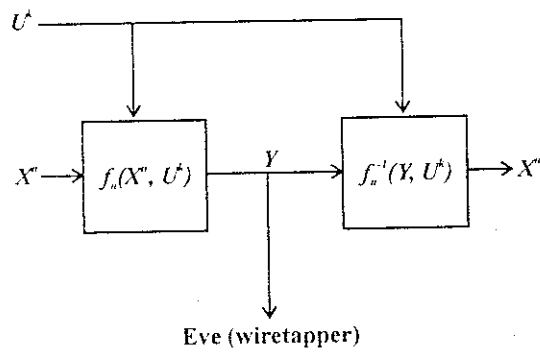
An Introduction to Guessing

Rajesh Sundaresan
Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore-560012
rajesh@ece.iisc.ernet.in

Abstract : Some interesting results on a guessing wiretapper's performance on a Shannon cipher system are summarized. The performance metrics are exponents of guessing moments and probability of large deviations. Connections with compression and the game of twenty questions are discussed. Asymptotic optimality of guessing attacks based on Lempel-Ziv coding lengths is highlighted.

1. Introduction

Consider the classical Shannon cipher system [1] shown in Fig. 1. Let $X^n = (X_1,\ldots,X_n)$ be a message where each letter takes values on a finite set $\mathbf{X}$. This alphabet set could be binary, the Latin characters, or letters from a standard qwerty key-board. The message is assumed to be put out by a source, an entity that emits strings of specified length (here n) according to a specified (or partially specified) probability law. The message should be communicated securely from a transmitter to a receiver. Both of these have access to a common secure key U^k of k purely random bits independent of X^n. The transmitter computes the cryptogram $Y = f_n(X^n, U^k)$ via an encryption function f_n and sends it to the receiver over a public channel. The cryptogram may be of variable length. The function f_n is invertible given U^k. The receiver, knowing Y and U^k, computes $X^n = f_n^{-1}(Y, U^k)$. The functions f and f_n^{-1} are published.



An attacker (wiretapper) has access to the cryptogram Y, knows f_n and f_n^{-1}, and attempts to identify X^n without knowledge of U^k. The attacker can use knowledge of the statistics of X^n. We assume that the attacker has a test mechanism that tells him whether a guess X^n is correct or not. For example, the attacker may wish to attack an encrypted password or personal information to gain access to, say, a computer account, or a bank account via internet, or a classified database [2]. In these situations, successful entry into the system or a failure provides the natural test mechanism. We assume that the attacker is

R. Sundaresan is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India
This work was supported by the Defence Research and Development Organisation, Ministry of Defence, Government of India, under the DRDO-IISc Programme on Advanced Research in Mathematical Engineering, and by the University Grants Commission under Grant Part (2B) UGC-CAS-(Ph.IV).

allowed an unlimited number of guesses. The key rate for the system is $R := k/n$ which represents the number of bits of key used to communicate one message letter.

The attacker stops as soon as he makes a correct guess. If a particular submitted guess is incorrect, he does not gather any further information on the realisation other than the fact that his prior guesses were incorrect. Thus, the set of strategies at the attacker's disposal is all the possible orderings of $\mathcal{X}^n$, each being an order in which he will submit his guesses. Depending on the cryptogram, the attacker may choose an ordering. A particular ordering, denoted $G(. \mid y)$, is thus a mapping from $\mathcal{X}^n$ to $\{1, 2, \ldots, |\mathcal{X}|^n\}$, one for each observed cryptogram.

The goal of the guessing attacker is to identify the realisation in as few guesses as possible. Typical quantities of interest are the expected number of guesses $\mathrm{E}[G(X^n \mid Y)]$, moments of guessing $\mathrm{E}[G(X^n \mid Y)^\rho]$ for $\rho > 0$, or the probability of exceeding a certain number of guesses $\mathrm{P}\{G(X^n \mid Y) \geq C\}$. The guessing attacker would choose $G(. \mid y)$ to make these as small as possible. On the other hand, the goal of the system designer is to make the aforementioned quantities as large as possible. The following questions arise :

1) The problem described above is reminiscent of the game of "twenty questions". What is the connection?
2) Given information on the secret message source, what is the best strategy for the attacker?
3) What is the attacker's performance using the optimal guessing strategy? In particular, how does it grow with n, and what is the relationship with the key rate R?
4) How sensitive is the attacker's performance to knowledge of source statistics?

The purpose of this review article is to give a summary of answers to the above questions. We will focus on only the expected number of guesses and guessing moments in this review.

There are two scenarios of interest. The first is the case of perfect secrecy where encryption is so strong that the cryptogram is practically useless to the attacker. This is the case when the key rate is large (for example, $R \geq 1$). The second is one where the key rate is sufficiently small that the attacker may exploit the constraint to reduce his guessing effort. We will look at both.

2. Relationship to "Twenty Questions"

The game of "twenty questions" has two players – a questioner and an answerer. The answerer has a subject in mind which is not known to the questioner. The questioner asks questions one after another and has to identify the subject, based only on the answerer's responses. The answerer is only allowed to say "yes" or "no". The questioner's goal is to minimise the number of questions asked. (More precisely, expected number of questions, moments of the number of questions, or probability that the number of questions exceeds a certain number, just as above).

The chosen subject is modeled as the realisation of a source. The questioner may ask any set-membership question: "Does $X^n 2 \in E_i$?" for $i = 1, 2, \ldots$, where $E_i \subset \mathcal{X}^n$. He picks this series of questions to minimize the number of questions asked, and exploits the statistical structure inherent in the choice of the subject. For example, if the questioner were to guess a famous personality, given the choice's geographical and temporal bias, the questioner ought to pick a subset of personalities based on recent news clippings. For example, "Is the person a politician or

a cricketeer?" is a good first question keeping in mind the recent elections in Gujarat, and the recently concluded cricket test series (This article was written in early December 2007).

The game of twenty questions has relevance to the classical problem of source compression. Huffman encoding can be thought of as yes-no answers to a certain series of set-membership questions. The sets E_i are picked so that an answer at each stage reduces the uncertainty to a set whose probability is half of that of the prior uncertainty set. For the specific case when all realisations are equi-probable, a binary search has exactly this property.

It is well-known that the minimum expected number of questions is $H(PX_n)$, the Shannon entropy of the source restricted to strings of length n. For a stationary source, this grows asymptotically linearly in n with a slope equal to the entropy rate of the system:

$$H := \lim_{n\to\infty} n^{-1} H(P_{X^n}).$$

Moreover, we know that there exists a compression strategy that work for a rather large class of sources. For example, the Lempel-Ziv encoding [3] is asymptotically optimal for every stationary and ergodic source. Such strategies are termed "universal", where the universality refers to asymptotic optimality for sources that belong to a specified class of sources (here, stationary and ergodic sources).

In the language of the game of twenty questions, the Lempel-Ziv coding strategy yields a series of questions that enable the questioner to discover the realisation in $nH + o(n)$ questions where $o(n)$ is a sequence that satisfies $\lim_{n\to\infty} o(n)/n = 0$. Asymptotically, the penalty for not knowing the exact parameters of the source within the class is the sub-linear quantity $o(n)$.

The game of guessing (in the perfect secrecy scenario) is the game of twenty questions with the added restriction that sets, membership to which is tested, are singletons, i.e., "Does $X^n \in \{x^n\}$?", or equivalently, "Is $X^n = x^n$?", and so on. Naturally, we anticipate that the expected number of guesses grows at a rate faster than nH. Stretching the connection a little further, we anticipate the existence of robust guessing attacks that do not depend on exact parameters of source statistics.

3. Attacker's Best Strategy and Performance

Given the probability mass function of X^n, the function f_n, and the cryptogram Y, the attacker can determine the posterior probabilities of the messages $P_{X^n|Y}(. \mid y)$ using Bayes rule. His best guessing strategy having observed $Y = y$ is then to guess in the decreasing order of these posterior probabilities $P_{X^n|Y}(. \mid y)$. As one might expect, this is the strategy that minimizes all the measures of performance given above.

Without loss of generality, we assume that the message alphabet is binary. Let us first consider the case $R \geq 1$ or $k \geq n$. One may set fn to be such that the output is the XOR of the n message bits and the first n key bits. Since the key bits are purely random, a simple application of Bayes rule shows that

$$P_{X^n|Y}(x^n \mid y) = P_{X^n}(x^n)$$

i.e., the cryptogram is useless to the attacker. The best attack strategy is to guess in the decreasing order of source probabilities. Let us denote this guessing strategy as G^*.

Fact 1: Consider a discrete memoryless source (DMS): a source where each letter is independent and identically distributed with a generic distribution P_X.

- (Arikan [4]) For any DMS, the expected number of guesses for the optimal guessing strategy grows exponentially in n. The rate of growth is given by the R´enyi entropy of order 1/2. More precisely,

$$\lim_{n\to\infty} n^{-1}\log E[G^*(X^n)] = H_{\frac{1}{2}}(P_X) \qquad (1)$$

where the R´enyi entropy of order α is

$$H_\alpha(P_X) := \frac{1}{1-\alpha}\log\Big(\sum_{a\in X} P_X(a)^\alpha\Big) \qquad (2)$$

- (Arikan [4]) More generally,

$$\lim_{n\to\infty} n^{-1}\log E[G^*(X^n)^\rho] = \rho H_{\frac{1}{1+\rho}}(P_X)$$

It is the form of (1) that is important. It tells us that in contrast to twenty questions, the expected number of guesses grows not linearly, but exponentially in n. This should not be surprising given the restriction on the type of questions asked (singleton-set membership questions).

Suppose now that the key rate is smaller than 1, i.e., we have fewer key bits than message bits. Consider the extreme case when $k = 1$, i.e., there are only two possible keys, 0 or 1. Clearly, given the cryptogram, the attacker can attempt decryption using keys 0 and 1 to get two possible messages. He will then submit the more probable of the two as the first guess and the other one next. He needs at most two guesses. If $k = 2$, there are four possible keys, and the attacker needs at most 22 guesses. More generally, when $k = nR$, this exhaustive key-search attack yields the correct guess in at most 2^{nR} guesses (which is less than 2^n). The system designer should choose f_n to make the expected number of guesses as close to this upper bound as possible. For an f_n, let G^*_{fn} be the best attack strategy.

The following result characterises the expected number of guesses as a function of R.

Fact 2 (Merhav & Arikan [2]): For a DMS, the optimal exponent of guessing moment is given by

$$\lim_{n\to\infty} \sup_{f_n} n^{-1}\log E[G^*_{f_n}(X^n \mid Y)^\rho] = E(R,\rho) = \max_Q[\rho \min\{H(Q), R\} - D(Q\,\|\,P)] \qquad (3)$$

Yet again, it is the qualitative conclusion that one can draw from (3) that is important, and not the actual expression itself. Firstly, $E(R, \rho)$ never exceeds ρR, the performance of an exhaustive key-search attack. Secondly, when $R \geq 1$, we have $R \geq H(Q)$, and therefore the minimum in the expression for $E(R, \rho)$ is $H(Q)$. We therefore anticipate that

$$E(\rho) := \max_Q[\rho H(Q) - D(Q\,\|\,P)] = \rho H_{\frac{1}{1+\rho}}(X),$$

a well-known identity in the information theory. Thirdly, (3) implies that $E(R, \rho)$ is a non-decreasing and concave function of R. The nondecreasing property is expected given the operational significance of $E(R, \rho)$ as the guessing moment's exponent. As the key rate increases, the number of key bits increases, and it should become more difficult to guess the

realisation. The interesting aspect is the concavity – the returns from adding an extra key bit diminish with the length of the key.

A precise characterisation of $E(R, \rho)$ is the following.

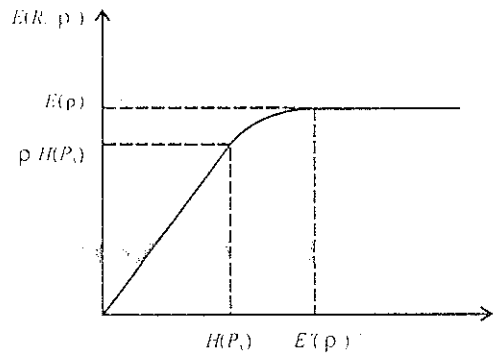


Fig. 2. Exponent of guessing moment as a function of R.

Fact 3 (Merhav & Arikan [2]): E(R, ρ) is given by the following expression. See Fig. 2.

$$E(R,\rho)=\begin{cases}\rho R, & R<H(P_X),\\ (\rho-\theta_0)R+E(\theta_0), & H(P_X)\le R\le E'(\rho),\\ E(\rho), & R>E'(\rho)\end{cases}$$

where θ_0 depends on R and belongs to $[0, \rho]$ in the second case.

Merhav and Arikan [2] further provide the following interpretation on the shape of this function. For low key rates, i.e., $R < H(P_X)$, an exhaustive key-search attack is the most effective attack (from the point of view of asymptotic guessing moments). On the other hand, when $R > E'(\rho)$, the key rate is sufficiently high to render the cryptogram useless. The attacker might as well discard the cryptogram and use only the source's statistical structure in submitting its guesses. In this regime, the key rate is high enough to attain perfect secrecy. For intermediate values of $R \in [H(P_X), E'(\rho)]$, the attacker should use a combination of the exhaustive key-search attack and an attack based on the source's statistical structure. In particular, the attacker should guess alternately from each list, skipping those sequences that have already been guessed. (Each sequence will occur twice in the interlaced list). This leads to a penalty of at most a factor of 2 in the expected number of guesses, yielding a loss of $n^{-1} \log 2 = o(1)$ in the exponent of the expected number of guesses. The quantity $E'(\rho)$ turns out to be the entropy of a "tilted" source. See [2] or [5] for more details. These results have been generalised to a class of sources with memory in [5].

4. Concluding Remarks

We highlighted some key results in guessing the realisation of a source. We presented results for both the perfect secrecy and the key-rate constrained scenarios. An interesting point is that in the traditional Shannon-theoretic sense, perfect secrecy is attained if $R > H(P,X)$. However

a slightly higher key rate $R > E'(\rho) > H(P_X)$ is needed for perfect secrecy under guessing. The difference is because guessing performance is significantly affected by large deviations behaviour leading to a more stringent demand on the key rate. Just as there are so-called "universal" compression strategies that are asymptotically optimal for a wide class of sources, are there guessing strategies that are asymptotically optimal for a similarly wide class of sources? It turns out that the answer is yes. In fact, the association between compression and guessing is rather tight – every good compression strategy leads to a good guessing strategy and every guessing strategy leads to a good compression strategy.

In particular, it is known that guessing in the increasing order of Lempel-Ziv lengths (or any other asymptotically optimal compression scheme) is a robust guessing strategy for a wide range of sources called unifilar[1] sources. In the key-rate constrained case, this strategy interlaced with the exhaustive key search attack does the trick. We refer the reader to [5] for details.

We end the paper with this short list of interesting open questions.

- While there are asymptotically optimal robust guessing attacks for unifilar sources, are there asymptotically optimal encryption strategies that do not depend on knowledge of source statistics?
- Are Lempel-Ziv based attacks asymptotically optimal for stationary and ergodic sources?
- What is the degradation on the attacker's performance when only a noisy version of the cryptogram is available to the attacker?

References

[1] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, no. 3, pp. 565–715, Oct. 1949.

[2] N. Merhav and E. Arikan, "The Shannon cipher system with a guessing wiretapper," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 1860–1866, Sep. 1999.

[3] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. 24, no. 5, pp. 530–536, Sept. 1978.

[4] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 99–105, Jan. 1996.

[5] R. Sundaresan, "Guessing based on length functions," DRDO-IISc Programme in Advanced Mathematical Engineering, Tech. Rep. TR-PME-2007-02, Feb. 2007.

[1] Unifilar sources are finite state sources where the next state is a deterministic invertible function of the previous state and current output. Markov sources whose state is a finite past form an important subclass of unifilar sources. See [5].