Quantifying Computational Security Subject to Source Constraints, Guesswork and Inscrutability

Ahmad Beirami,[†] Robert Calderbank,[†] Ken Duffy,[*] Muriel Médard[‡]

[†]Department of Electrical and Computer Engineering, Duke University, USA
[*]Hamilton Institute, National University of Ireland Maynooth, Ireland
[‡]Research Laboratory of Electronics, Massachusetts Institute of Technology, USA

Email: [†]{ahmad.beirami, robert.calderbank}@duke.edu, [*]ken.duffy@nuim.ie, [‡]medard@mit.edu

***Abstract*—Guesswork forms the mathematical framework for quantifying computational security subject to brute-force determination by query. In this paper, we consider guesswork subject to a per-symbol Shannon entropy budget. We introduce inscrutability rate as the asymptotic rate of increase in the exponential number of guesses required of an adversary to determine one or more secret strings. We prove that the inscrutability rate of any string-source supported on a finite alphabet $\mathcal{X}$, if it exists, lies between the per-symbol Shannon entropy constraint and $\log|\mathcal{X}|$. We further prove that the inscrutability rate of any finite-order Markov string-source with hidden statistics remains the same as the unhidden case, i.e., the asymptotic value of hiding the statistics per each symbol is vanishing. On the other hand, we show that there exists a string-source that achieves the upper limit on the inscrutability rate, i.e., $\log|\mathcal{X}|$, under the same Shannon entropy budget.**

***Index Terms*—Brute-force attack; Guesswork; Inscrutability; Rényi entropy; Universal methods; Large deviations.**

I. Introduction

In recent years, data storage has experienced a shift toward cloud storage where data is stored in a diversity of sites, each hosted at multiple locations. Cloud service providers assume responsibility for availability, accessibility, and most important, the security, of the stored data. But how secure is the cloud? The vulnerabilities of the cloud storage services have been exploited in several recent incidents resulting in the compromise of very private data stored on the cloud. The security guarantees advertised by individual sites typically assume an isolated attack. However the actual vulnerability is to a coordinated attack, where an attacker with access to more than one site combines partial information to compromise overall security.

Guesswork, which forms the mathematical framework for quantifying computational security subject to brute-force determination by query, was first considered in a short paper by Massey [1] who demonstrated that the number of guesses expected of an attacker bears little relation to the Shannon entropy. Arikan [2] then proved that this guesswork grows exponentially with an exponent that is a specific Rényi entropy for iid processes. His result has been generalized to ergodic Markov chains [3] and a wide range of stationary sources [4], [5]. Arikan and Merhav [6] have also derived fundamental limits on guessing, subject to an allowable distortion. Sundaresan [7] considered guessing on iid processes with unknown statistics and showed that the growth rate of the average guesswork is related to a specific Rényi entropy. Recently, Christiansen and Duffy [8] showed that guesswork satisfies a large deviations principle, completely characterizing the rate function, and providing an approximation to the distribution of guesswork. Finally, in [9], the idea of guesswork was extended beyond guessing a single secret string to a setup in which an attacker wishes to guess one or more out of many secret strings drawn independently from not necessarily identical string-sources. It was shown in [9] that when the individual string-sources are stationary, under some regularity conditions, guesswork satisfies a large deviations principle whose rate function is not necessarily convex.

In this paper, in a setup similar to [9], we consider V secret strings drawn independently from identical string-sources that are constrained to satisfy a given per-symbol Shannon entropy budget. We define inscrutability as the exponent of the average number of guesses required of an adversary to determine a secret string by query. Accordingly, per-symbol inscrutability is the contribution of each symbol in a string to the exponent of average guesswork. Finally, inscrutability rate (if it exists) is defined as the asymptotic rate of increase in the exponent incurred by each additional symbol in the string. Our contributions in this paper are summarized in the following:

- We show that the inscrutability rate of a constrained string-source, if it exists, lies between the per-symbol Shannon entropy constraint and the logarithm of the size of the support, i.e., $\log|\mathcal{X}|$.
- We consider guesswork on finite-memory stationary string-sources[1] with hidden statistics. We show that when the inquisitor does not know the statistics of a finite-memory string-source, he can devise a *universal* guessing strategy that is asymptotically optimal in the sense that it achieves the same inscrutability rate as the string-source with unhidden statistics.
- Finally, we establish that the upper bound on the inscrutability rate is tight by showing that there exists a string-source that achieves an inscrutability rate of $\log|\mathcal{X}|$ under the same Shannon entropy budget.

The work of Robert Calderbank was supported in part by AFOSR under grant FA 9550-13-1-0076; and the work of Muriel Médard was supported in part by Netapp through a Faculty Fellowship.

[1] This is a viable model for the case where the secret strings are chosen as chunks of a compressed file.

II. Problem Setup and Related Work

Let $\mathcal{X} = \{a_1, \ldots, a_{|\mathcal{X}|}\}$ be a finite alphabet of size $|\mathcal{X}|$. Denote $x_k^{n+k-1} = x_k x_{k+1} \ldots x_{n+k-1} \in \mathcal{X}^n$ as a n-string over $\mathcal{X}$. Further, let $x^n = x_1^n$ and for $i > n$, $x_i^n = \varnothing$, where $\varnothing$ denotes the null string. Let μ^n denote a probability measure on $\mathcal{X}^n$. We refer to $\{\mu^n\}_{n=1}^{\infty}$ as a string-source. We use the notation $\{\mu^n\}$ to denote $\{\mu^n\}_{n=1}^{\infty}$ as well. Note that the marginals of $\{\mu^n\}$ might be position dependent, i.e., $\sum_{x_n \in \mathcal{X}} \mu^n(x^n)$ is not necessarily equal to $\mu^{n-1}(x^{n-1})$. A string-source is said to be stationary if $\sum_{x_1,\ldots,x_k} \mu^{n+k}(x^{n+k}) = \mu^n(x_{k+1}^n)$. Let $X^n \in \mathcal{X}^n$ be a random n-string drawn from μ^n.

Some of the results in this paper are derived for finite-memory parametric string-sources.

Definition 1 (finite-memory parametric string-source): A finite-memory parametric string-source is parametrized with a d-dimensional parameter vector $\theta = (\theta_1, ..., \theta_d)$. Let $\Lambda \subset \mathbb{R}^d$ be a d-dimensional *open* set where the d parameters live. Then, μ_θ^n denotes a parametric probability measure defined by the parameter vector θ on n-strings. We assume that $\{\mu_\theta^n\}$ is a stationary string-source for all $\theta \in \Lambda$. We also assume that the source has a finite memory of at most h, i.e., the probability of observing each symbol at any position at most depends on the symbols in the previous h positions. For the convenience of notation, we assume that x_{-h+1}^0 is a run of length h of symbol a_0. We denote $\mathcal{P}_\Lambda$ as the *family* of parametric string-sources such that $\theta \in \Lambda$, i.e., $\mathcal{P}_\Lambda = \{\{\mu_\theta^n\} : \theta \in \Lambda\}$.

The finite-memory parametric models include all iid and finite-memory Markov string-sources. The simplest parametric model is a binary iid string-source with $\mathcal{X} = \{0, 1\}$ and $\theta = P\{X_i = 1\}$ is the single source parameter, which lives in $\Lambda = (0, 1)$. Note that we exclude the boundaries. For example, $\mu_\theta(1, 1, 0) = \theta^2(1 - \theta)$. Consider a binary Markov source as another parametric model on $\mathcal{X} = \{0, 1\}$ with $d = 2$ parameters

$$(\theta_1, \theta_2) = (P\{X_i = 1|X_{i-1} = 0\}, P\{X_i = 1|X_{i-1} = 1\}),$$

that live in $\Lambda = (0, 1) \times (0, 1)$. For example, $\mu_\theta(1, 1, 0) = \theta_1\theta_2(1 - \theta_2)$ since we assume that $x_0 = 0$. Finally, consider order r Markov processes over alphabet $\mathcal{X}$. In this case, the source parameters are the *non-zero* transition probabilities given the previous r symbols, and hence, $d = |\mathcal{X}|^r(|\mathcal{X}| - 1)$.

Let $H^n(\mu^n)$ denote the Shannon entropy of a random n-string drawn from μ^n, i.e.,[2]

$$H^n(\mu^n) = -E\left\{\log \mu^n(X^n)\right\} = \sum_{x^n \in \mathcal{X}^n} \mu^n(x^n) \log\left(\frac{1}{\mu^n(x^n)}\right).$$

Further, let $H(\{\mu^n\})$ be the Shannon entropy rate of the string-source (if it exists), i.e., $H(\{\mu^n\}) \triangleq \lim_{n\to\infty} \frac{1}{n} H^n(\mu^n)$.

Similar to [9], we consider V strings, denoted by $\mathbf{x}^{n,V} = (x^n(1), \ldots, x^n(V))$, that are drawn independently from an identical string-source $\{\mu^n\}$. This extends the guesswork problem to a multi-string system with V strings where an inquisitor wishes to identify U out of V strings. The case where $V = U = 1$ corresponds to a single-string guesswork problem and has been studied extensively.

We have the following assumptions on the attacker and chooser:

- The length n of the chosen strings is known to the attacker.
- The chooser draws V strings *independently* from the string-source $\{\mu^n\}$.
- $\{\mu^n\}$ is known to the attacker. This assumption will be dropped for finite-memory parametric string-sources in Section IV.
- At each time, the attacker is allowed to pick one of the systems, say system i, and ask "Is $X^n(i) = y^n$?". He continues this process until he correctly guesses U of the V randomly drawn strings $\mathbf{x}^{n,V}$.
- In Sections III and V, we assume that the chooser is constrained to choose a string-source $\{\mu^n\} \in \Delta_{H_\mathcal{X}}$, where $\Delta_{H_\mathcal{X}}$ is the set of all string-sources supported on the finite alphabet $\mathcal{X}$ that satisfy a per-symbol entropy constraint of $H_\mathcal{X}$ for all $n \geq 1$. That is $(1/n)H^n(\mu^n) = H_\mathcal{X}$. We also assume that $H_\mathcal{X} > 0$.

In the single-string special case, it is straightforward to see that when the probability distribution μ^n is known to the attacker, the optimal strategy (that stochastically dominates any other strategy) would be to order all possible n-strings from the most likely outcome to the least likely (breaking ties arbitrarily), and then query the strings one by one from the top of the list until the correct password has been guessed.

In [9], it was proved that an asymptotically optimal strategy for the multi-string guesswork would be to round-robin the single-string optimal strategies. That is to query the most likely string of system 1 followed by the most likely string of system 2 and so forth till system V, before moving to the second most likely string of each system.

In the multi-string case, let $G_{\mu^n}(U, V, \mathbf{x}^{n,V})$ denote the number of queries required of an attacker to guess U out of V of sequences $\mathbf{x}^{n,V} = (x^n(1), \ldots, x^n(V))$ using the asymptotically optimal strategy described above. In the single-string case, we further use the short-hand $G_{\mu^n}(x^n)$ to denote $G_{\mu^n}(1, 1, \mathbf{x}^{n,1})$. We use the subscript μ^n in $G_{\mu^n}(\cdot)$ to emphasize that it is dependent on the specific string-source probability measure μ^n. The average guesswork $E\{G_{\mu^n}(U, V, \mathbf{X}^{n,V})\}$ quantifies the average number of guesses required of an attacker to identify U out of V of the secret strings, where the expectation is taken with respect to the iid copies of μ^n on each string.

Massey [1] demonstrated that the average guesswork in the single-string case is lower bounded by

$$E\{G_{\mu^n}(X^n)\} \geq (1/4)2^{H^n(\mu^n)} + 1.$$

The bound is tight up to a factor of $4/e$ for a geometric distribution (on an infinite support). Massey also proved that an upper bound on the average guesswork in terms of the Shannon entropy does not exist proving that average guesswork bears little relation to the Shannon entropy of the string-source in general.

In [2], Arikan considered an iid process and proved that the exponent of the average growth rate of the average guesswork (which we refer to as the inscrutability) is the specific Rényi

[2] In this paper $\log(\cdot)$ always denotes the logarithm in base 2.

entropy with parameter $\alpha = (1/2)$. In other words,

$$H^n_{1/2}(\mu^n) - \log(1+\log|\mathcal{X}|) \le \log E\{G_{\mu^n}(X^n)\} \le H^n_{1/2}(\mu^n),$$

where $H^n_\alpha(\mu^n)$ is the Rényi entropy of order α ($\alpha > 0$, $\alpha \neq 1$) defined as

$$H^n_\alpha(\mu^n) = \frac{1}{1-\alpha}\log\left(\sum_{x^n\in\mathcal{X}^n}\mu^n(x^n)^\alpha\right).$$

Further, if it exists, the Rényi entropy rate of the string-source is defined as $H_\alpha(\{\mu^n\}) = \lim_{n\to\infty}\frac{1}{n}H^n_\alpha(\mu^n)$. Note that $H_\alpha(\{\mu^n\})$ if it exists converges to $H(\{\mu^n\})$ as $\alpha \to 1$.

Definition 2 (inscrutability): The inscrutability of identifying U out of V of the V random n-strings $\mathbf{X}^{n,V}$, denoted by $S^n(U, V, \mu^n)$ is defined as

$$S^n(U, V, \mu^n) \triangleq \log E\{G_{\mu^n}(U, V, \mathbf{X}^{n,V})\}.$$

The inscrutability rate of a string-source, denoted by $S(U, V, \{\mu^n\})$, if it exists, is defined as

$$S(U, V, \{\mu^n\}) \triangleq \lim_{n\to\infty}\frac{1}{n}S^n(U, V, \mu^n).$$

In particular, it can be concluded from Arikan's result that for an iid string-source $\{\mu^n\}$ the inscrutability rate for $U = V = 1$ is

$$S(1, 1, \{\mu^n\}) = H_{1/2}(\{\mu^n\}).$$

Arikan's result was later generalized to ergodic Markov chains [3] and a wide class of stationary sources [4], [5], for which the inscrutability rate can be related to the specific Rényi entropy rate with parameter $(1/2)$ under those setups as well. Recently, the authors in [9] derived the inscrutability rate for arbitrary U and V as the specific Rényi entropy rate with parameter $(V - U + 1)/(V - U + 2)$. That is

$$S(U, V, \{\mu^n\}) = H_{(V-U+1)/(V-U+2)}(\{\mu^n\}). \tag{1}$$

In particular, it can be deduced from this result that in the large system limit when $V \to \infty$, if U/V stays bounded away from 1 the inscrutability rate converges to the specific Shannon entropy rate. This is stated in the following proposition.

Proposition 1: If U scales with V in such a way that $U/V < (1 - \delta)$ for some $\delta > 0$, then

$$\lim_{V\to\infty} S(U, V, \{\mu^n\}) = H(\{\mu^n\}).$$

Proof: This is an immediate consequence of (1). ■

The authors in [9] further showed that the guesswork $G_{\mu^n}(U, V, \mathbf{X}^{n,V})$ satisfies a large deviations principle and identified its rate function which is stated in Lemma 4 of [9].

III. Lower Bound on Inscrutability

In this section, we consider a multi-string system with secret strings drawn *independently* from the string-source $\{\mu^n\}$. We assume that $\{\mu^n\} \in \Delta_{H_\mathcal{X}}$. First, we identify the string-source in $\Delta_{H_\mathcal{X}}$, denoted by $\{\underline{\mu}^n\}$, that achieves the smallest inscrutability for all $n \ge 1$.

Theorem 2: For any $1 \le U \le V$, the inscrutability of identifying U out of V secret strings chosen from any string-source $\{\mu^n\} \in \Delta_{H_\mathcal{X}}$ is bounded from below by

$$S^n(U, V, \mu^n) \ge S^n(U, V, \underline{\mu}^n), \tag{2}$$

where $\underline{\mu}^n$ is a truncated geometric distribution on the support $\mathcal{X}^n$ that satisfies the per-symbol entropy constraint. Further, the inscrutability rate exists for the string-source $\{\underline{\mu}^n\}$ and is equal to the per-symbol Shannon entropy constraint. That is

$$S(U, V, \{\underline{\mu}^n\}) = \lim_{n\to\infty}\frac{1}{n}S^n(U, V, \underline{\mu}^n) = H_\mathcal{X}. \tag{3}$$

Sketch of the proof: It turns out this is an optimization problem with concave constraints where the set over which the optimization is performed is convex, hence, the minimizing distribution is either found by the method of Lagrange multipliers or lies on the boundary of the set. In this case, it can be shown that the minimizer is obtained by forming the Lagrangian. ■

By considering Proposition 1 and Theorem 2, when $\{\mu^n\}$ is a finite-memory parametric string-source, if U scales with V such that $U/V < (1 - \delta)$, then

$$\lim_{V\to\infty} S(U, V, \{\mu^n\}) = S(U, V, \{\underline{\mu}^n\}) = H(\{\mu^n\}).$$

This shows, as V grows large, the inscrutability rate of any finite-memory parametric string-source with a given Shannon entropy rate approaches the lowest limit of the inscrutability rate. Observe that inscrutability rate is defined as the asymptotic limit as $n \to \infty$ of the per-symbol inscrutability and in the above statement the limits as $n \to \infty$ and $V \to \infty$ are not interchangeable.

IV. Inscrutability of Finite-Memory Parametric String-Sources with Hidden Statistics

In this section, we investigate the impact of hiding the string-source statistics on the inscrutability of identifying U out of V secret strings drawn independently from a parametric string-source $\{\mu^n_\theta\}$. Note that the round-robin of single-string optimal strategies is an asymptotically optimal strategy for the multi-string system [9], and hence, we only need to find an asymptotically optimal single-string guessing strategy. This is the essence of the universal ordering problem first studied by Weinberger *et al.* [10]. Later Arikan and Merhav [6] proposed a universal ordering based on the empirical entropy of iid sources, which also bears great similarity with Kosut and Sankar's recently proposed universal type-size coding [12] (universal compression without prefix constraint) on iid processes.

A. Universal Type-Size Guessing Strategy

We shall provide a guessing strategy for parametric string-sources using the method of types (see [11]) and in particular by considering more general notion of types [17].

The type class of sequence x^n is defined as

$$T_\Lambda(x^n) = \{y^n \in \mathcal{X}^n : \mu^n_\theta(y^n) = \mu^n_\theta(x^n)\ \forall\theta \in \Lambda\}. \tag{4}$$

Further, $|T_\Lambda(x^n)|$ denotes the size of the type class of x^n, i.e., the total number of sequences with the same type as x^n.

Single-string universal guessing strategy:

- Order all sequences based on the size of their corresponding type classes in an ascending fashion and break ties arbitrarily.
- Let $G_\star(x^n)$ be the order in which the sequence x^n appears in the above list. Clearly, the sequence x^n may appear before y^n only if $|T_\Lambda(x^n)| \leq |T_\Lambda(y^n)|$.

Our main result on the universal type-size guessing strategy described above is the following.

Theorem 3: Let Λ denote the set of parametric sources over finite alphabet $\mathcal{X}$. Let $G_{\mu^n_\theta}$ be an optimal non-universal guessing strategy for parametric source with parameter vector θ, such that in $G_{\mu^n_\theta}$ ties are broken in favor of guessing sequences with smaller type-sizes first and if there is a tie in the size of the type the tie is broken arbitrarily. Then for any individual sequence x^n, the universal guessing function $G_\star(x^n)$ obeys:

$$\frac{G_\star(x^n)}{G_{\mu^n_\theta}(x^n)} = O(n^{d+1}), \tag{5}$$

where d is the number of source parameters.

Sketch of the proof: Let N_n denote the total number of types associated with n-strings. It is straightforward to show that $N_n = O(n^d)$ (see [11]). Observe that using the defined guessing strategy, only sequences whose type size is at most $|T_\Lambda(x^n)|$ are guessed before x^n. Hence, $G_\star(x^n) \leq Cn^d|T_\Lambda(x^n)|$ for some absolute constant C. If x^n is not of highest likelihood and smallest type-size, we can change one symbol in x^n to achieve a sequence y^n, which is guessed earlier than x^n in the optimal non-universal strategy $G_{\mu^n_\theta}$. Therefore, $G_{\mu^n_\theta}(x^n) \geq |T_\Lambda(y^n)|$. On the other hand, since y^n is obtained by changing only one symbol in x^n, we have $\frac{|T_\Lambda(x^n)|}{|T_\Lambda(y^n)|} < n$. Putting these together completes the proof. ■

Theorem 3 is pointwise and hence can be invoked to prove the large deviations principle for the multi-string system with unknown statistics. The derivation is omitted due to page limit. This was expected in light of the analysis of the single-string universal strategies in [7], and the recent results on the large deviations for multi-string guesswork [9]. Let the universal inscrutability rate of the universal type-size guessing strategy be defined as

$$S_\star(U, V, \{\mu^n\}) = \lim_{n\to\infty} \frac{1}{n} \log E\{G_\star(U, V, \mathbf{X}^{n,V})\}.$$

Here, we also obtain the multi-string counterpart of Sundaresan's Theorem 16 of [7] on the growth rate of the average universal guesswork.

Corollary 4: The universal inscrutability rate of the universal type-size guessing strategy is given by

$$S_\star(U, V, \{\mu^n\}) = H_{(V-U+1)/(V-U+2)}(\{\mu^n\}). \tag{6}$$

This is straightforward by from the large deviations principle and Corollary 1 of [9].

This establishes that the inscrutability rates for a finite-memory parametric sources with hidden and unhidden statistics are equal.

B. Universal Bayesian Guessing Strategy

In this section, we present a Bayesian viewpoint on universal guesswork. The Bayesian construction assumes the least-favorable Jeffreys' prior in the context of universal compression (see [13]) and is readily applicable to finite-memory sources, such as, finite-state machines [15] and context trees [16]. Let $\mathcal{I}(\theta)$ be the Fisher information matrix associated with the parameter vector θ, i.e.,

$$\mathcal{I}_{i,j}(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n\log e} E\left\{ \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\left(\frac{1}{\mu^n_\theta(X^n)}\right)\right\}. \tag{7}$$

We assume that the source is ergodic such that the above limit exists. Let Jeffreys' prior, denoted by p_Λ, be

$$p_\Lambda(\theta) \triangleq \frac{|\mathcal{I}(\theta)|^{\frac{1}{2}}}{\int_\Lambda |\mathcal{I}(\lambda)|^{\frac{1}{2}} d\lambda}. \tag{8}$$

Let μ^n_Λ denote the mixture distribution with Jeffreys' prior:

$$\mu^n_\Lambda(x^n) = \int_\Lambda \mu^n_\theta(x^n) p_\Lambda(\theta) d\theta. \tag{9}$$

Let $G_{\mu^n_\Lambda}$ be the optimal procedure for the distribution μ^n_Λ.

Theorem 5: $G_{\mu^n_\Lambda}$ and $G_\star$ are asymptotically equivalent, and hence are both asymptotically optimal.
The proof follows the same lines of Theorem 6 of [14].

C. Twice Universal Guesswork on Finite-Memory Sources

Thus far, we assumed that the source parameters of a finite memory source were unknown but Λ is known. We further extend to twice universal guessing on finite-memory sources, where in addition to the source statistics being unknown to the inquisitor, the (finite) source model is also unknown (see [18] for a formal definition).

Let $h : \mathbb{N} \to \mathbb{N}$ be any function such that $h(n) = o(\log n)$ and $h(n) = \omega(1)$. For any $n \geq 1$, let the unknown source model be described by a Markov source of order $h(n)$, which defines a parametric source with $d(n) = (|\mathcal{X}| - 1)|\mathcal{X}|^{h(n)}$ parameters. Let $\Lambda_{d(n)}$ denote the space of parameter vectors for the model. Note that using this strategy we asymptotically overestimate the number of unknown source parameters of any finite-order process as the number of source parameters is growing unboundedly. On the other hand, even with this model we can achieve the inscrutability of known source parameter vector case. Let $\mu^n_{\Lambda_{d(n)}}$ be defined similar to (9). We use h instead of $h(n)$ when it is clear from the context. Let $G_{\mu^n_{\Lambda_{d(n)}}}$ be the order in which x^n appears when the sequences are sorted based on $\mu^n_{\Lambda_{d(n)}}$ in a descending fashion. Then, similar to Theorem 3 and considering the growth rate of $h(n)$ we have

$$\frac{1}{n} \log\left(\frac{G_{\mu^n_{\Lambda_{d(n)}}}(x^n)}{G_{\mu^n_\theta}(x^n)}\right) = o(1).$$

Alternatively, a universal strategy could be achieved by using type-size coding using the twice universal types defined in [18], which would also be asymptotically equivalent to the aforementioned Bayesian strategy. Let $S_{\star\star}(U, V, \{\mu^n\})$ be the twice universal inscrutability rate of guessing U out of V

secret strings chosen from an unknown Markov string-source with unknown finite order.

Corollary 6: For any $1 \leq U \leq V$, the twice universal inscrutability rate of a finite-memory Markov source with unknown order and unknown parameters is the specific Rényi entropy rate satisfies:

$$S_{\star\star}(U, V, \{\mu^n\}) = H_{(V-U+1)/(V-U+2)}(\{\mu^n\}).$$

Observe that when the number of unknown parameters grows, the class of probability distributions that can be well approximated using the parametric model becomes richer. On the other hand, the cost associated with universality also grows linearly with the number of unknown parameters resulting in a fundamental tradeoff. Although our results show that the cost of universality is asymptotically negligible, this overhead can be quite large for moderate problem sizes when the class of distributions is fairly complex. This is analogous to the cost of universal compression that can be quite large for small to moderate sequence lengths while universal compression is known to asymptotically achieve the Shannon entropy.

V. Upper Bound on Inscrutability

Thus far, we showed that with a constrained Shannon entropy budget, choosing strings independently from a stationary string-source, corresponds to the minimum inscrutability rate against adversarial attacks. Furthermore, if the string-source is finite-memory parametric, hiding the string-source statistics is not asymptotically a remedy in the sense that it does not decrease the inscrutability rate. A natural question is whether there exists a string source in $\Delta_{H_{\mathcal{X}}}$ that has a larger inscrutability rate than the Shannon entropy rate. This is answered in the following theorem.

Theorem 7: For any $1 \leq U \leq V$, the inscrutability of identifying U out of V strings drawn independently from $\{\mu^n\} \in \Delta_{H_{\mathcal{X}}}$ is bounded from above by

$$S^n(U, V, \mu^n) \leq S^n(U, V, \overline{\mu}^n), \tag{10}$$

where $\overline{\mu}^n$ is such that all symbols but one are uniform and the probability measure is distributed between the most probable symbol and the rest of the uniform symbols such that the Shannon entropy budget $H_{\mathcal{X}}$ is satisfied. Further, the inscrutability rate exists for the string-source $\{\overline{\mu}^n\}$ and is equal to $\log |\mathcal{X}|$. That is

$$S(U, V, \{\overline{\mu}^n\}) = \lim_{n\to\infty} \frac{1}{n} S^n(U, V, \overline{\mu}^n) = \log |\mathcal{X}|. \tag{11}$$

The proof is omitted due to page limit.

Theorem 7 indeed reveals that given any non-zero Shannon entropy budget $H_{\mathcal{X}}$, the inscrutability rate of the string-source $\{\overline{\mu}^n\}$ is equal to that of a uniform distribution on the entire support set. In light of Theorems 2 and 7, if the inscrutability rate exists for a string-source $\{\mu^n\} \in \Delta_{H_{\mathcal{X}}}$, then for all $1 \leq U \leq V$ it satisfies:

$$H_1(\{\mu^n\}) = H_{\mathcal{X}} \leq S(U, V, \{\mu^n\}) \leq \log |\mathcal{X}| = H_0(\{\mu^n\}). \tag{12}$$

VI. Conclusion

In this paper, we considered multi-string guesswork subject to source constraints. We defined inscrutability as the average growth rate of the exponential number of guesses required of an inquisitor to determine one or more secret strings out of many in a brute-force attack. We established that the inscrutability rate lies between the Shannon entropy rate (Rényi entropy rate of order 1) and the logarithm of the size of the support set (Rényi entropy rate of order 0) and showed that there exist sources that achieve either end of the range. Finally, we also proved that hiding the statistics of any finite-memory string-source does not provide larger inscrutability rate, i.e., the per-symbol gain of hiding the statistics of a finite-memory string-source is asymptotically vanishing.

Acknowledgements

The authors are thankful to Ali Makhdoumi for several discussions and careful reading of the paper.

References

[1] J. L. Massey, "Guessing and entropy," in *1994 IEEE International Symposium on Information Theory Proceedings*. IEEE, 1994, p. 204.

[2] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 99–105, Jan. 1996.

[3] D. Malone and W. G. Sullivan, "Guesswork and entropy," *IEEE Trans. Inf. Theory*, vol. 50, no. 3, pp. 525–526, Mar. 2004.

[4] C. E. Pfister and W. G. Sullivan, "Renyi entropy, guesswork moments, and large deviations," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2794–2800, Nov. 2004.

[5] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 70–78, Jan. 2011.

[6] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[7] R. Sundaresan, "Guessing under source uncertainty," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 269–287, Jan. 2007.

[8] M. M. Christiansen and K. R. Duffy, "Guesswork, large deviations, and Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 796–802, Feb. 2013.

[9] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard, "Quantifying the computational security of multi-user systems," *arXiv preprint arXiv:1405.5024*, May 2014.

[10] M. J. Weinberger, J. Ziv, and A. Lempel, "On the optimal asymptotic performance of universal ordering and of discrimination of individual sequences," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 380–385, Mar. 1992.

[11] I. Csiszár, "The method of types," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[12] O. Kosut and L. Sankar, "New results on third-order coding rate for universal fixed-to-variable source coding: Converse and prefix codes," in *2014 International Symposium on Information Theory (ISIT 2014)*, Jul. 2014.

[13] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453 –471, May 1990.

[14] A. Beirami and F. Fekri, "Fundamental limits of universal lossless one-to-one compression of parametric sources," in *2014 IEEE Information Theory Workshop (ITW '14)*, Nov. 2014, pp. 212–216.

[15] A. Martin, G. Seroussi, and M. Weinberger, "Linear time universal coding and time reversal of tree sources via FSM closure," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1442 – 1468, Jul. 2004.

[16] F. Willems, Y. Shtarkov, and T. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.

[17] A. Martin, G. Seroussi, and M. J. Weinberger, "Type classes of context trees," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4077–4093, Jul. 2012.

[18] Á. Martín, N. Merhav, G. Seroussi, and M. J. Weinberger, "Twice-universal simulation of markov sources and individual sequences," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4245–4255, Sept. 2010.