^{(1)}\left(W_k^{(1)}\right), \ldots, G^{(V)}\left(W_k^{(V)}\right)\right). \tag{6.2}$$

This can be thought of as allowing the inquisitor to query, for each n in turn, the n^{th} most likely string for all users while only accounting for a single guess and thus it does not correspond to a valid strategy.

Lemma 6.3 *For any strategy S and any $U \in \{1, \ldots, V\}$, $G_{opt}(U, \vec{W}_k)$ is stochastically dominated by $G_S(U, \vec{W}_k)$. That is, for any any $U \in \{1, \ldots, V\}$ and any $n \in \{1, \ldots, m^k\}$*

$$P(G_{opt}(U, \vec{W}_k) \leq n) \geq P(G_S(U, \vec{W}_k) \leq n).$$

Proof: For any strategy S,

$$G_S(U, \vec{w}_k) \geq \text{U-min}\left(N_S\left(1, S\left(1, w_k^{(1)}\right)\right), \ldots, N_S\left(V, S\left(V, w_k^{(V)}\right)\right)\right).$$

As for each $v \in \{1, \ldots, V\}$, $G^{(v)}(W_k^{(v)})$ is stochastically dominated by all other strategies,

$$P\left(G^{(v)}\left(W_k^{(v)}\right) \leq n\right) \geq P\left(N_S\left(v, S\left(1, W_k^{(v)}\right)\right) \leq n\right),$$

which, using equation (6.2), implies that

$$\begin{aligned} P(G_{\text{opt}}(U, \vec{W}_k) \leq n) &\geq P(\text{U-min}(N_S(1, S(1, W_k^{(1)})), \ldots, N_S(V, S(V, W_k^{(V)}))) \leq n) \\ &\geq P(G_S(U, \vec{W}_k) \leq n) \end{aligned}$$

as required.

■

The strategy that we construct that will asymptotically meet the performance of the lower bound is to query the most likely string of each user in a round-robin fashion, followed by the second most likely, and so on. An upper bound on this strategy's performance is to consider only stopping at the end of a round of such queries, even if they reveal more than U strings, which gives

$$VG_{\text{opt}}(U, \vec{W}_k), \tag{6.3}$$

where $G_{\text{opt}}(U, \vec{W}_k)$ is defined in (6.2).

In large deviations parlance the stochastic processes $\{k^{-1} \log G_{\text{opt}}(U, \vec{W}_k)\}$ and $\{k^{-1} \log(VG_{\text{opt}}(U, \vec{W}_k))\}$ arising from equations (6.2) and (6.3) are asymptotically equivalent, e.g. Section 4.2.2 [16], as $\lim_{k\to\infty} k^{-1} \log V = 0$. As a result, if one satisfies the LDP then the other does and thus it proves sufficient to establish the large deviation properties of behavior of $\{k^{-1} \log G_{\text{opt}}(U, \vec{W}_k)\}$ in order to determine those of this asymptotically optimal strategy.

6.4 Asymptotic performance

We begin by assuming that the guesswork of individual users possess properties that have been established to hold in substantial generality.

For each individual user $v \in \{1, \ldots, V\}$, denote the specific Rényi entropy by $R^{(v)}(\beta)$.

For each $v \in \{1, \ldots, V\}$, the scaled Cumulant Generating Function (sCGF) of $\{k^{-1} \log G^{(v)}(W_k^{(v)})\}$ exists and can be identified in terms of specific Rényi entropy:

$$\begin{aligned} \Lambda_G^{(v)}(\alpha) &= \lim_{k\to\infty} \frac{1}{k} \log E\left(\exp\left(\alpha \log G^{(v)}(W_k^{(v)})\right)\right) \\ &= \begin{cases} \alpha R^{(v)}\left(\dfrac{1}{1+\alpha}\right) & \text{if } \alpha > -1 \\ -R^{(v)}(\infty) & \text{if } \alpha \leq -1. \end{cases} \end{aligned} \tag{6.4}$$

If, in addition, $R^{(v)}(\beta)$ is differentiable and has a continuous derivative, it is established in Chapter 3 that the process $\{k^{-1} \log G^{(v)}(W_k^{(v)})\}$ satisfies a Large Deviation Principle

with a convex rate function

$$\Lambda_G^{(v)^*}(x) = \sup_{\alpha \in \mathbb{R}} \left(x\alpha - \Lambda_G^{(v)}(\alpha) \right). \tag{6.5}$$

This LDP is used to deduce the the following approximation

$$P(G^{(v)}(W_k^{(v)}) = n) \approx \frac{1}{n} \exp\left(-k\Lambda_G^{(v)^*}\left(\frac{1}{k} \log n \right) \right) \tag{6.6}$$

for large k and $n \in \{1, \ldots, m^k\}$.

The following theorem establishes the fundamental analogues of these results for an optimal strategy in the setting where user strings may have distinct probabilistic properties.

Theorem 6.4 *Assume that the components of $\{\vec{W}_k\}$ are independent and that for each $v \in \{1, \ldots, V\}$ $R^{(v)}(\beta)$ exists for all $\beta > 0$, is differentiable and has a continuous derivative, and that equation (6.4) holds. Then the process $\{k^{-1} \log G_{opt}(U, \vec{W}_k)\}$, and thus any asymptotically optimal strategy, satisfies a large deviation principle. Defining*

$$\delta^{(v)}(x) = \begin{cases} \Lambda_G^{(v)^*}(x) \text{ if } x \leq R^{(v)}(1) \\ 0 \text{ otherwise,} \end{cases}$$

$$\text{and } \gamma^{(v)}(x) = \begin{cases} \Lambda_G^{(v)^*}(x) \text{ if } x \geq R^{(v)}(1) \\ 0 \text{ otherwise,} \end{cases}$$

the rate function is

$$I_{G_{opt}}(U, V, x) = \min_{v_1, \ldots, v_V} \left(\Lambda_G^{(v_1)^*}(x) + \sum_{i=2}^{U} \delta^{(v_i)}(x) + \sum_{i=U+1}^{V} \gamma^{(v_i)}(x) \right), \tag{6.7}$$

which is lower semi-continuous and has compact level sets, but may not be convex. The sCGF capturing how the moments scale is

$$\begin{aligned} \Lambda_{G_{opt}}(U, V, \alpha) &= \lim_{k \to \infty} \frac{1}{k} \log E(\exp(\alpha \log G_{opt}(U, \vec{W}_k))) \\ &= \sup_{x \in [0, Vm]} \left(\alpha x - I_{G_{opt}}(U, V, x) \right). \end{aligned} \tag{6.8}$$

PROOF: Under the assumptions of the theorem, for each $v \in \{1, \ldots, V\}$, $\{k^{-1} \log G^{(v)}(W_k^{(v)})\}$ satisfies the LDP with the rate function given in equation (3.4). As users' strings are selected independently, the sequence of vectors

$$\left\{ \left(\frac{1}{k} \log G^{(1)}(W_k^{(1)}), \ldots, \frac{1}{k} \log G^{(V)}(W_k^{(V)}) \right) \right\}$$

satisfies the LDP in $\mathbb{R}^V$ with rate function $I(y^{(1)}, \ldots, y^{(V)}) = \sum_{v=1}^{V} {\Lambda_G^{(v)}}^*(y^{(v)})$, the sum of the rate functions given in equation (3.4).

Within our setting, the contraction principle, e.g. Theorem 4.2.1 [16], states that if a sequence of random variables $\{X_n\}$ taking values in a compact subset of $\mathbb{R}^V$ satisfies a LDP with rate function $I : \mathbb{R}^V \mapsto [0, \infty]$ and $f : \mathbb{R}^V \mapsto \mathbb{R}$ is a continuous function, then the sequence $\{f(X_n)\}$ satisfies the LDP with rate function $\inf_{\vec{y}}\{I(\vec{y}) : f(\vec{y}) = x\}$.

Assume, without loss of generality, that $\vec{x} \in \mathbb{R}^V$ is such that $x^{(1)} \leq x^{(2)} < \cdots \leq x^{(V)}$, so that U-min$(\vec{x}) = x^{(U)}$. Next define $|\vec{x} - \vec{y}| = min_i |x^{(i)} - y(i)|$. Note that $\vec{y}$ may not have the same ordering as $\vec{x}$. Now $|\vec{x} - \vec{y}| < \delta$ implies that $y^{(1)}, \ldots, y^{(U)} \leq$ U-min$(\vec{x}) + \delta$ so that U-min$(\vec{y}) \leq$ U-min$(\vec{x}) + \delta$. Similarly $y^{(U)}, \ldots, y^{(V)} \geq$ U-min$(\vec{x}) - \delta$ so that U-min$(\vec{y}) \geq$ U-min$(\vec{x}) - \delta$. This implies that for any $\epsilon > 0$ there exists a δ such that $|\vec{x} - \vec{y}| < \delta$ implies that $|$U-min$(\vec{y}) -$ U-min$(\vec{x})| < \epsilon$. Hence U-min : $\mathbb{R}^V \to \mathbb{R}$ is a continuous function and that a LDP holds follows from an application of the contraction principle, giving the rate function

$$I_{G_{\text{opt}}}(U, V, x) = \inf \left\{ \sum_{v=1}^{V} {\Lambda_G^{(v)}}^*(y_v) : \text{U-min}(y_1, \ldots, y_V) = x \right\}.$$

This expression simplifies to that in equation (6.7) as there has to be $U - 1$ of the y_is such that $y_v \leq x$, to minimise ${\Lambda_G^{(v)}}^*(y_v)$ under the condition that $y_v \leq x$ we set to 0 if possible under this condition or if not we use the fact that ${\Lambda_G^{(v)}}^*(y_v)$ is decreasing if a such that ${\Lambda_G^{(v)}}^*(a) = 0$ has $a > x$. This gives the $\delta^{(v)}(x)$ while a similar argument holds for $\gamma^{(v)}(x)$. Next we need at least one i such that $y_i = x$. Finally as each user may be picking using a different process we take the minimum over all possible combinations. The sCGF result follows from an application of Varahadan's Lemma, e.g [16, Theorem 4.3.1].

■

The expression for the rate function in equation (6.7) lends itself to a useful interpretation. In the long string-length asymptotic, the likelihood that an inquisitor has identified U of the V users' strings after approximately $\exp(kx)$ queries is contributed to by three distinct groups of identifiable users. For given x, the argument in the first term (v_1) identifies the last of the U users whose string is identified. The second summed term is contributed to by the collection of users, (v_2) to (v_U), whose strings have already been identified prior to $\exp(kx)$ queries, while the final summed term corresponds to those users, (v_{U+1}) to (v_V), whose strings have not been identified.

The reason for using the notation $I_{G_{\text{opt}}}(U, V, \cdot)$ in lieu of $\Lambda^*_{G_{\text{opt}}}(U, V, \cdot)$ for the rate function in Theorem 6.4 is that $I_{G_{\text{opt}}}(U, V, \cdot)$ is not convex in general, which we will demonstrate by example, and so is not always the Legendre-Fenchel transform of the sCGF $\Lambda_{G_{\text{opt}}}(U, V\cdot)$. Instead

$$\Lambda^*_{G_{\text{opt}}}(U, V, x) = \sup_{\alpha} \left(\alpha x - \Lambda_{G_{\text{opt}}}(U, V, \alpha)\right)$$

forms the convex hull of $I_{G_{\text{opt}}}(U, V, \cdot)$. In particular, this means that we could not have proved Theorem 6.4 by establishing properties of $\Lambda_{G_{\text{opt}}}(U, V, \cdot)$ alone, which was the successful route taken for the $U = V = 1$ setting, and instead needed to rely on the LDP proved in Chapter 3.

The potential lack of convexity in the rate function of Theorem 6.4, equation (6.7), only arises if users' string statistics are asymptotically distinct. The significance of this lack of convexity on the phenomenology of guesswork can be understood in terms of the asymptotically optimal round-robin strategy: if the rate function is not convex, there is no single set of users whose strings are most vulnerable. That is, if U strings are recovered after a small number of guesses, they will be from one set of users, but after a number of guesses corresponding to a transition from the initial convexity they will be from another set of users. This is made explicit in the following corollary to Theorem 6.4.

Corollary 6.1 *If $I_{G_{opt}}(U, V, x)$ is not convex in x, then there is there is no single set of users whose strings will be identified in the long string length asymptotic.*

PROOF: We prove the result by establishing the converse: if a single set of users is always most vulnerable, then $I_{G_{\text{opt}}}(U, V, x)$ is convex. Recall the expression for $I_{G_{\text{opt}}}(U, V, x)$

given in equation (6.7)

$$I_{G_{\text{opt}}}(U,V,x) = \max_{v_1,\ldots,v_V} \left({\Lambda_G^{(v_1)}}^*(x) + \sum_{i=2}^{U} \delta^{(v_i)}(x) + \sum_{i=U+1}^{V} \gamma^{(v_i)}(x) \right),$$

As explained after Theorem 6.4, for given x the set of users $\{(v_1),\ldots,(v_U)\}$ corresponds to those users whose strings, on the scale of large deviations, will be identified by the inquisitor after approximately $\exp(kx)$ queries. If this set is unchanging in x, i.e. the same set of users is identified irrespective of x, then both of the functions

$$\left({\Lambda_G^{(v_1)}}^*(x) + \sum_{i=2}^{U} \delta^{(v_i)}(x) \right) \text{ and } \sum_{i=U+1}^{V} \gamma^{(v_i)}(x)$$

are sums of functions that are convex in x, and so are convex themselves. Thus the sum of them, $I_{G_{\text{opt}}}(U,V,x)$, is convex.

■

This is most readily illustrated by an example that falls within the two-user setting considered in [42, 26] with $U = 1$, $V = 2$ and one of the strings is chosen uniformly, while the authors directly identify $\Lambda_{G_{\text{opt}}}(1,2,\alpha)$ for $\alpha > 0$, one cannot establish a full LDP from this approach as the resulting rate function is not convex.

For an explicit illustration, that falls within the setting in [42], let $\mathbb{A} = \{0,\ldots,7\}$, $U = 1$ and $V = 2$. Assume both sources are i.i.d., with

$$P(W_1^{(1)} = i) = \begin{cases} 1/2 & \text{if } i \in \{0,1\} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } P(W_1^{(2)} = i) = \begin{cases} 0.55 & \text{if } i = 0 \\ 0.1 & \text{if } i \in \{1,2\} \\ 0.05 & \text{if } i \in \{3,\ldots,7\} \end{cases}$$

For these values, Figures 6.1 and 6.2 plot the rate functions for guessing each of the user's strings individually as well as the rate function for guessing one out of two, which is simply the minimum of the two rate function when they are finite. Taking the Legendre Fenchel transform of the sCGF results in the convex hull of this non-convex function

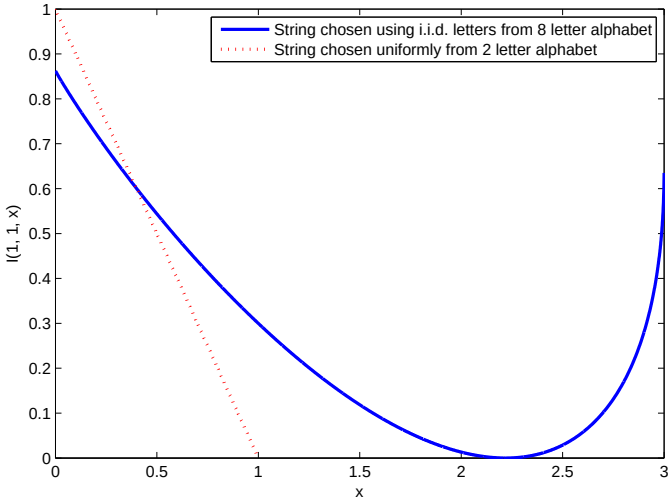

Figure 6.1: User 1 picks a string uniformly using the alphabet $\{0, 1\}$ and User 2 picks each character of a string in an i.i.d. fashion using $P(W_1 = 0) = 0.55$, $P(W_1 = 1) = P(W_1 = 2) = 0.1$ and $P(W_1 = 3) = P(W_1 = 4) = P(W_1 = 5) = P(W_1 = 6) = P(W_1 = 7) = 0.05$. The figure displays the rate function for guessing each user's strings individually.

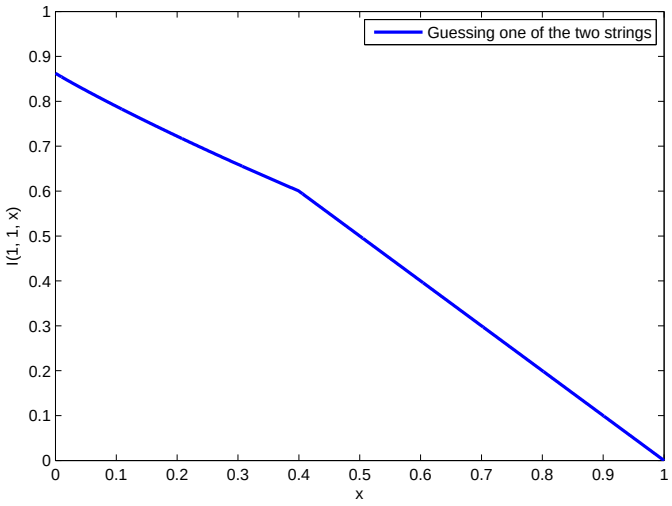

Figure 6.2: User 1 picks a string uniformly using the alphabet $\{0, 1\}$ and User 2 picks each character of a string in an i.i.d. fashion using $P(W_1 = 0) = 0.55$, $P(W_1 = 1) = P(W_1 = 2) = 0.1$ and $P(W_1 = 3) = P(W_1 = 4) = P(W_1 = 5) = P(W_1 = 6) = P(W_1 = 7) = 0.05$. The figure displays $I_{G_{\text{opt}}}(1, 2, x)$.

and so, while the sCGF correctly captures how the asymptotic moments scale, it does not contain sufficient information to establish the LDP, explaining the necessity for the distinct approach taken here.

Convexity is ensured, however, if all users select using the same stochastic properties. Indeed, the results in Theorem 6.4 simplify greatly and we have the following corollary.

Corollary 6.2 *If, in addition to the assumptions of Theorem 6.4,* $\Lambda_G^{(v)}(\cdot) = \Lambda_G(\cdot)$ *for all* $v \in \{1, \dots, V\}$ *with corresponding Rényi entropy* R*, then the rate function in equation* (6.5) *simplifies to the convex function*

$$\Lambda^*_{G_{opt}}(U, V, x) = \begin{cases} U\Lambda_G{}^*(x) & \text{if } x \leq R(1) \\ (V - U + 1)\Lambda_G{}^*(x) & \text{if } x \geq R(1) \end{cases} \tag{6.9}$$

where $R(1)$ *is the specific Shannon entropy, and the sCGF in equation* (6.8) *simplifies to*

$$\Lambda_{G_{opt}}(U, V, \alpha) = \begin{cases} U\Lambda_G\left(\dfrac{\alpha}{U}\right) & \text{if } \alpha \leq 0 \\ (V - U + 1)\Lambda_G\left(\dfrac{\alpha}{V - U + 1}\right) & \text{if } \alpha \geq 0. \end{cases} \tag{6.10}$$

In particular, with $\alpha = 1$ *we have*

$$\begin{aligned} &\lim_{k\to\infty} \frac{1}{k} \log E\left(G_{opt}(U, \vec{W}_k)\right) = \Lambda_{G_{opt}}(1) \\ &= (V - U + 1)\Lambda_G\left(\frac{1}{V - U + 1}\right) = R\left(\frac{V - U + 1}{V - U + 2}\right), \end{aligned} \tag{6.11}$$

which is a convex, decreasing function of $V - U$*.*

Proof: The simplification follows readily from equation (6.7). To establish convexity, using equation (6.11) it suffices to show that $x\Lambda_G(1/x)$ is convex for $x > 0$. This can be seen by noting that for any $a \in (0, 1)$ and $x_1, x_2 > 0$,

$$\begin{aligned} (ax_1 + (1-a)x_2)\Lambda_G\left(\frac{1}{ax_1 + (1-a)x_2}\right) &= (ax_1 + (1-a)x_2)\Lambda_G\left(\eta\frac{1}{x_1} + (1-\eta)\frac{1}{x_2}\right) \\ &\leq ax_1\Lambda_G\left(\frac{1}{x_1}\right) + (1-a)x_2\Lambda_G\left(\frac{1}{x_2}\right), \end{aligned}$$

where $\eta = ax_1/(ax_1 + (1-a)x_2) \in (0, 1)$ and we have used the convexity of Λ_G. That

$R(x) \downarrow R(1)$ as $x \uparrow 1$ is a general property of specific Rényi entropy, and so the monotonicity follows.

■

As the growth rate, $R((V-U+1)/(V-U+2))$, is convex and decreasing in $V-U$, there is a law of diminishing returns where the greatest decrease in the average guesswork growth rate is through the provision of one additional user. Note that in these results we cannot take the limit as $(V-U) \to \infty$ as to do so would involve an interchange of limits. As $R\,(x)$ is greater than the specific Shannon entropy of the source for all $x < 1$, however, in the multi-user setting the specific Shannon entropy of the source is a universal lower bound on the exponential growth rate of the expected guesswork that is tight for large $V-U$.

Regardless of whether the rate function $I_{G_{\text{opt}}}(U,V,\cdot)$ is convex, the following lemma justifies the approximation

$$P(G_{\text{opt}}(U,\vec{W}_k)=n) \approx \frac{1}{n}\exp\left(-kI_{G_{\text{opt}}}\left(U,V,\frac{1}{k}\log n\right)\right) \tag{6.12}$$

for large k and $n \in \{1,\ldots,m^k\}$. Equation (6.12) is analogous to that in equation (6.6) in Chapter 3, but there are additional difficulties that must be overcome to establish it. In particular, if $U=V=1$, the likelihood that the string is identified is decreasing per guess, but this is not true in the more general case. As a simple example, consider $U=V=2$, $\mathbb{A}=\{0,1\}$, strings of length 1 and strings chosen uniformly. Here the probability of guessing both strings in one guess is $1/4$, but at the second guess the probability is $3/4$. Despite this lack of monotonicity, the approximation still holds in the following sense.

Lemma 6.5 *Under the assumptions of Theorem 6.4, for any $x \in [0,\log(m))$ we have*

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\liminf_{k\to\infty}\frac{1}{k}\log\inf_{n\in K_k(x,\epsilon)}P(G_{opt}(U,\vec{W}_k)=n)\\
&=\lim_{\epsilon\downarrow 0}\limsup_{k\to\infty}\frac{1}{k}\log\sup_{n\in K_k(x,\epsilon)}P(G_{opt}(U,\vec{W}_k)=n)\\
&=-I_{G_{opt}}(U,x)-x,
\end{aligned}$$

where

$$K_k(x,\epsilon) = \{n : n \in (\exp(k(x-\epsilon)), \exp((k(x+\epsilon)))\}$$

is the collection of guesses made in a log-neighborhood of x.

PROOF: The proof follows the ideas in Chapter 3, Corollary 3.1, but with the added difficulty resolved. The trick being to isolate the last string that is likely to be guessed and to use the monotonicity of its individual likelihood of being identified.

Consider for $x \in (0, \log(m))$

$$\begin{aligned}
&\sup_{n\in K_k(x,\epsilon)} P(G_{\text{opt}}(U,\vec{W}_k) = n) = \sup_{n\in K_k(x,\epsilon)} \sum_{(v_1,\ldots,v_V)} P(G^{(v_1)}(W_k^{(v_1)}) = n) \\
&\prod_{i=2}^{u} P(G^{(v_i)}(W_k^{(v_i)}) \le n) \prod_{i=u+1}^{V} P(G^{(v_i)}(W_k^{(v_i)}) \ge n) \\
&\le \sup_{n\in K_k(x,\epsilon)} \max_{(v_1,\ldots,v_V)} (V!) P(G^{(v_1)}(W_k^{(v_1)}) = n) \\
&\prod_{i=2}^{u} P(G^{(v_i)}(W_k^{(v_i)}) \le n) \prod_{i=u+1}^{V} P(G^{(v_i)}(W_k^{(v_i)}) \ge n) \\
&\le \sup_{n\in K_k(x,\epsilon)} \max_{(v_1,\ldots,v_V)} (V!) P(G^{(v_1)}(W_k^{(v_1)}) = n) \\
&\prod_{i=2}^{u} P\left(\frac{1}{k}\log G^{(v_i)}(W_k^{(v_i)}) \le x+\epsilon\right) \prod_{i=u+1}^{V} P\left(\frac{1}{k}\log G^{(v_i)}(W_k^{(v_i)}) \ge x-\epsilon\right) \\
&\le \inf_{n\in K_k(x-2\epsilon,\epsilon)} \max_{(v_1,\ldots,v_V)} (V!) P\left(\frac{1}{k}\log G^{(v_1)}(W_k^{(v_1)}) = n\right) \\
&\prod_{i=2}^{u} P\left(\frac{1}{k}\log G^{(v_i)}(W_k^{(v_i)}) \le x+\epsilon\right) \prod_{i=u+1}^{V} P\left(\frac{1}{k}\log G^{(v_i)}(W_k^{(v_i)}) \ge x-\epsilon\right).
\end{aligned}$$

The first equality holds by definition of $G_{\text{opt}}(U,\cdot)$. The first inequality follows from the union bound over all possible permutations of $\{1,\ldots,V\}$. The second inequality utilizes $k^{-1}\log n \in (x-\epsilon, x+\epsilon)$ if $n \in K_k(x,\epsilon)$, while the third inequality uses the monotonic decreasing probabilities in guessing a single user's string.

Taking $\lim_{\epsilon\downarrow 0} \limsup_{k\to\infty} k^{-1}\log$ on both sides of the inequality, interchanging the order of the max and the supremum, using the continuity of $\Lambda_G^{(v)}(\cdot)$ for each $v \in \{1,\cdots,V\}$,

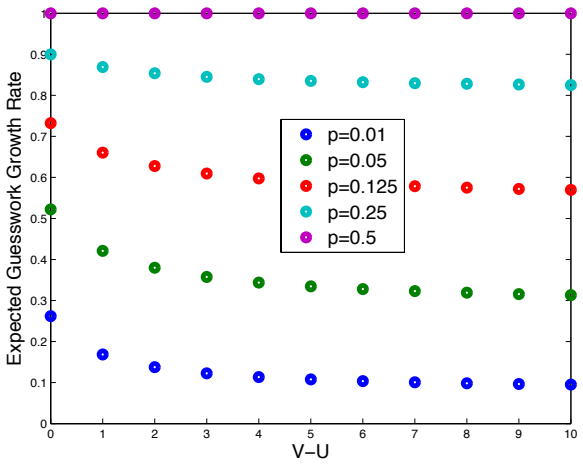

Figure 6.3: Binary alphabet, $\mathbb{A} = \{0, 1\}$, Bernoulli selection with $P(W_1 = 1) = p$ in figure legend. Average guesswork growth rate as a function of $V - U$, the excess number of guessable strings.

and the representation of the rate function $I_{G_{\text{opt}}}(U, V, \cdot)$ in equation (6.7), gives the upper bound

$$\lim_{\epsilon \downarrow 0} \limsup_{k \to \infty} \frac{1}{k} \log \sup_{n \in K_k(x,\epsilon)} P(G_{\text{opt}}(\vec{W}_k) = n) \leq -I_{G_{\text{opt}}}(U, V, x) - x.$$

Considering the least likely guesswork in the ball leads to a matching lower bound. The other case, $x = 0$, follows similar logic, leading to the result.

■

6.5 Examples

To illustrate the reduction in computational security that comes from having multiple users, in Figure 6.3 the average guesswork growth rate for an asymptotically optimal strategy is plotted for the simplest case, a binary alphabet and an i.i.d. Bernoulli string source for each user. The x-axis is the excess number of guessable strings, $V - U$, and the y-axis is the $\log_2$ growth rate. If the source is perfectly uniform (i.e. characters are chosen with a Bernoulli $1/2$ process), then the average guesswork growth rate is maximal

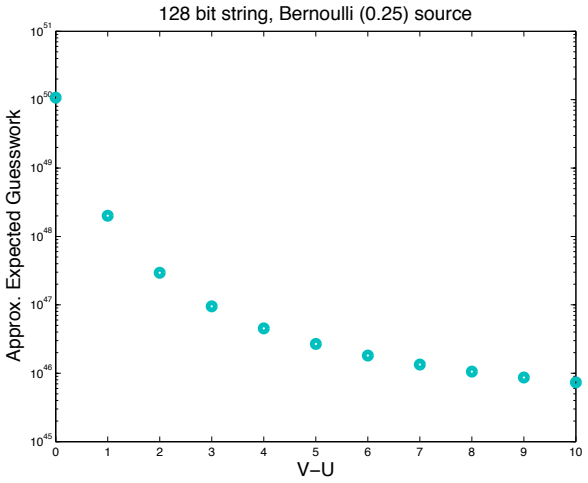

Figure 6.4: Binary alphabet, $\mathbb{A} = \{0, 1\}$, Bernoulli source with $P(W_1 = 1) = 0.25$. Approximate average guesswork for a 128 bit string as a function of $V - U$, the excess number of guessable strings.

and unchanging in $V - U$. If the source is not perfectly uniform, then the growth rate decreases as the number of excess guessable strings $V - U$ increases, with a lower bound of the source's Shannon entropy.

For a string of length 128 bits and a Bernoulli (0.25) source, Figure 6.4 displays the impact that the change in this exponent has, approximately, on the average number of guesses required to determine U strings.

In Chapter 3 it is shown that $\Lambda(\cdot)$ is constant for $\alpha < -1$ and increasing thereafter. The derivative of Λ is continuous for $\alpha > -1$ and 0 for $\alpha < -1$, though it may not exist at $\alpha = -1$. Something similar holds for the multi-user case, though with a discontinuous derivative at $\alpha = -U$.

In Figure 6.5, $\mathbb{A} = \{0, 1, 2\}$ is used with each character in a string picked in an i.i.d. fashion and $P(W_1 = 0) = P(W_1 = 1) = 0.4$, $P(W_1 = 2) = 0.2$ for each user. For the $U = 1, V = 2$ case it agrees with Chapter 3 in that the discontinuity in that the derivative of the sCGF occurs at $\alpha = -1$ and $\Lambda(1, 2, \alpha)$ is constant for $\alpha \leq -1$, but the other two cases illustrate that this is dependent on U and not on V in the multi-user case. It can be seen that the $U = 2, V = 2$ and $U = 2, V = 4$ both display the discontinuity in the derivative of the sCGF referred to in Chapter 3 at $\alpha = -U$ as opposed to -1 and both

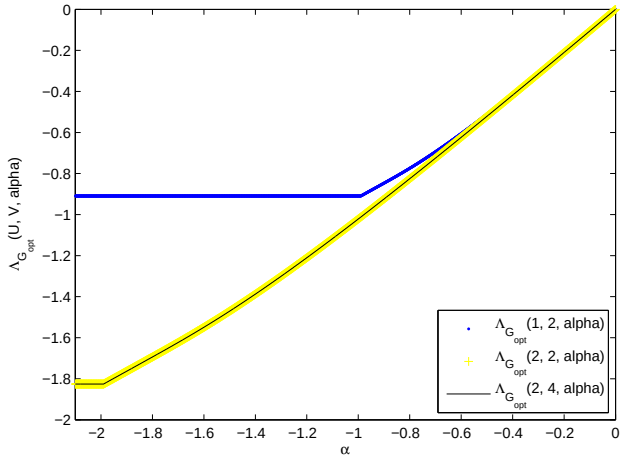

Figure 6.5: Trinary alphabet, $\mathbb{A} = \{0, 1, 2\}$, Bernoulli string selection with $P(W_1 = 0) = P(W_1 = 1) = 0.4$, $P(W_1 = 2) = 0.2$. Illustrates the difference in the rate functions for $U = 1, V = 2$, $U = 2, V = 2$ and $U = 2, V = 4$ for $\alpha \leq 0$. This shows that the value below which the sCGF is constant is dependent on U but not V.

a constant below this point.

Throughout the rest of this section we shall use $\mathbb{A} = \{0, 1\}$, assume that all users have the same source statistics, and revert to using log base 2. In this case, the maximum average guesswork growth rate is 1.

We now consider a number of metrics including the ultimate security gap between having one user and having many, $R(1/2) - R(1)$ and the one additional user security gap, $R(1/2) - R((V - U + 1)/(V - U + 2))$. These measure the drop in average guesswork growth rate possible from having an arbitrarily large number of users and having $V - U$ additional users, respectively.

For Bernoulli(p) sources, while Figure 6.3 shows how the security gaps change for fixed p as a function of the excess number of guessable strings, $V - U$, Figure 6.6 shows how for fixed $V - U$ the gap behaves as p is changed.

If that source statistics of all strings are governed by a Markov chain with transition

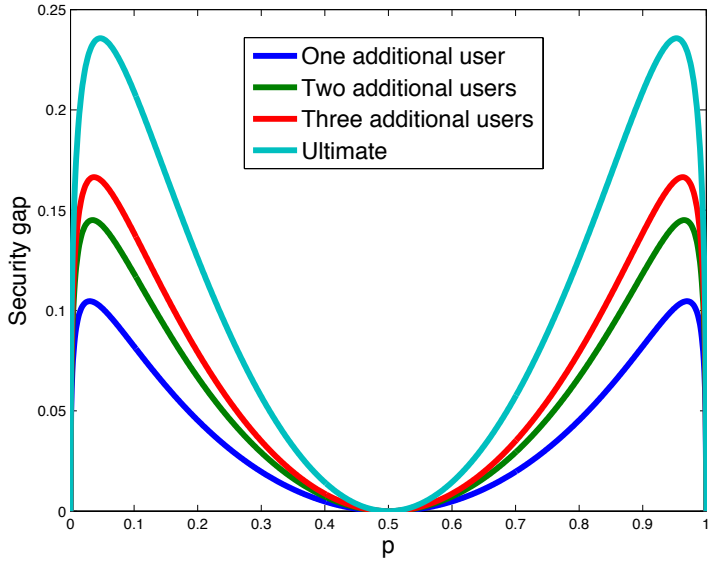

Figure 6.6: Binary alphabet, $\mathbb{A} = \{0, 1\}$, Bernoulli string selection with $P(W_1 = 1) = p$ in figure legend. Computational security gap for a range of $V - U$, the excess number of guessable strings.

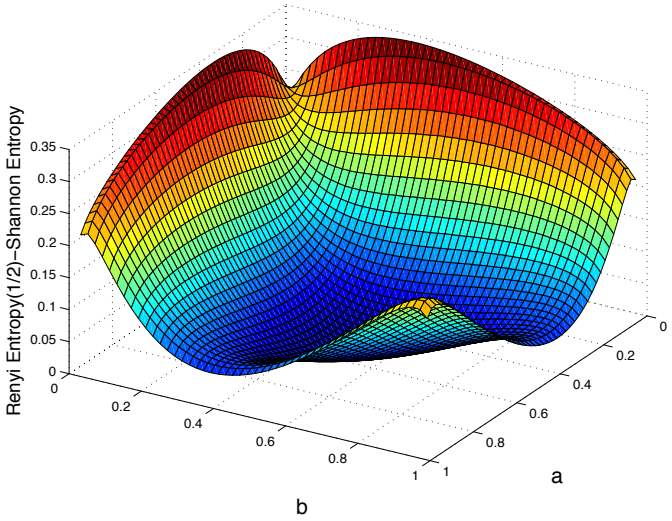

Figure 6.7: Binary alphabet, $\mathbb{A} = \{0, 1\}$, Markov string selection with $P(W_2 = 1|W_1 = 0) = a$ and $P(W_2 = 0|W_1 = 1) = b$ in figure legend. Computational security gap, Rényi entropy (1/2) less Shannon entropy, as a function of $V - U$, the excess number of guessable strings.

matrix

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix},$$

where $a, b \in (0,1)$, then the specific Rényi entropy can be evaluated for $\beta \neq 1$ to be

$$R(\beta) = \frac{1}{1-\beta} \log \left(\frac{(1-a)^\beta + (1-b)^\beta}{2} + \frac{\sqrt{((1-a)^\beta - (1-b)^\beta)^2 + 4(ab)^\beta}}{2} \right)$$

and $R(1)$ is the Shannon entropy

$$R(1) = \frac{b}{a+b} H(a) + \frac{a}{a+b} H(b),$$
$$\text{where } H(a) = -a \log(a) - (1-a) \log(1-a).$$

If $b = 1-a$, then the source is i.i.d. and if $b = 1-a = 0.5$ then the average guesswork is maximized with rate 1, as is the ultimate security gap. The ultimate security gap is plotted in Figure 6.7 as a and b vary, with largest gap being approximately 0.346 he one additional user security gap is plotted in Figure 6.8, which has a maximal value of approximately 0.156.

6.6 Discussion

Since Massey [39] posed the original guesswork problem and Arikan [1] introduced his long string asymptotic, generalizations have been used to quantify the computational security of several systems, including being related to questions of lossless compression. Here we have considered what appears to be one of the most natural extensions of that theory, that of multi-user computational security. As a consequence of the inherent non-convex nature of the guesswork rate function unless string source statistics are equal for all users, this development wasn't possible prior to the Large Deviation Principle proved in [10]. The results therein themselves relied on the earlier work that determined the scaled cumulant generating function for the guesswork for a broad class of process [1, 37, 46].

The fact that rate functions can be non-convex encapsulates that distinct subsets of users are likely to be identified depending on how many unsuccessful guesses have been made.

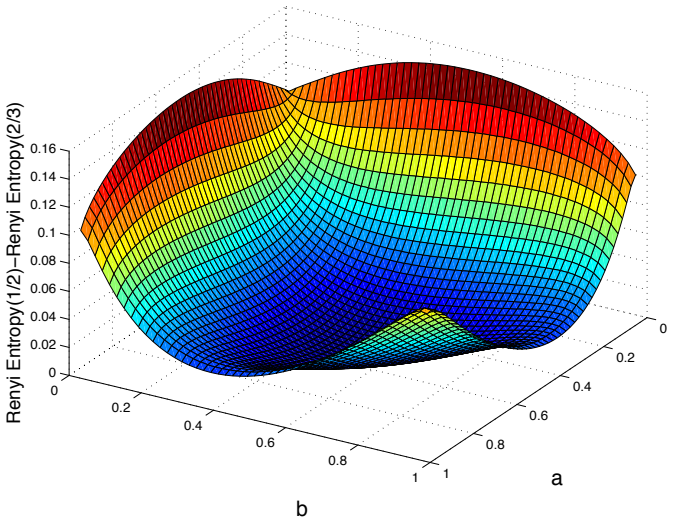

Figure 6.8: Binary alphabet, $\mathbb{A} = \{0, 1\}$, Markov string selection with $P(W_2 = 1|W_1 = 0) = a$ and $P(W_2 = 0|W_1 = 1) = b$ in figure legend. Computational security gap in having one extra hackable user: Rényi entropy $(1/2)$ less Rényi entropy $(2/3)$ as a function of $V - U$, the excess number of guessable strings.

As a result, a simple ordering of string guessing difficulty is inappropriate in multi-user systems and suggests that quantification of multi-user computational security is inevitably nuanced.

The original analysis of the asymptotic behavior of single user guesswork identified an operational meaning to specific Rényi entropy. In particular, the average guesswork grows exponentially in string length with an exponent that is the specific Rényi entropy of the character source with parameter $1/2$. When users' string statistics are the same, the generalization to multi-user guesswork identifies a surprising operational rôle for specific Rényi entropy with parameter $n/(n + 1)$ for each $n \in \mathbb{N}$ when n is the excess number of strings that can be guessed. Moreover, while the specific Shannon entropy of the string source was found in the single user problem to have an unnatural meaning as the growth rate of the expected logarithm of the guesswork, in the multi-user system it arises as the universal lower bound on the average guesswork growth rate.

For the asymptote at hand, the key message is that there is a law of diminishing returns for an inquisitor as the number of users increases. For a multi-user system designer, in contrast to the single character, single user system introduced in [39], Shannon entropy

is the appropriate measure of expected guesswork for systems with many users.

7 Reverse Guesswork

7.1 Introduction

Unlike previous chapters, this one is somewhat speculative, pointing towards future work and enhancements to the quality of estimates on the guesswork probability mass function. Building on earlier results, in Chapter 3 for single user guesswork it was proven that if $\{W_k\}$ is a stochastic process, with $W_k : \Omega \mapsto \mathbb{A}^k = \{0, \ldots, m-1\}^k$, in one of the classes considered by [1, 37, 46], and with G being an optimal single-source guesswork ordering, then the process

$$\left\{ \frac{1}{k} \log G(W_k) \right\} \tag{7.1}$$

satisfies a Large Deviation Principle (LDP) with rate function

$$I(x) = \sup_{\alpha \in \mathbb{R}} (\alpha x - \Lambda(\alpha)), \text{ where } \Lambda(\alpha) = \begin{cases} (1+\alpha) R\left(\frac{1}{1+\alpha}\right) & \text{if } \alpha > -1 \\ -R(\infty) & \text{if } \alpha \leq -1 \end{cases}$$

and R is the specific Rényi entropy of $\{W_k\}$

$$R(\beta) = \lim_{k \to \infty} \frac{1}{1-\beta} \log \sum_{w \in \mathbb{A}^k} P(W_k = w)^\beta.$$

Based on this result, in Corollary 3.1, the following direct estimate on the guesswork probability mass function (PMF) is proposed:

$$P(G(W_k) = n) \approx \frac{1}{n} \exp\left(-kI\left(k^{-1} \log n\right)\right), \text{ for } n \in \{1, \ldots, m^k\}. \tag{7.2}$$

While this estimate has formal backing and appears reasonable (e.g. Figure 7.2), we felt it could be improved upon. What follows below is a result in the simplest of cases, i.e. binary i.i.d. string sources, and a conjecture in that setting. A new LDP that will give a distinct approximation to the guesswork probability mass function will be formally established. From it, a heuristic and some evidence for a conjectured result is provided.

7.2 Reverse Guesswork

The key observation that suggests considering alternate approximations to the guesswork PMF relates to the non-linear scaling in eq. (7.1), i.e. the inner log. As a result of it, the LDP essentially provides information regarding

$$P\left(\frac{1}{k}\log G(W_k) \in (x-\epsilon, x+\epsilon)\right) = P\left(G(W_k) \in (e^{k(x-\epsilon)}, e^{k(x+\epsilon)})\right).$$

For x small, this contains information about the likelihood the guesswork is in a relatively small neighbourhood around $\exp(kx)$ guesses and, indeed, it is effectively perfect for $x = 0$. As x increases, however, it contains information about the likelihood the guesswork is in an exponentially expanding neighbourhood of guess numbers. In the extreme case where $x = \log|\mathbb{A}|$ it is almost inquiring "what is the likelihood the guesswork is in the second half of the strings?" and the estimate is, therefore, poor for any particular individual string in this second half.

Here we provide an initial approach at obtaining an estimate where the scaling focuses on a different part of the guesswork PMF. We do this by considering guessing in the pessimal order, the absolute reverse of an optimal order, which we dub reverse guesswork. Consider an inquisitor that asks questions from the least likely to most likely. If $w_k^i \in \mathbb{A}^k$ for $i = 1, \ldots, m^k$, let

$$G(w_k^1) > G(w_k^2) > \cdots > G(w_k^{m^k})$$

denote an optimal guessing order. The inquisitor reverses the order and defines the new order

$$G^R(w_k^1) < G^R(w_k^2) < \cdots < G^R(w_k^{m^k}).$$

Establishing a LDP for the reverse guesswork,

$$\left\{\frac{1}{k}\log G^R(W_k)\right\},$$

provides a bad approximation for likely strings, but a finer estimate for highly unlikely strings. Better estimates of the likelihood of these unlikely words may be of interest, for example, in understanding lossless Huffman codes.

Analysing reverse guesswork turns out to be a challenge for one main reason, which is also what makes it technically interesting: the resulting rate function will be concave and so dual methods, determining how the moments scale, cannot tell us how the probabilities scale.

As we can't start via the sCGF route, we need a more direct argument about the probabilities themselves.

7.3 Reverse Guesswork and i.i.d binary sources

We restrict to i.i.d. binary sources, $\mathbb{A} = \{0, 1\}$, and denote $P(W_1 = 0) = p > 1/2$, for which we have the following, whose proof appears in Section 7.4.

Theorem 7.1 *Define the (pseudo-)inverse*

$$H^{-1}(x) := \max\{b \in [0, 1] : H(b) = x\}, \tag{7.3}$$

where $H(b) = -b \log b - (1 - b) \log(1 - b)$ is the binary Shannon entropy. Then guesswork $\{k^{-1} \log G(W_k)\}$ satisfies the LDP with a rate function that admits the following characterisation:

$$I^H(x) = -x - H^{-1}(x) \log p - (1 - H^{-1}(x)) \log(1 - p) \text{ for } x \in [0, \log(2)]. \tag{7.4}$$

The reverse guesswork, $\{k^{-1} \log G^R(W_k)\}$, satisfies the LDP with rate function

$$I^R(x) = -x - (1 - H^{-1}(x)) \log p - H^{-1}(x) \log(1 - p) \text{ for } x \in [0, \log(2)], \tag{7.5}$$

which is strictly concave.

For $p = 0.8$, Figure 7.1 plots both the rate function for guesswork $I^H(x)$ and reverse guesswork $I^R(x)$ vs. x. The first thing to note is that $I^R(x)$ is concave and so while its Legendre Fenchel transform is the scaled Cumulant Generating Function (sCGF), it is not the Legendre Fenchel transform of the sCGF. Indeed, the double Legendre-Fenchel transform would be the rate function's convex hull, which in this case is a straight line. Thus the approach taken heretofore in the study of guesswork is not of use for this

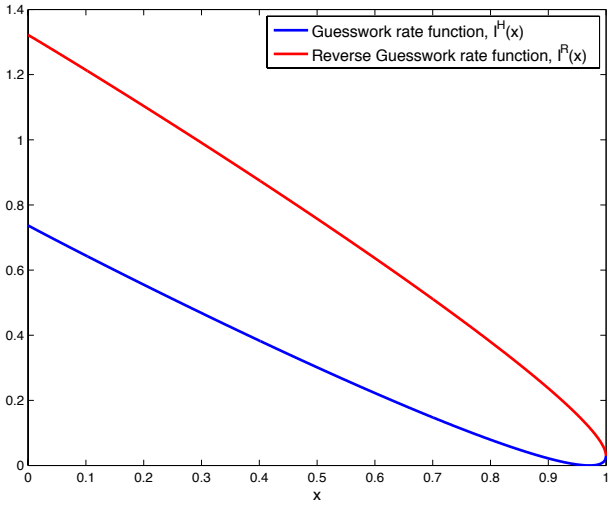

Figure 7.1: Guesswork rate function, $I^H(x)$, and reverse guesswork rate function, $I^R(x)$, for a binary i.i.d. source and $p = P(W_1 = 0) = 0.8$. Note that these are guessing in distinct orders: the least likely guess in guesswork is at $x = \log(2)$ while this is the most-likely guess in reverse guesswork. This is suggestive that the latter is picking up where the former finishes.

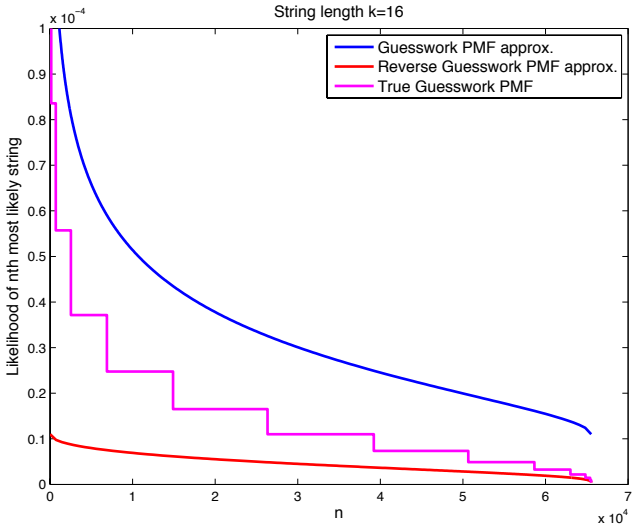

Figure 7.2: Guesswork and reverse guesswork PMF approximations, which do not consider any potential factors missed through the asymptote, versus the true guesswork PMF for $p = 0.8$ and string length $k = 16$.

process. The second thing to note is that guesswork is read most likely to least likely left-to-right on this figure, while reverse guesswork is the other way around. In some sense, it appears as if reverse guesswork takes up where the guesswork leaves off.

In the same way that the original LDP suggests an approximation to the guesswork PMF, Theorem 7.1 can be formally shown to suggest the following:

$$\begin{aligned} P(G(W_k) = 2^k - n + 1) &= P(G^R(W_k) = n) \\ &\approx \frac{1}{n} \exp(-kI^R(k^{-1} \log n)). \end{aligned} \tag{7.6}$$

For binary strings of length $k = 16$, Figure 7.2 plots the true guesswork PMF along with the two estimates obtained from guesswork and reverse guesswork asymptotes. One might imagine that the two in some way converge from above and below. If, however, one recalls our original observation that guesswork is providing estimates for the first half of guesses and reverse guesswork for the second half, one can look to see what would happen if there was a missing factor of 2 in our approximation. I.e. guesswork is giving us an approximation for the first half of the guesses and reverse guesswork for the latter half.

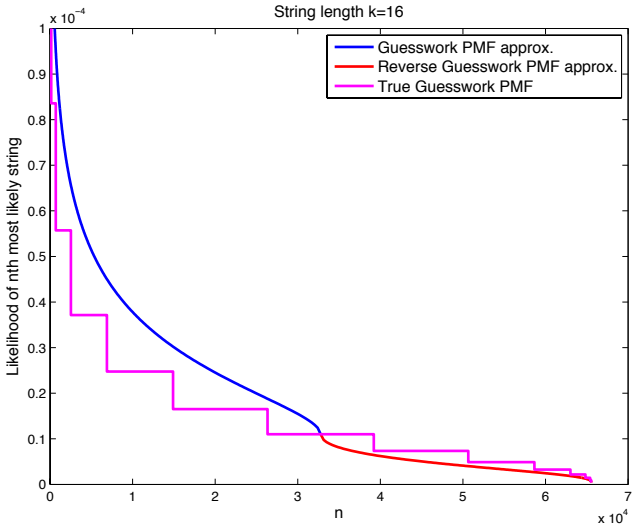

Figure 7.3: Conjecture of a better approximation guesswork and reverse guesswork PMF approximations that incorporate a conjectured factor that would be lost in the asymptotic scaling versus true guesswork PMF for $p = 0.8$ and string length $k = 16$.

To achieve this approximation we replace eq. (7.2) with the following for $n \in \{1, \ldots, 2^{k-1}\}$

$$P(G(W_k) = n) \approx \frac{1}{n} \exp(-kI(k^{-1} \log(2n))).$$

and eq. (7.6) for $n \in \{2^{k-1} + 1, 2^k\}$ with

$$\begin{aligned} P(G(W_k) = n) &= P(G^R(W_k) = 2^k - n + 1) \\ &\approx \frac{1}{n} \exp\left(-kI^R(k^{-1} \log(2(2^k - n + 1)))\right). \end{aligned}$$

To be clear, this conjectured factor is one that the asymptote couldn't capture. As presently formed, it would be eliminated by the scaling and so the approximation is as valid as the ones we've previously established. The result, however, is plotted Figure 7.3. This suggests to us that heuristic is correct and that if we wish to get a handle on the guessing probabilities of the unlikely strings, a deeper analysis of reverse guesswork would be desirable, for which the present results provide an initial investigation.

7.4 Sketch proof

The characterisation of the rate function for guesswork in terms of inverse Shannon entropy, equation (7.4), is most readily seen by algebraic manipulation. One can show I^H is convex and so to establish that

$$I^H(x) = I(x) = \sup_{\alpha \in \mathbb{R}} (\alpha x - \Lambda(\alpha)),$$

where

$$\Lambda(\alpha) = \begin{cases} (1+\alpha)\log\left(p^{1/(1+\alpha)} + (1-p)^{1/(1+\alpha)}\right) & \text{if } \alpha > -1 \\ \log(p) & \text{if } \alpha \leq -1 \end{cases},$$

it suffices to prove that

$$\Lambda^H(\alpha) = \sup_{x \in \mathbb{R}} (x\alpha - I^H(x)) = \Lambda(\alpha).$$

This can be achieved by calculus recalling the inverse differential rule. The strict concavity of I^R can also be shown directly by calculus.

The i.i.d. binary assumption gives us the following symmetry, which we rely on in the proof and so an alternate approach would need to be developed in general.

Lemma 7.2 *For a binary, i.i.d. source we have*

$$P(G(W_k) = n) = p^{ck}(1-p)^{(1-c)k}$$

for $c \in [0,1]$, if and only if

$$P(G^R(W_k) = n) = p^{(1-c)k}(1-p)^{ck}.$$

Proof: Let

$$C_j = \{w \in \{0,1\}^k : P(W_k = w) = p^{k-j}(1-p)^j\}$$

and note that in guesswork the sets are asked in order from C_1 to C_k, with the ordering of strings within them broken arbitrarily. In reverse guesswork, they are asked in the

order C_k to C_1, but the size of the set C_j is the same as that of C_{k-j},

$$|C_j| = \binom{k}{j} = \binom{k}{k-j} = |C_{k-j}|,$$

and so the result follows.

■

From this symmetry, the proof of the LDP for reverse guesswork follows the argument of Theorem 3.3 in Chapter 3 for guesswork, showing that the upper and lower deviation functions coincide. As the argument is near identical, the details are omitted.

Bibliography

[1] E. Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Trans, Inf. Theory*, 42(1):99–105, 1996.

[2] E. Arikan and S. Boztas. Guessing with lies. In *Proc. ISIT*, 2002.

[3] E. Arikan and N. Merhav. Guessing subject to distortion. *IEEE Trans. Inf. Theory*, 44:1041–1056, 1998.

[4] P. Baldi. Large deviations and stochastic homogenization. *Ann. Mat. Pura Appl. (4)*, 151:161–177, 1988.

[5] P. Billinsley. *Probability and Measure*. John Wiley and Sons, 1995.

[6] M. Bloch and J. Barros. *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge University Press, 2011.

[7] S. Boztas. Comments on an inequality on guessing and its application to sequential decoding. *IEEE Trans, Inf. Theory*, 43(6):2062–2063, 1997.

[8] S. Boztas. Oblivious distributed guessing. In *Proc. ISIT*, pages 2161–2165, 2012.

[9] L. L. Campbell. A coding theorem and Rényi's entropy. *Information and Control*, 8:423–429, 1965.

[10] M. M. Christiansen and K. R. Duffy. Guesswork, large deviations and Shannon entropy. *IEEE Trans. Inf. Theory*, 59(2):796–802, 2013.

[11] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard. Brute force searching, the typical set and Guesswork. In *Proc. ISIT*, 2013.

[12] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard. Guessing a password over a wireless channel (on the effect of noise non-uniformity). In *Proc. Asilomar*, 2013.

[13] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard. Quantifying the computational security of multi-user systems, 2014. `http://arxiv.org/pdf/1405.5024.pdf`.

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

[15] L. Czap, V. M. Prabhakaran, C. Fragouli, and S. Diggavi. Secret message capacity of erasure broadcast secret message capacity of erasure broadcast channels with feedback. In *Proc. Information Theory Workshop*, 2011.

[16] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer-Verlag, 1998.

[17] M. Denuit, J. Dhaene, M. Goovaerts, and R. Kass. *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Wiley, 2006.

[18] S. S. Dragomir. Some new estimates for the moments of guessing mappings. *Math. Comm.*, 4:177–190, 1999.

[19] S. Draper, A.Khisti, E. Martinian, A. Vetro, and J. Yedidia. Secure storage of fingerprint biometrics using Slepian-Wolf codes. In *Proc. ITA Workshop*, 2007.

[20] F. du Pin Calmon, M. Médard, L. Zegler, J. Barros, M. Christiansen, and K. Duffy. Lists that are smaller than their parts: A coding approach to tunable secrecy. In *Proc.* 50^{th} *Allerton Conference*, 2012.

[21] K. Duffy and A. P. Metcalfe. The large deviations of estimating rate-functions. *J. Appl. Probab.*, 42(1):267–274, 2005.

[22] A. R. Ghazaryan and E. C. van der Meulen. Optimal strategies for hierarchical guessing problem. In *Proc. 25th Symp. on Inf. Th. in the Benelux*, 2004.

[23] S. Gollakota, H. Hassanieh, B. Ransford, D. Katabi, and K. Fu. They can hear your

heartbeats: non-invasive security for implantable medical devices. In *Proc. ACM SIGCOMM*, pages 2–13, 2011.

[24] M. K. Hanawal and R. Sundaresan. Guessing and compression subject to distortion. Technical report, Division of Electrical Sciences, Indian Institute of Science, 2010.

[25] M. K. Hanawal and R. Sundaresan. Guessing revisited: A large deviations approach. *IEEE Trans. Inf. Theory*, 57(1):70–78, 2011.

[26] M. K. Hanawal and R. Sundaresan. The Shannon cipher system with a guessing wiretapper general sources. *IEEE Trans. Inf. Theory*, 57(4):2503–2516, 2011.

[27] E. A. Haroutunian and A. R. Ghazaryan. Guessing subject to distortion and reliability criteria. *Trans. of the Inst. for Inform. and Autom. Problem of the NAS of RA and of the Y.S.U., Armenia, Math. prob. of cs*, 21:83–90, 2000.

[28] Y. Hayashi and H. Yamamoto. The coding theorems for the Shannon cipher system with a guessing wiretapper and correlated source outputs. In *Proc. ISIT*, 2006.

[29] K. Knopp. *Theory and Application of Infinite Series.* Balckie and Son Limited, 1954.

[30] E. L. Lehmannr. Ordered families of distributions. *Ann. Math. Statis.*, 26:399–419, 1955.

[31] J. T. Lewis and C.-E. Pfister. Thermodynamic probability theory: some aspects of large deviations. *Russian Math. Surveys*, 50(2):279–317, 1995.

[32] J. T. Lewis, C.-E. Pfister, and W. G. Sullivan. Entropy, concentration of probability and conditional limit theorems. *Markov Process. Related Fields*, 1(3):319–386, 1995.

[33] J. T. Lewis, C.-E. Pfister, and W. G. Sullivan. Reconstruction sequences and equipartition measures: An examination of the asumptotic equipartitian property. *IEEE Trans. Inform. Theory*, 43:1935–1947, 1997.

[34] C. Lundin and S. Lindskog. Extending the definition of Guesswork. In *Proc. of International Conference on Information Assurance, and Security*, 2010.

[35] C. Lundin and S. Lindskog. Security implications of selective enctryption. In *Proc. of Metric*, 2010.

[36] D. Malone and K. Maher. Investigating the distribution of password choices. In *Proc. WWW*, 2012.

[37] D. Malone and W. G. Sullivan. Guesswork and entropy. *IEEE Trans. Inf. Theory*, 50(4):525–526, 2004. `http://www.maths.tcd.ie/~dwmalone/p/guess02.pdf`.

[38] D. Malone and W. G. Sullivan. Guesswork is not a substitute for entropy. In *Proc. of the Inf. Technology and Telecom. Conf.*, 2004.

[39] J. L. Massey. Guessing and entropy. *IEEE Int. Symo. Inf Theory*, pages 204–204, 1994.

[40] U. Maurer. Secret key agreement by public discussion from common information. *IEEE Trans. Inf. Theory*, 39(3):733–742, 1993.

[41] A. Menezes, S. Vanstone, and P. Van Oorschot. *Handbook of Applied Cryptography*. CRC Press, Inc., 1996.

[42] N. Merhav and E. Arikan. The Shannon cipher system with a guessing wiretapper. *IEEE Trans. Inform. Theory*, 45(6):1860–1866, 1999.

[43] N. Merhav, R. N. Roth, and E. Arikan. Hierarchical guessing with a fidelity criterion. *IEEE Trans. Inform. Theory*, 45(1):330–337, 1999.

[44] P. F. Oliveira, L. Lima, T. T. V. Vinhoza, J. Barros, and M. Médard. Coding for trusted storage in untrusted networks. *IEEE Trans. Inf. Forensics Security*, 7(6):1890–1899, 2012.

[45] H. H. Permuter, P. Cuff, B. Van Roy, and T. Weissman. Capacity of the trapdoor channel with feedback. *IEEE Trans. Inf. Theory*, 54(7):3150–3165, 2008.

[46] C-E. Pfister and W. Sullivan. Rényi entropy, Guesswork moments and large deviations. *IEEE Trans. Inf. Theory*, 50(11):2794–2800, 2004.

[47] J. Pliam. On the incomparability of entropy and marginal Guesswork in brute-force

attacks. In *Proc. INDOCRYPT*, pages 67–79, 2000.

[48] R. T. Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[49] R. Sundaresan. Guessing based on length functions. In *Proc. ISIT*, 2007.

[50] R. Sundaresan. Guessing under source uncertainty. *IEEE Trans, Inf. Theory*, 53:269–287, 2007.

[51] Y. Sutcu, S. Rane, J. S. Yedidia, S. C. Draper, and A. Vetro. Feature extraction for a Slepian-Wolf biometric system using LDPC codes. In *Proc. ISIT*, 2008.

[52] E. Tekin and A. Yener. The general Gaussian multiple-access and two-way wiretap channels: Achievable rates and cooperative jamming. *IEEE Trans. Inf. Theory*, 54(6):2735–2751, 2008.

[53] J. P. Vilela, M. Bloch, J. Barros, and S. W. McLaughlin. Wireless secrecy regions with friendly jamming. *IEEE Trans. Inf. Forensics Security*, 6(2):256–266, 2011.

[54] P. Walters. *An introduction to ergodic theory*. Springer-Verlag, New York-Berlin, 1982.

[55] A. D. Wyner. The wire-tap channel. *Bell System Technical Journal*, 54(8):1355–1387, 1975.

[56] H. Yamamoto. On secret sharing communication systems with two or three channels. *IEEE Trans. Inf. Theory*, 32(3):387–393, 1986.

[57] H. Yamamoto. A coding theorem for secret sharing communication systems with two Gaussian wiretap channels. *IEEE Trans. Inf. Theory*, 37(3):634–638, 1991.

[58] L. Zang, R. Yates, and W. Trappe. Secrecy capacity of independent parallel channels. In *Proc. Allerton Conference on Communicatoin, Control and Computation*, 2006.