3.3. Estimating the key length of a Vigenère cipher

Stochastic model:

\( X = \{0, \ldots, m^2 - 1\} \) Alphabet

- \( k \) key word length
- \( u \) message length, \( kn \)

\( M = (M_1, \ldots, M_k, M_{k+1}, \ldots, M_{kn} ) \)

\( K = (K_1, \ldots, K_k, K_{k+1}, \ldots, K_{kn} ) \)

\( C = (C_1, \ldots, C_k, C_{k+1}, \ldots, C_{kn} ) \)

- \( M_i \) i.i.d., \( P(M_i = e) = \rho_e \) (known)
- \( K_i \) i.i.d., \( P(K_i = e) = \frac{1}{m} \)

\( I_c \) index of coincidence,

\[
I_c = \frac{1}{(u^2 - u)} \sum_{i<j} X_{ij}, \quad X_{ij} = \begin{cases} 1, & C_i = C_j \text{ or } \gamma_i = \gamma_j \\ 0, & \text{otherwise} \end{cases}
\]

\[
K_M = \sum_{e=0}^{u-1} \rho_e^2
\]

\[
\text{Lemma 3.5.} \quad E(I_c) = \frac{1}{k(u-1)} \left[ (u-k) K_M + u(k-1) \frac{1}{m} \right] (*)
\]
Outline of the proof.

Consider 2 cases:

1.) \( i \equiv j \pmod{k} \)
\[
E(Y_{ij}) = \sum_{E\in E} p_e^k = k_M
\]

2.) \( i \not\equiv j \pmod{k} \)
\[
E(Y_{ij}) = \frac{1}{m}
\]

Finally:
\[
E(I_c) = \frac{1}{(\frac{1}{2})} \sum_{i \neq j} E(Y_{ij})
\]
\[
= \frac{1}{(\frac{1}{2})} \left[ \sum_{i \neq j} E(Y_{ij}) + \sum_{i \neq j} E(Y_{ij}) \right]
\]
\[
= (\star)
\]

We are interested in \( k \). Solve (\star) for \( k \):
\[
(n-1) E(I_c) = \frac{1}{k} \left( n(k_M - \frac{1}{m}) - (k_M - \frac{n}{m}) \right)
\]
\[
k = \frac{n \left( k_M - \frac{1}{m} \right)}{(n-1) E(I_c) + k_M - \frac{n}{m}}
\]
Application: Estimate \( E(I_c) \) by \( I_c \)

\[
I_c = \frac{1}{u(u-1)} \sum_{e=1}^{u} u_e (u_e - 1)
\]

By Lemma 3.3. \( \Pr.: I_c \rightarrow E(I_c) \) (u->oo) a.e.

In German: \( k_M = 0.0762, u = 26 \)

Hence:

\[
h = \frac{0.0377u}{(u-1)I_c - 0.0385u + 0.0762}
\]

If \( k \) is known, write \( C \) as follows

\[
\begin{pmatrix}
C_1 & \ldots & C_k \\
C_{k+1} & \ldots & C_{2k} \\
\vdots & \ddots & \vdots \\
C_{5k+1} & \ldots & C_{n}
\end{pmatrix}
\]

The columns are monoalphabetic, apply frequency analysis to the columns.
3.4. Vigenère cipher with running key

\[ a_1 \quad a_2 \quad \ldots \quad a_n \]

\[ S_1 \quad S_2 \quad \ldots \quad S_n \quad \text{(taken from a book)} \]

\[ C_1 \quad C_2 \quad \ldots \quad C_n \]

Frequency attack is possible, if \((S_1, \ldots, S_n)\)

is from a nat. language.

Model: \(M_i\) r.v.s occurrence of plaintext char \(\sim \) stock.
\(K_i\) r.v.s 4 key char.

Consider most freq. char. : \(E, T, A, O, I, N, S\) (57%)

\[ P(M_i \mid K_i) \in \{E, \ldots, S\}^2 \approx 0.57^2 = 0.3249 \]

About \(\frac{1}{3}\) of all cipher char. are obtained

by "adding" 2 of the most frequent char.
Example:

```
ITISNICETOLEARNABOUTCRYPTOGRAPHY
ETAISANELEMENTOFTHEGREEKALPHABET
MMIAFIPIIESXINKBFUVYZTVCTZTVYAYQLR
```

Investigating the first five ciphertext characters:

```
ABCDEFGHIJKLMNOPQRSTUVWXYZ
MLKJIHGFEDCBAZYXWVUTSRQPON
MMMMMMMMMMMMMMMMMMMMMMMMMMM

ABCDEFGHIJKLMNOPQRSTUVWXYZ
IHZFEDCBAZYXWVUTSRQPONMLKJ
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ABCDEFGHIJKLMNOPQRSTUVWXYZ
AZYXWVUTSRQPONMLKJIHGFE
AAAAAAA AAA AAA AAA AAA AAA AAA AAA AAA AAA

ABCDEFGHIJKLMNOPQRSTUVWXYZ
FEDCBAZYXWVUTSRQPONMLKJIHG
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

There are $3 \cdot 3 \cdot 4 \cdot 2 = 72$ pairs out of the most frequent characters.

Some of them are:

```
EIAAN ... ITAAN ... ITAINN ... ITISN
IEIAS ... ETIAS ... ETESS ... ETANS ... ETAIS
```
Defense against this attack: random key stream

→ one time pad

However, never use the same key twice. Otherwise:

\[(a_1, \ldots, a_n) \oplus (k_1, \ldots, k_n) = (c_1, \ldots, c_n)\]
\[(b_1, \ldots, b_n) \oplus (k_1, \ldots, k_n) = (a_1, \ldots, a_n)\]

Oscar:

\[(c_i - a_i) \mod 26 = (a_i - b_i) \mod 26\]

vulnerable to the above attack.
4. Entropy and Perfect Secrecy

4.1. Entropy

Consider random experiments, e.g.,

\[(0.9, 0.05, 0.05)\]
\[(0.33, 0.33, 0.34)\]

We aim at a measure of

\[\begin{align*}
\text{uncertainty about the outcome (before)} \\
\text{information gained by the outcome (after)}
\end{align*}\]

The right measure was introduced by Shannon (1949).

Formal description

\[X: \text{discrete r.v. with finite support } X = \{x_1, \ldots, x_m\}\]

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

Def. 4.1. Let \(c > 1\) constant,

\[H(X) = -\sum_{i=1}^{m} p_i \log p_i = -\sum_{i=1}^{m} P(X=x_i) \log_e P(X=x_i)\]

is called entropy of \(X\) (or \((p_1, \ldots, p_m))\).

Convention: \(0 \cdot \log 0 = 0\), omit \(c\) but fix it.
Analogously for 2-dim. random variables 

\((X, Y)\) with support \(X \times Y = \{x_1, \ldots, x_m\} \times \{y_1, \ldots, y_n\}\)

distribution \(P(X=x_i, Y=y_j) = p_{ij}\)

**Def. 4.2.**

a) \(H(X, Y) = - \sum_{ij} p(X=x_i, Y=y_j) \log P(X=x_i, Y=y_j)\)

\[= - \sum_{ij} p_{ij} \log p_{ij}\]

is called (joint) entropy of \(X, Y\).

b) \(H(X|Y) = - \sum_{i} p(X=x_i|Y=y_j) \sum_{j} p(X=x_i|Y=y_j) \log P(X=x_i|Y=y_j)\)

\[= - \sum_{ij} P(X=x_i, Y=y_j) \log P(X=x_i|Y=y_j)\]

is called conditional entropy or equivocation.
Theorem 4.3.

(1) \( 0 \leq H(X) \leq \log m \)

a) \( i ) \iff \exists x_i : P(X=x_i) = 1 \)

b) \( \begin{align*}
(1) & \Rightarrow H(X,Y) \leq H(X) \\
(2) & \Rightarrow H(X,Y) \leq H(X) + H(Y) \\
\end{align*} \)

(2) \( \begin{align*}
(1) & \iff \text{Y is fully dependent on } X \\
(2) & \iff X, Y \text{ stoch. independent} \\
\end{align*} \)

(2) \( \begin{align*}
H(X,Y) & = H(X) + H(Y|X) = H(Y) + H(X|Y) \\
& \text{(chain rule)} \\
\end{align*} \)

Proof. any book on information theory.