3.3. Estimating the keyword length of a Vigenère cipher

Stochastic model:

$\mathcal{A} = \{0, \dots, m-1\}$ Alphabet

k keyword length, n message length, $k \mid n$

$$M = (M_1, \dots, M_k, M_{k+1}, \dots, M_{2k}, M_{2k+1}, \dots, M_{3k}, \dots, M_n)$$

$$\oplus \; K = (K_1, \dots, K_k, K_1, \dots, K_k, K_1, \dots, K_k \dots K_k)$$

$$C = (C_1, \dots, C_k, C_{k+1}, \dots, C_2 \; - \; - \; - \; - \; C_n)$$

Assumption: M_i i.i.d., $P(M_i = \ell) = p_\ell$ (known)

K_i i.i.d., $P(K_i = \ell) = \frac{1}{m}$

I_C : index of coincidence, $I_C = \frac{1}{\binom{n}{2}} \sum_{i<j} Y_{ij}$

$$K_M = \sum_{\ell=0}^{m-1} p_\ell^2 \qquad Y_{ij} = \begin{cases} 1, & C_i = C_j \\ 0, & \text{otherwise} \end{cases}$$

Lemma 3.5.

$$E(I_C) = \frac{1}{k(n-1)} \left[(n-k) K_M + n(k-1) \frac{1}{m} \right] \quad (*)$$

Outline of proof:

Consider 2 cases

1. $i \equiv j \pmod k \quad \Leftrightarrow \quad i \sim j$

$$E(Y_{ij}) = \sum_{\ell=0}^{u-1} p_\ell^2 = \kappa_M$$

2. $i \not\equiv j \pmod k \quad \Leftrightarrow \quad i \not\sim j$

$$E(Y_{ij}) = \frac{1}{u}$$

Finally:

$$E(\bar{I}_c) = \frac{1}{\binom{u}{2}} \sum_{i<j} E(Y_{ij})$$

$$= \frac{1}{\binom{u}{2}} \left[\sum_{\substack{i<j \\ i\sim j}} E(Y_{ij}) + \sum_{\substack{i<j \\ i\not\sim j}} E(Y_{ij}) \right]$$

$$= (*) \qquad \blacksquare$$

Resolve $(*)$ for k

$$k = \frac{u\left(\kappa_M - \frac{1}{u}\right)}{(u-1)\,E(\bar{I}_c) + \kappa_M - \frac{u}{u}}$$

Application: Given a ciphertext $(c_1, \dots, c_u)$

Estimate $E(\bar{I}_c)$ by $\hat{\bar{I}}_c = \frac{1}{u(u-1)} \sum_{\ell=0}^{u-1} u_\ell (u_\ell - 1)$

$\hat{\bar{I}}_c \to E(\bar{I}_c)$

Use $\hat{I}_c$ in the formula above.

$$\hat{k} = \frac{n\left(\kappa_M - \frac{1}{n}\right)}{(n-1)\,\hat{I}_c + \kappa_M - \frac{n}{n}}$$

In German: $\kappa_M = 0.0762$, $n = 26$

Hence:

$$\hat{k} = \frac{0.0377n}{(n-1)\,\hat{I}_c - 0.0385n + 0.0762}$$

If k is known, write C as follows

$$\tilde{C} = \begin{pmatrix} c_1 & \cdots & c_k \\ c_{k+1} & \cdots & c_{2k} \\ \vdots & & \vdots \\ c_{sk+1} & \cdots & c_n \end{pmatrix}$$

The columns are monoalphabetic, apply frequency analysis to the columns.

3.4. Vigenere cipher with running key

$$\begin{array}{ll} & a_1 \ldots\ldots a_n \\ \oplus & s_1 \ldots\ldots s_n \quad \text{(taken from a book)} \\ \hline & c_1 \ldots\ldots c_n \end{array}$$

Frequency analysis is still possible, if $s_1, \ldots, s_n$ is from a language.

Model: M_i r.v., model occurrence of plaintext char.
K_i r.v., " " " key char.

Consider: E, T, A, O, I, N, S (58%)

$P(M_i \in \{E, \ldots, S\}) \approx 0.58 \approx P(K_i \in \{E, \ldots, S\})$

Hence:

$P((M_i, K_i) \in \{E, \ldots, S\}^2) = 0.58^2 = 0.3249$

Appr. 1/3 of all ciphertext char. is generated by a combination of most frequent char.

Defense against this attack:

random key stream → one time pad

However, never use a key twice.

$(a_1, \ldots, a_n) \oplus (k_1, \ldots, k_n) = (c_1, \ldots, c_n)$

$(b_1, \ldots, b_n) \oplus (k_1, \ldots, k_n) = (d_1, \ldots, d_n)$

Oscar / Eve

$(c_i - d_i) \bmod 26 = (a_i - b_i) \bmod 26$

vulnerable to the above frequency attack.

4. Entropy and Perfect Secrecy

4.1. Entropy

Consider random experiments, e.g.,

(0.9, 0.05, 0.05)

(0.33, 0.33, 0.34)

A measure of uncertainty or information.

The right measure was introduced by Shannon (49)

Formal description

X : discrete r.v. with finite support $\mathcal{X} = \{x_1, \dots, x_m\}$,

distribution : $P(X = x_i) = p_i$, $i = 1, \dots, m$

D.4.1. Let $c > 1$ be an arbitrary constant.

$$H(X) = -\sum_{i=1}^{m} p_i \log_c p_i$$

$$= -\sum_{i=1}^{m} P(X = x_i) \log_c P(X = x_i)$$

is called entropy (of X).

Convention: $0 \cdot \log 0 \doteq 0$, omit c, but fix c.

Analogous definition for
(X,Y) with support $\mathcal{X}\times\mathcal{Y} = \{x_1,\dots,x_m\}\times\{y_1,\dots,y_d\}$
distribution $P(X=x_i, Y=y_j) = p_{ij}$

D. 4.2.

a) $$H(X,Y) = -\sum_{i,j} P(X=x_i, Y=y_j) \log P(X=x_i, Y=y_j)$$
$$= -\sum_{i,j} p_{ij} \log p_{ij}$$
is called (joint) entropy of (X,Y).

b) $$H(X|Y) = -\sum_{j=1}^{d} P(Y=y_j) \sum_{i=1}^{m} P(X=x_i|Y=y_j) \log P(X=x_i|Y=y_j)$$
$$= -\sum_{i,j} P(X=x_i, Y=y_j) \log P(X=x_i|Y=y_j)$$
is called conditional entropy of equivocation.

Th. 4.3. (Properties)

a) $0 \overset{(i)}{\le} H(X) \overset{(ii)}{\le} \log m$

"$=$" in (i) $\Leftrightarrow \exists x_i : P(X=x_i) = 1$ (singleton distribution)

"$=$" in (ii) $\Leftrightarrow P(X=x_i) = \frac{1}{m}$ (uniform distribution) for all $i = 1,\dots,m$

b) $0 \overset{(i)}{\le} H(X|Y) \overset{(ii)}{\le} H(X)$

"$=$" in (i) $\Leftrightarrow P(X=x_i|Y=y_j) = 1$ $\forall i,j : P(X=x_i, Y=y_j) > 0$

"$=$" in (ii) $\Leftrightarrow$ X, Y stoch. independent.