

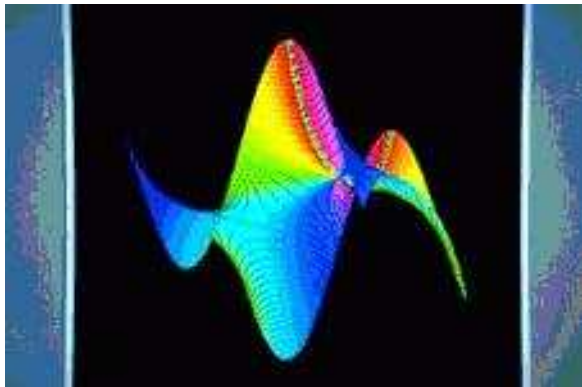
Analytic Information Theory: Analysis, Algorithms, and Beyond*

W. Szpankowski

Department of Computer Science
Purdue University
W. Lafayette, IN 47907

June 25, 2010

AofA and **IT** logos



AofA School, Vienna, 2010

*Research supported by NSF [Science & Technology Center](#), and [Humboldt Foundation](#).

Outline

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
 - Asymptotic Equipartition Theorem
 - Jointly typical Sequences
 - Capacity
3. Source Coding
4. Redundancy: Known Sources
 - Shannon and Huffman Coding (sequences modulo 1)
 - Non-Prefix Codes (saddle point method)
 - Tunstall Code (Mellin transform)
5. Minimax Redundancy: Universal (unknown) Sources
 - (a) Universal Memoryless Sources (Tree-like gen. func.)
 - (b) Universal Markov Sources (Balance matrices)
 - (c) Universal Renewal Sources (Combinatorial calculus)
6. Appendix A: Mellin Transform
7. Appendix B: Saddle Point Method

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
 - Asymptotic Equipartition Theorem
 - Jointly typical Sequences
 - Capacity
3. Source Coding
4. Redundancy: Known Sources
5. Minimax Redundancy: Universal (unknown) Sources

Three Jewels of Shannon

Theorem 1 & 3. (Shannon 1948; Lossless & Lossy Data Compression)

Lossless Compression: compression **bit rate** \geq source **entropy** $H(X)$;

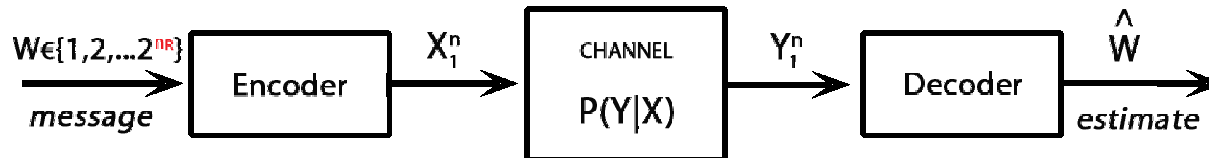
Lossy Compression: For distortion level D :
lossy **bit rate** \geq **rate distortion** function $R(D)$

Theorem 2. (Shannon 1948; Channel Coding)

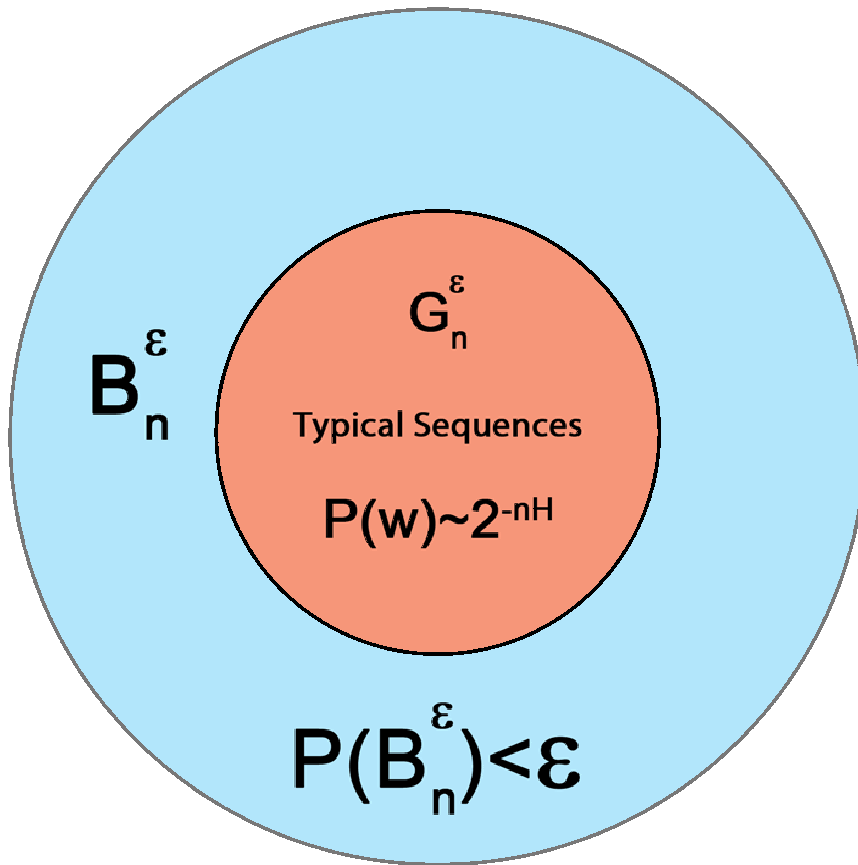
In Shannon's words:



It is possible to **send information** at the **capacity** through the channel **with as small a frequency of errors as desired** by proper (**long**) encoding. This statement is **not true** for any rate **greater than** the capacity.



Theorem 1: AEP and Typical Sequences



Shannon-McMillan-Breiman:

$$-\frac{1}{n} \log P(X_1^n) \rightarrow H(X) \quad (\text{pr.})$$

$H(X)$ is the **entropy rate**.

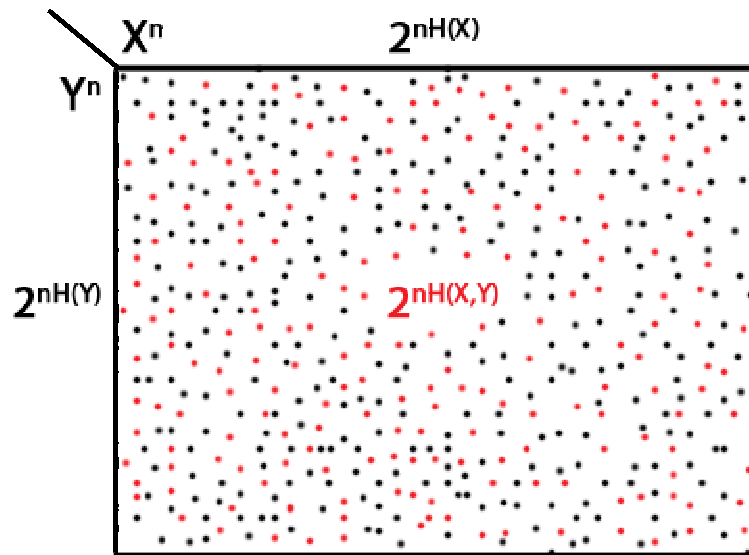
Asymptotic Equipartition Property: Sequences of length n can be partitioned into

good set G_n^ϵ $P(w) \sim 2^{-nH(X)}$, $w \in G_n^\epsilon$

bad set B_n^ϵ $P(B_n^\epsilon) < \epsilon$.

Also, $|G_n^\epsilon| \sim 2^{nH(X)}$.

Theorem 2: Shannon Random Decoding Rule



There are $2^{nH(X)}$ X -typical sequences

There are $2^{nH(Y)}$ Y -typical sequences

There are $2^{nH(X,Y)}$ jointly X,Y -typical pair of sequences

Decoding Rule: Declare that sequence sent X is the one that is jointly typical with the received sequence Y provided there is unique X satisfying this property!

Sketch of Proof: Channel Capacity Theorem

1. With high probability (whp), there is a **jointly typical** pair (X, Y) .
2. The probability that there is another **jointly typical** pair is $2^{-nI(X,Y)}$.
Indeed:
 - there are $2^{nH(X)}$ and $2^{nH(Y)}$ typical sequences X^n and Y^n , respectively.
 - there are $2^{nH(X,Y)}$ **jointly typical** pairs (X, Y) .The probability of error (more than one typical pair is):

$$\frac{2^{nH(X,Y)}}{2^{nH(X)+H(Y)}} = 2^{-nI(X,Y)}.$$

3. Probability of **error** when 2^{nR} messages are sent is approximately

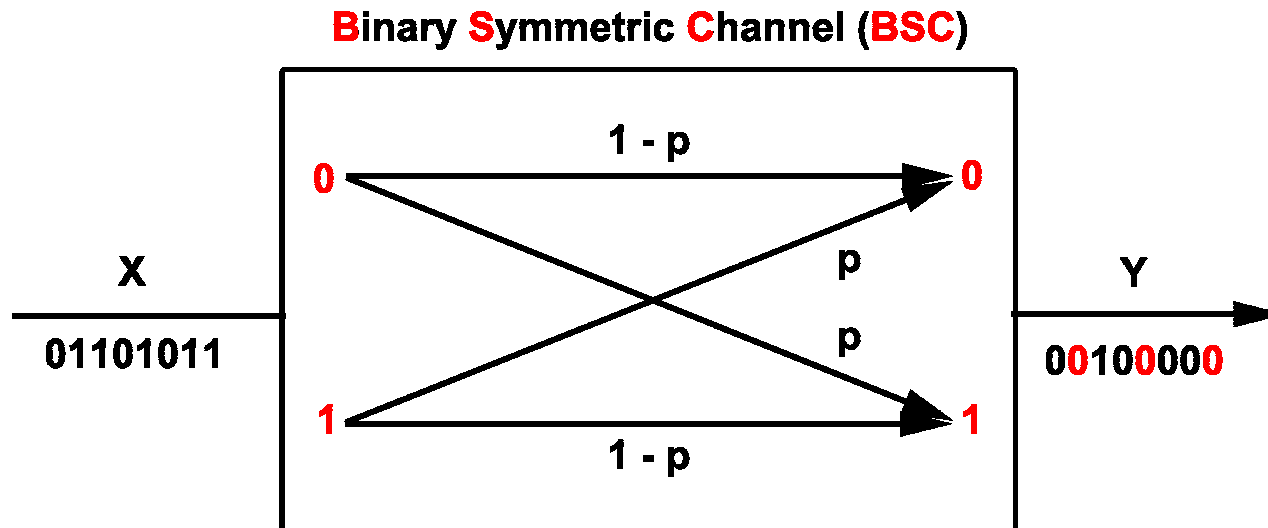
$$\min P(\text{error}) \sim 2^{-n(\sup_{P(X)} I(X,Y) - R)} = 2^{-n(C-R)}.$$

4. In conclusion:

$$R < C \quad P(\text{error}) \sim 2^{-n\delta}$$

$$R > C \quad P(\text{error}) \rightarrow 1.$$

Capacity of BSC



Capacity:

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p). \end{aligned}$$

The **capacity** is achieved for the **uniform** input distribution. Thus

$$C = 1 - H(p).$$

Temporal Capacity for BSC (Exercise)

1. Each bit incurs a delay, T with known probability distribution: $F(t) = P(T < t)$. If a bit arrives after a given deadline τ , it is dropped.
2. The longer it takes to send a bit, the lower the probability of success, which we denote by $\Phi(\varepsilon, t)$ for $t < \tau$ (e.g., $\Phi(\varepsilon, t) = (1 - \varepsilon)^t$).
3. Define $P(x|x) = \int_0^\tau \Phi(\varepsilon, t) dF(t)$: prob. of a successfully transmission:

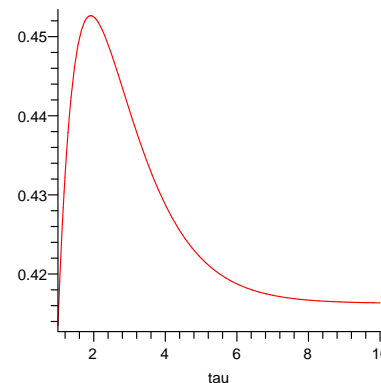
$$P(y|x) = \begin{cases} \alpha := 1 - F(\tau) & y = \text{erasure} \\ P(x|x) & \text{if } x = y \\ 1 - \alpha - P(x|x) & \text{if } x \neq y. \end{cases}$$

4. Define: $\alpha = 1 - F(\tau)$ and $\rho := \frac{P(x|x)}{(1-\alpha)}$.

$$H(Y|X) = H(\alpha) + (1 - \alpha)H(\rho) \text{ and}$$

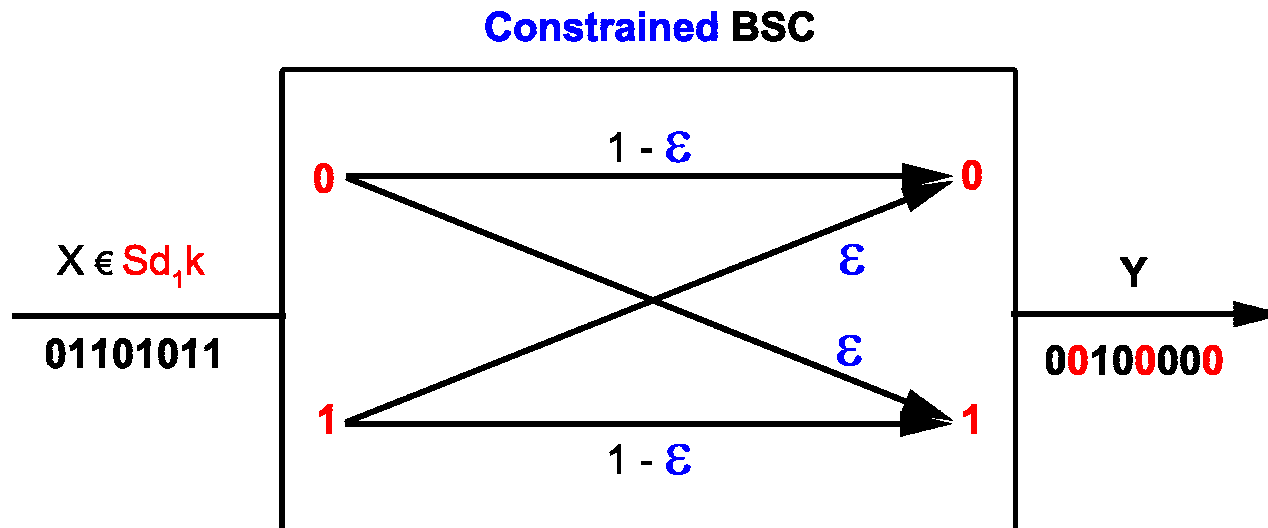
$$H(Y) = H(\alpha) + (1 - \alpha)H(p\rho + \bar{p}\bar{\rho}).$$

Then:



$$C(\tau) = [(1 - P(T > \tau)][1 - H(\rho)].$$

Noisy Constrained Channel

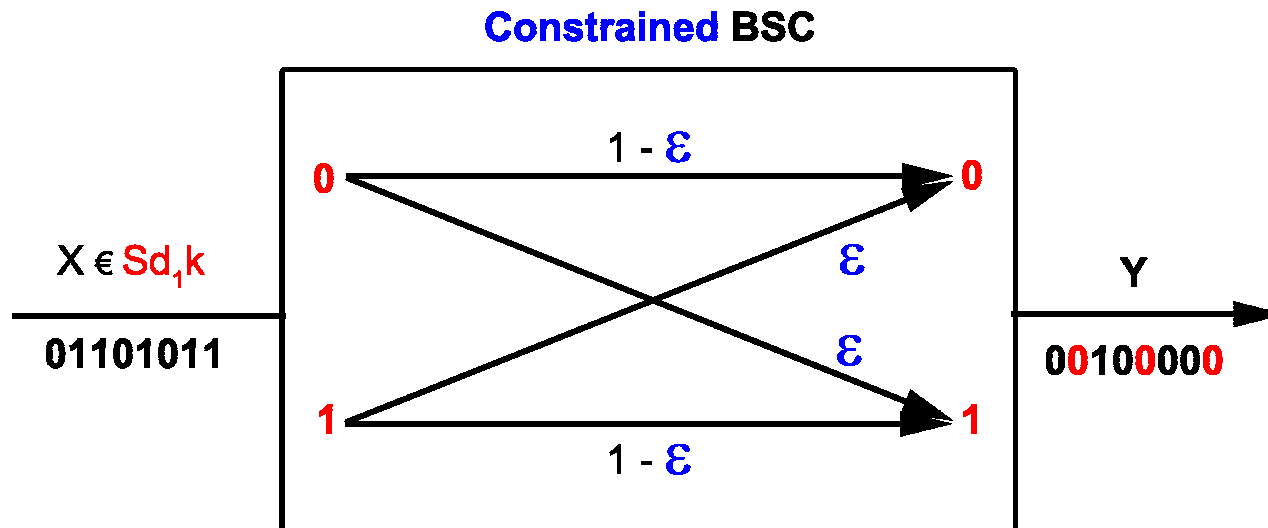


Let \mathcal{S} denote the set of binary **constrained sequences** of length n . Here:

$$\mathcal{S}_{d,k} = \{(d,k) \text{ sequences}\},$$

i.e., **no** sequence contains **a run of zeros shorter than d** or **longer than k** .

Noisy Constrained Channel



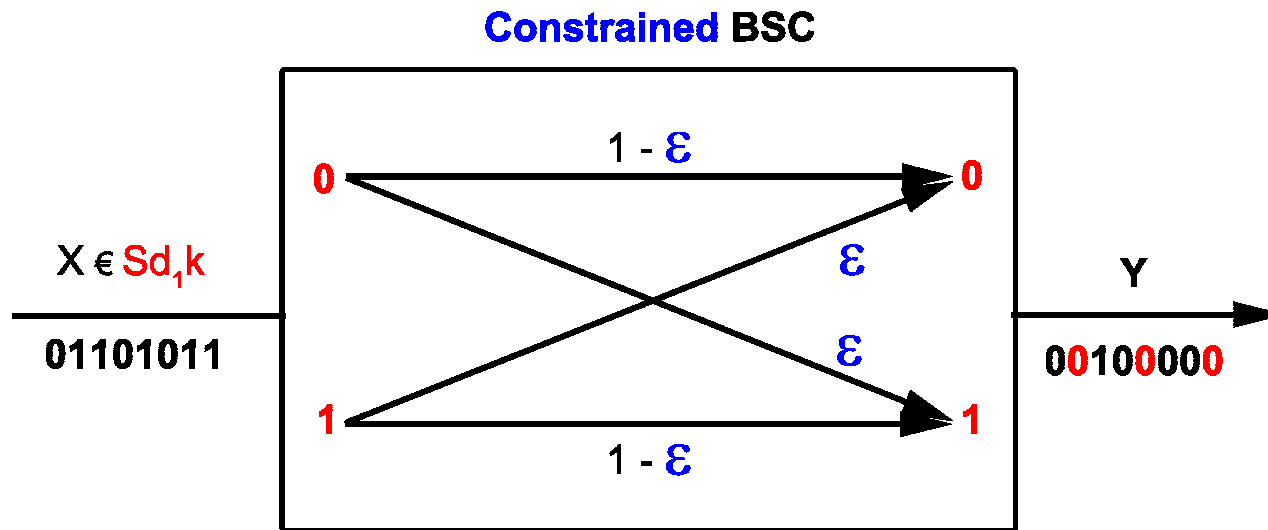
Let \mathcal{S} denote the set of binary **constrained sequences** of length n . Here:

$$\mathcal{S}_{d,k} = \{(d,k) \text{ sequences}\},$$

i.e., **no** sequence contains **a run of zeros shorter than d** or **longer than k** .

Sequence $X \in \mathcal{S}_{(d,k)}$ can be represented as a **Markov process.**

Noisy Constrained Channel



Let \mathcal{S} denote the set of binary **constrained sequences** of length n . Here:

$$\mathcal{S}_{d,k} = \{(d,k) \text{ sequences}\},$$

i.e., **no** sequence contains **a run of zeros shorter than d or longer than k** .

Sequence $X \in \mathcal{S}_{(d,k)}$ can be represented as a Markov process.

$C(\mathcal{S}, \varepsilon)$ – **noisy constrained capacity** defined as

$$C(\mathcal{S}, \varepsilon) = \sup_{X \in \mathcal{S}} I(X; Y) = \lim_{n \rightarrow \infty} \frac{1}{n} \sup_{X_1^n \in \mathcal{S}_n} I(X_1^n, Y_1^n).$$

This is/was an open problem since Shannon.

Entropy of HMM

1. Mutual information

$$I(X; Y) = H(Y) - H(Y|X).$$

where

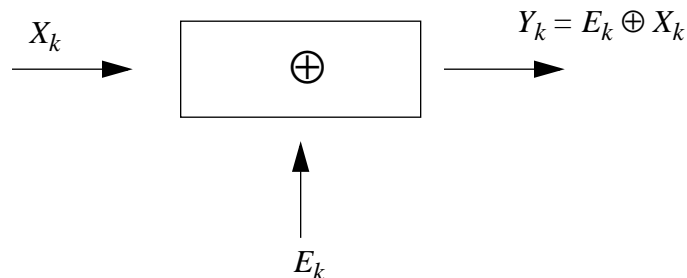
$$Y_k = X_k \oplus E_k, \quad k \geq 1, \quad (\oplus \text{ is exclusive-or}),$$

$E = \{E_k\}_{k \geq 1}$ is a binary i.i.d. representing **noise** such that $P(E_i = 1) = \varepsilon$.

2. Observe that

$$H(Y|X) = H(E) = H(\varepsilon)$$

hence we **must** find the **entropy** of $H(Y)$, that is, the entropy of a **hidden Markov process** since a (d, k) sequence can be generated as an output of a k th order Markov process.



How to compute the **entropy** $H(Y)$?

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. [Source Coding](#)
4. Redundancy: Known Sources
5. Minimax Redundancy: Universal (unknown) Sources

Source Coding

A **source code** is a **bijjective mapping**

$$C : \mathcal{A}^* \rightarrow \{0, 1\}^*$$

from sequences over the alphabet \mathcal{A} to set $\{0, 1\}^*$ of binary sequences.

The **basic problem** of **source coding** (i.e., *data compression*) is to **find codes with shortest descriptions (lengths)** either on *average* or for *individual sequences*.

Three Basic Types of Source Coding:

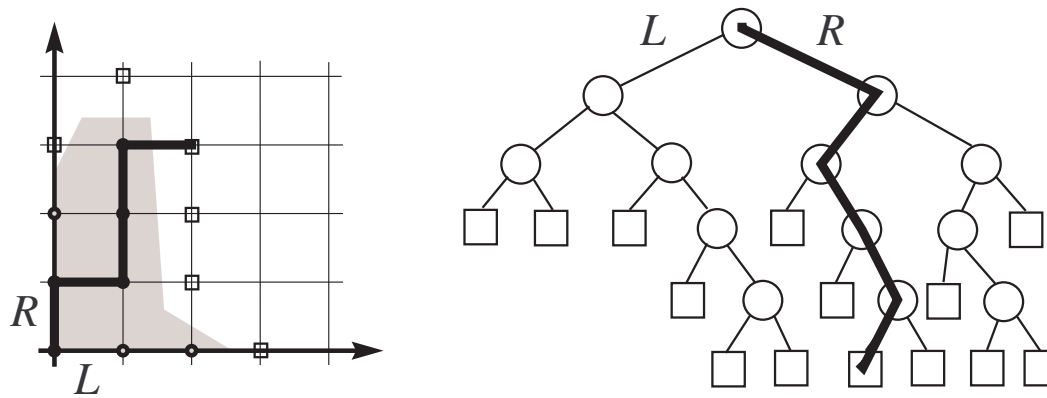
- **Fixed-to-Variable (FV)** length codes (e.g., **Huffman** and **Shannon** codes).
- **Variable-to-Fixed (VF)** length codes (e.g., **Tunstall** and **Khodak** codes).
- **Variable-to-Variable (VV)** length codes (e.g., **Khodak VV** code).



Preliminary Results

Prefix code is such that no codeword is a prefix of another codeword.

Tree and lattice representations:



Notation: For a source model \mathcal{S} and a code \mathcal{C} we let:

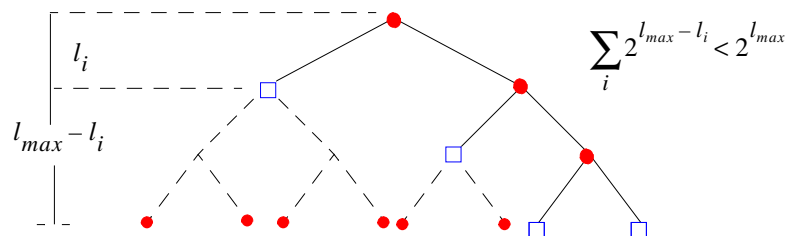
- $P(x)$ be the probability of $x \in \mathcal{A}^*$;
- $L(\mathcal{C}, x)$ be the code length for the source sequence $x \in \mathcal{A}^*$;
- Entropy $H(P) = - \sum_{x \in \mathcal{A}^*} P(x) \lg P(x)$.

Quantities are expressed in binary logarithms written $\lg := \log_2$.

Prefix Codes

Kraft's Inequality

A **binary** code is a **prefix code** iff the code lengths l_1, l_2, \dots, l_N satisfy



$$\sum_{i=1}^N 2^{-l_i} \leq 1.$$

Barron's lemma

For any sequence a_n of positive constants satisfying $\sum_n 2^{-a_n} < \infty$

$$\Pr\{L(X) < -\log P(X) - a_n\} \leq 2^{-a_n},$$

and therefore

$$L(X) \geq -\log P(X) - a_n \quad (\text{a.s.}).$$

Proof: We argue as follows:

$$\begin{aligned} \Pr\{L(X) < -\log_2 P(X) - a_n\} &= \sum_{x: P(x) < 2^{-L(x) - a_n}} P(x) \\ &\leq \sum_{x: P(x) < 2^{-L(x) - a_n}} 2^{-L(x) - a_n} \\ &\leq 2^{-a_n} \sum_x 2^{-L(x)} \leq 2^{-a_n}. \end{aligned}$$

Shannon Lower Bound

Shannon First Theorem

For any prefix code the average code length $\mathbf{E}[L(C, X)]$ cannot be smaller than the entropy of the source $H(P)$, that is,

$$\mathbf{E}[L(C_n, X)] \geq H(P).$$

Proof: Let $K = \sum_x 2^{-L(x)} \leq 1$, and $L(C, x) := L(C)$. Then

$$\begin{aligned} \mathbf{E}[L(C, X)] & - H(P) = \\ & = \sum_{x \in \mathcal{A}^*} P(x) L(x) + \sum_{x \in \mathcal{A}^*} P(x) \log P(x) \\ & = \sum_{x \in \mathcal{A}^*} P(x) \log \frac{P(x)}{2^{-L(x)}/K} - \log K \\ & \geq 0 \end{aligned}$$

since $\log x \leq x - 1$ for $0 < x \leq 1$ or the divergence is nonnegative, while $K \leq 1$ by Kraft's inequality.

Exercise: There exists at least one sequence \tilde{x}_1^n such that $L(\tilde{x}_1^n) \geq -\log_2 P(\tilde{x}_1^n)$.

Redundancy

Known Source P .

The pointwise redundancy $R(x)$ and the average redundancy \bar{R} :

$$\begin{aligned}R(x) &= L(C, x) + \lg P(x) \\ \bar{R} &= \mathbf{E}[L(C, X)] - H(P) \geq 0\end{aligned}$$

Optimal Code:

$$\min_L \sum_x L(x) P(x) \quad \text{subject to} \quad \sum_x 2^{-L(x)} \leq 1.$$

Solution: By Lagrangian multipliers we find $L^{opt}(x) = -\lg P(x)$.

The smaller the redundancy is, the better (closer to the optimal) the code is.

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
 - Shannon and Huffman Coding
 - Non-Prefix Codes
 - Tunstall Code
5. Minimax Redundancy: Universal (unknown) Sources

Redundancy for Huffman's Code

We consider **fixed-to-variable length codes**; in particular, **Huffman's code**.

For a **known** source P , we consider **fixed** length sequences $x_1^n = x_1 \dots x_n$.

Huffman Code: The following **optimization problem**

$$\bar{R}_n = \min_{C_n \in \mathcal{C}} \mathbf{E}_{x_1^n} [L(C_n, x_1^n) + \log_2 P(x_1^n)].$$

is solved by **Huffman's code**.

We study the **average redundancy** for a **binary memoryless sources** with p denoting the probability of generating "0" and $q = 1 - p$.

In 1994 **Stubbley** proposed the following for **Huffman's average redundancy**

$$\bar{R}_n^H = 2 - \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \langle \alpha k + \beta n \rangle - 2 \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} 2^{-\langle \alpha k + \beta n \rangle} + o(1).$$

where

$$\alpha = \log_2 \left(\frac{1-p}{p} \right), \quad \beta = \log_2 \left(\frac{1}{1-p} \right)$$

and $\langle x \rangle = x - \lfloor x \rfloor$ is the **fractional part** of x .

Main Result

Theorem 1 (W.S., 2000). Consider the *Huffman block* code of length n over a *binary memoryless source* with $p < \frac{1}{2}$. Then as $n \rightarrow \infty$

$$\bar{R}_n^H = \begin{cases} \frac{3}{2} - \frac{1}{\ln 2} + o(1) \approx 0.057304 & \alpha \text{ irrational} \\ \frac{3}{2} - \frac{1}{M} \left(\langle \beta M n \rangle - \frac{1}{2} \right) - \frac{1}{M(1-2^{-1/M})} 2^{-\langle n \beta M \rangle / M} + O(\rho^n) & \alpha = \frac{N}{M} \end{cases}$$

where N, M are integers such that $\gcd(N, M) = 1$ and $\rho < 1$.

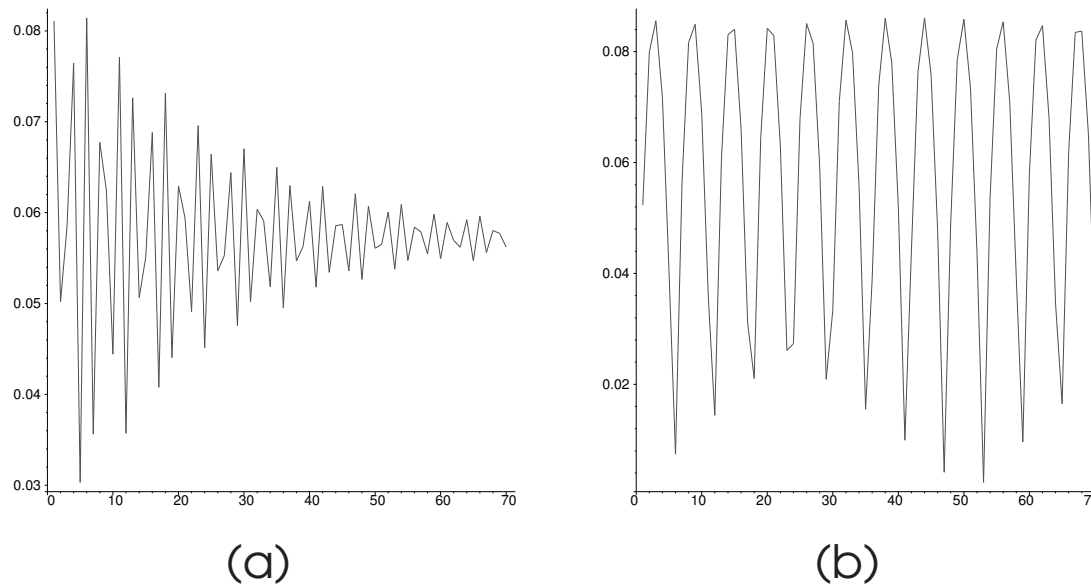


Figure 1: The average redundancy of Huffman codes versus block size n for: (a) irrational $\alpha = \log_2(1 - p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1 - p)/p$ with $p = 1/9$.

Why Two Modes: Shannon Code

Consider the **Shannon code** that assigns the length

$$L(C_n^S, x_1^n) = \lceil -\lg P(x_1^n) \rceil$$

to the **source sequence** x_1^n . Observe that

$$P(x_1^n) = p^k (1 - p)^{n-k}$$

where p is **known** probability of generating 0 and k is the number of 0s.

The **Shannon code redundancy** is

$$\begin{aligned} \bar{R}_n^S &= \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} \left(\lceil -\log_2(p^k (1 - p)^{n-k}) \rceil \right. \\ &\quad \left. + \log_2(p^k (1 - p)^{n-k}) \right) \\ &= 1 - \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} \langle \alpha k + \beta n \rangle \end{aligned}$$

where $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x , and

$$\alpha = \log_2 \left(\frac{1 - p}{p} \right), \quad \beta = \log_2 \left(\frac{1}{1 - p} \right).$$

Sketch of Proof

We need to understand **asymptotic behavior** of the following sum (cf. **Bernoulli distributed sequences modulo 1**)

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha k + y \rangle)$$

for fixed p and some Riemann integrable function $f : [0, 1] \rightarrow \mathbb{R}$.

Lemma 1. Let $0 < p < 1$ be a fixed real number and α be an **irrational number**. Then for every **Riemann integrable function** $f : [0, 1] \rightarrow \mathbb{R}$

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha k + y \rangle) = \int_0^1 f(t) dt,$$

where the convergence is uniform for all shifts $y \in \mathbb{R}$.

Lemma 2. Let $\alpha = \frac{N}{M}$ be a **rational number** with $\gcd(N, M) = 1$. Then for bounded function $f : [0, 1] \rightarrow \mathbb{R}$

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} f(\langle \alpha k + y \rangle) = \frac{1}{M} \sum_{l=0}^{M-1} f\left(\frac{l}{M} + \frac{\langle My \rangle}{M}\right) + O(\rho^n)$$

uniformly for all $y \in \mathbb{R}$ and some $\rho < 1$.

Uniformly Distributed Sequences Mod 1

1. Uniformly Distributed Sequences Mod 1: A sequence $x_n \in \mathbf{R}$ is said to be **Bernoulli uniformly distributed** modulo 1 (in short: B-u.d. mod 1) if for $0 < p < 1$

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} \chi_I(\langle x_k \rangle) = \lambda(I)$$

holds for every interval $I \subset \mathbf{R}$, where $\chi_I(x_n)$ is the characteristic function of I (i.e., it equals to 1 if $x_n \in I$ and 0 otherwise) and $\lambda(I)$ is the **Lebesgue measure** of I .

2. Weyl's Criterion:

A sequence x_n is B-u.d. mod 1 if and only if

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} e^{2\pi i m x_k} = 0$$

holds for all non-zero $m \in \mathbf{Z} - \{0\}$.

Proof. The proof is standard. Basically, it is based on the fact that by Weierstrass's *approximation theorem* every Riemann integrable function f of period 1 can be uniformly approximated by a trigonometric polynomial (i.e., a finite combination of functions of the type $e^{2\pi i m x}$).

Shannon Code: The Irrational Case

3. Let us return to the Shannon code redundancy. Two cases: α irrational and α rational.

4. We first consider α irrational.

To apply our previous results, we must show that $\langle \alpha k \rangle$ is B -u.d. mod 1. By Weyl's criterion

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} e^{2\pi i m(k\alpha)} &= \lim_{n \rightarrow 0} \left(p e^{2\pi i m \alpha} + q \right)^n \\ &= 0 \end{aligned}$$

provided α is irrational. Hence, by the previous theorem, with $f(t) = t$ and $y = \beta n$, we immediately obtain

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \langle \alpha k + \beta n \rangle = \int_0^1 t dt = \frac{1}{2}.$$

This proves that for α irrational

$$R_n^S = \frac{1}{2} + o(1).$$

Shannon Redundancy – Rational Case

Assume $\alpha = N/M$ where $\gcd(N, M) = 1$. Denote $p_{n,k} = \binom{n}{k} p^k q^{n-k}$.

$$\begin{aligned} S_n &= \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \left\langle k \frac{N}{M} + \beta n \right\rangle = \sum_{\ell=0}^{M-1} \sum_{m: k=\ell+mM \leq n} p_{n,k} \left\langle \ell \frac{N}{M} + N + \beta n \right\rangle \\ &= \sum_{\ell=0}^{M-1} \left\langle \frac{\ell}{M} + \beta n \right\rangle \sum_{m: k=\ell+mM \leq n} p_{n,k}. \end{aligned}$$

Lemma 3. For fixed $\ell \leq M$ and M , there exist $\rho < 1$ such that

$$\sum_{m: k=\ell+mM \leq n} \binom{n}{k} p^k (1-p)^{n-k} = \frac{1}{M} + O(\rho^n).$$

Proof. Let $\omega_k = e^{2\pi i k/M}$ for $k = 0, 1, \dots, M-1$ be the M th root of unity.

$$\frac{1}{M} \sum_{k=0}^{M-1} \omega_k^n = \begin{cases} 1 & \text{if } M|n \\ 0 & \text{otherwise.} \end{cases}$$

where $M|n$ means that M divides n . Then

$$\sum_{m: k=\ell+mM \leq n} \binom{n}{k} p^k q^{n-k} = \frac{1 + (p\omega_1 + q)^{n-\ell} + \dots + (p\omega_{M-1} + q)^{n-\ell}}{M} = \frac{1}{M} + O(\rho^n),$$

since $|(p\omega_r + q)| = p^2 + q^2 + 2pq \cos(2\pi r/M) < 1$ for $r \neq 0$.

Finishing the Rational Case

We shall use the following Fourier series; for real x

$$\langle x \rangle = \frac{1}{2} - \sum_{m=1}^{\infty} \frac{\sin 2\pi m x}{m\pi} = \frac{1}{2} - \sum_{m \in \mathbb{Z} - \{0\}} c_m e^{2\pi i m x}, \quad c_m = -\frac{i}{2\pi m},$$

Continuing the derivation and using the above lemma we obtain

$$\begin{aligned} S_n &= \frac{1}{M} \sum_{\ell=0}^{M-1} \left(\frac{1}{2} - \sum_{m \neq 0} c_m e^{2\pi i m (\ell/M + \beta n)} \right) = \frac{1}{2} - \sum_{m \neq 0} c_m e^{2\pi i m n \beta} \frac{1}{M} \sum_{\ell=0}^{M-1} e^{2\pi i m \frac{\ell}{M}} \\ &= \frac{1}{2} - \frac{1}{M} \sum_{m=kM \neq 0} c_{kM} e^{2\pi i kM \beta n} = \frac{1}{2} - \frac{1}{M} \left(\frac{1}{2} - \langle \beta n M \rangle \right). \end{aligned}$$

Now, having the above two results we easily establish that

$$\bar{R}_n^S = \begin{cases} \frac{1}{2} + o(1) & \alpha \text{ irrational} \\ \frac{1}{2} - \frac{1}{M} (\langle M n \beta \rangle - \frac{1}{2}) + O(\rho^n) & \alpha = \frac{N}{M} \end{cases}$$

Exercise. Using Stubbley's formula and tools discussed above, derive the redundancy of the Huffman code.

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
 - Shannon and Huffman Coding
 - Non-Prefix Codes
 - Tunstall Code
5. Minimax Redundancy: Universal (unknown) Sources

Do We Really Need Prefix FV Codes?

As argued in the [XXVIII Shannon Lecture](#), it is possible to attain average compression for FV codes lower than Huffman's code (P and n are known):

1. Avoiding symbol-by-symbol compression, we design **fixed-to-variable** code for the **whole file**, hence eliminating the need for **prefix codes**.

2. Applying **prefix condition** to **symbol-by-symbol** or **n -block supersymbol** is only **optimal** for the **linear term** but **not** for the **sublinear** term.

3. The **optimal FV** code performs **no blocking** but **encodes** a table that listing sequences in **decreasing probabilities**:

Define: $\pi_X(x) = \ell$ if x is the ℓ -th **most probable** element according to distribution P_X . The **the minimum average code length** is

$$L(X) = \mathbf{E}[\lceil \log_2 \pi_X(X) \rceil].$$

Do We Really Need Prefix FV Codes?

As argued in the [XXVIII Shannon Lecture](#), it is possible to attain average compression for FV codes lower than Huffman's code (P and n are known):

1. Avoiding symbol-by-symbol compression, we design **fixed-to-variable** code for the whole file, hence eliminating the need for prefix codes.
2. Applying prefix condition to symbol-by-symbol or n -block supersymbol is only optimal for the linear term but not for the sublinear term.
3. The optimal FV code performs no blocking but encodes a table that listing sequences in decreasing probabilities:

Define: $\pi_X(x) = \ell$ if x is the ℓ -th most probable element according to distribution P_X . The **the minimum average code length** is

$$L(X) = \mathbf{E}[\lfloor \log_2 \pi_X(X) \rfloor].$$

Example: Binary source with $p < 1 - p := q$; sequence $x_1^n = x_1 \dots x_n$:

$$\begin{array}{ccccccc}
 q^n \left(\frac{p}{q}\right)^0 & \geq & q^n \left(\frac{p}{q}\right)^1 & \geq & \dots & \geq & q^n \left(\frac{p}{q}\right)^n \\
 00 \dots 0 & & 00 \dots 1 & & \dots & & 11 \dots 1 \\
 \lfloor \log_2(1) \rfloor & & \lfloor \log_2(2) \rfloor & & \dots & & \lfloor \log_2(2^n) \rfloor
 \end{array}$$

Bounds on the Average Rate of Fixed-to-Variable Codes

Upper Bound: $\mathbf{E}[L(X)] \leq H(X)$ (Wyner).

Lower Bounds:

Theorem 2 (W.S., Verdu, 2009). Define the *monotonically increasing function* by:

$\psi(x) = x + (1 + x) \log_2(1 + x) - x \log_2 x$. Then

$$\psi^{-1}(H(X)) \leq L(X).$$

Looser Bounds: Notice that: $\psi(x) \leq x + \log_2(e + ex)$,
then **Alon & Orlicsky** bound follows:

$$H(X) - \log_2(H(X) + 1) - \log_2 e \leq L(X)$$

By *monotonic increasing*: $h(x) = (1 + x) \log(1 + x) - x \log x$, we conclude

$$L(X) \geq H(X) - (1 + H(X)) \log_2(1 + H(X)) - H(X) \log_2 H(X)$$

This can be also found in **Blundo & de Prisco**.

Binary Memoryless Source (WS, 2005)

Consider again a binary memoryless source with p probability of transmitting a 1. Let $x_1^n = x_1 \dots x_n$.

There are $\binom{n}{k}$ equal probabilities $P(x_1^n) = p^k q^{n-k}$, k number of 1's. Define

$$A_k = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{k}, \quad A_{-1} = 0.$$

Starting from A_{k-1} the next $\binom{n}{k}$ probabilities $P(x_1^n)$ are the same. The average code length is

$$L_n = \sum_{k=0}^n p^k q^{n-k} \sum_{j=A_{k-1}+1}^{A_k} \lfloor \log_2(j) \rfloor = \sum_{k=0}^n p^k q^{n-k} \sum_{i=1}^{\binom{n}{k}} \lfloor \log_2(A_{k-1} + i) \rfloor.$$

In 2005 we proved the following asymptotic results:

- $p = \frac{1}{2}$

$$L_n = n - 2 + 2^{-n}(n + 2).$$

- $p < \frac{1}{2}$

$$L_n = nh(p) - \frac{1}{2} \log n + O(1).$$

Binary Case: More Precise Main Result

Theorem 3 (W.S., 2005). For a *binary memoryless source*, let $p < \frac{1}{2}$. Then

$$L_n = nH(p) - \frac{1}{2} \log_2 n - \frac{3 + \ln(2)}{2 \ln(2)} + \log_2 \frac{1-p}{1-2p} \frac{1}{\sqrt{2\pi p(1-p)}} \\ + \frac{p}{1-2p} \log_2 \left(\frac{2(1-p)}{p} \right) + F(n) + o(1)$$

where $H(p) = -p \log_2 p - (1-p) \log_2(1-p)$, and

- $F(n) = 0$ if $\log_2 \frac{1-p}{p}$ is *irrational*;
- $F(n)$ is an *oscillating function* if $\log_2 \frac{1-p}{p} = N/M$ is *rational*,

$$F(n) = \frac{1-p}{1-2p} H_M(n\beta)[x] - \frac{p}{1-2p} H_M(n\beta-\alpha)[-x] - \frac{2(1-3p)}{1-2p} H_M(n\beta)[2^{-x}] + \frac{p}{1-2p} H_M(n\beta-\alpha)[2^x]$$

where

$$H_M(\mathbf{y})[f] := \frac{1}{M\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \left(\left\langle M \left(\mathbf{y} - \log_2 \left(\frac{1-2p}{1-p} \sqrt{2\pi p q n} \right) - \frac{x^2}{2 \ln 2} \right) \right\rangle - \int_0^1 f(t) dt \right) dx$$

for some Riemann function f and $\beta = -\log_2(1-p)$.

Some Oscillations

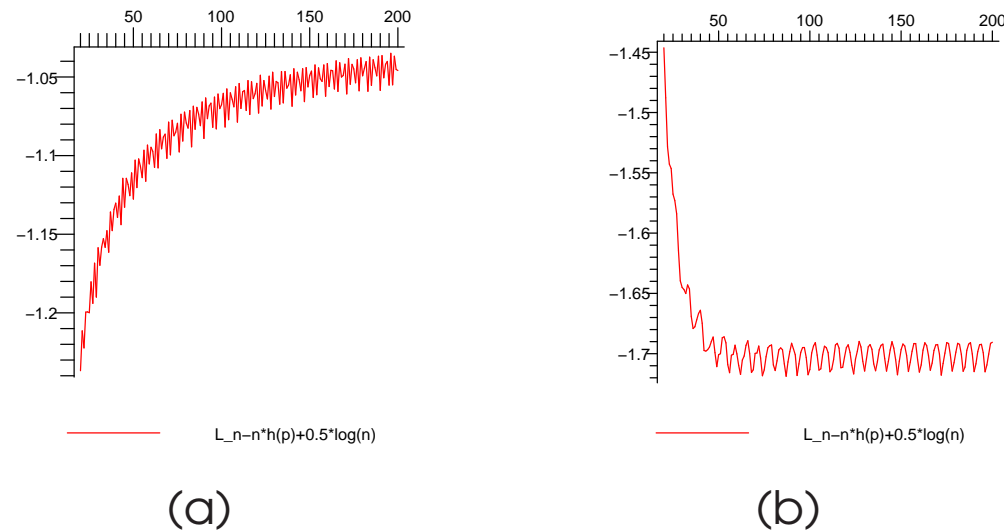


Figure 2: Plots of $L_n - n h(p) + 0.5 \log(n)$ (y-axis) versus n (x-axis) for: (a) **irrational** $\alpha = \log_2(1 - p)/p$ with $p = 1/\pi$; (b) **rational** $\alpha = \log_2(1 - p)/p$ with $p = 1/9$.

Sketch of Proof

1. Using the following identity to handle **floor functions** (partial summation; cf. Knuth)

$$\sum_{j=1}^N a_j = N a_n - \sum_{j=1}^{N-1} (a_{j+1} - a_j)$$

we can reduce L_n to the sums of the following form

$$\begin{aligned} S_n &= \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \lfloor \log_2 A_k \rfloor \\ &= \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \log_2 A_k - \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle \\ &= a_n + b_n \end{aligned}$$

where

$$\begin{aligned} a_n &= \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \log_2 A_k, \\ b_n &= \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle. \end{aligned}$$

Asymptotics of A_n : Saddle point Method

Lemma 5. For large n and $p < 1/2$

$$A_{np} = \frac{1-p}{1-2p} \frac{1}{\sqrt{2\pi np(1-p)}} 2^{nH(p)} \left(1 + O(n^{-1/2})\right).$$

More precisely, for an $\varepsilon > 0$ and $k = np + \Theta(n^{1/2+\varepsilon})$ we have (Exercise)

$$A_k = \frac{1-p}{1-2p} \frac{1}{\sqrt{2\pi np(1-p)}} \left(\frac{1-p}{p}\right)^k \frac{1}{(1-p)^n} \exp\left(-\frac{(k-np)^2}{2p(1-p)n}\right) \left(1 + O(n^{-\delta})\right).$$

Proof. Notice that $A_n(z) = \sum_{k=0}^n A_k z^k = \frac{(1+z)^n - 2^n z^{n+1}}{1-z}$. Apply the saddle point method to the Cauchy formula $A_k = [z^n] A_n(z)$.

$$A_k = \frac{1}{2\pi i} \oint \frac{(1+z)^n - 2^n z^{n+1}}{1-z} \frac{dz}{z^{k+1}} = \frac{1}{2\pi i} \oint \frac{1}{1-z} 2^{n \log(1+z) - (k+1) \log z} dz.$$

The saddle point $z_0 = (k+1)/(n-k+1) = p/(1-p)$ and $H''(z_0) = q^3/p$.

$$A_k = \frac{1}{1-z_0} \frac{1}{\sqrt{2\pi n H''(z_0)}} 2^{nH(z_0)} \left(1 + O(n^{-1/2})\right).$$

Heuristic: Note that $\binom{n}{k-i} = (p/q)^i \binom{n}{k}$.

Returning to b_n

3. We also need asymptotics of

$$b_n = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle.$$

From previous lemma we conclude that

$$\log A_k = \alpha k + n\beta - \log_2 \omega \sqrt{n} - \frac{(k - np)^2}{2pqn \ln 2} + O(n^{-\delta})$$

for some $\omega > 0$ and $\alpha = \log p / (1 - p)$.

Thus we need asymptotics of the following sum

$$\sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \left\langle \alpha k + n\beta - \log_2 \omega \sqrt{n} - \frac{(k - np)^2}{2pqn \ln 2} \right\rangle.$$

We must now resort to theory of **Bernoulli sequences modulo 1**.

Final Lemma

Lemma 6 (Drmota, Hwang, W.S., 2004). Let $0 < p < 1$ be a fixed real number and $f : [0, 1] \rightarrow \mathbf{R}$ be a Riemann integrable function.

(i) If α is *irrational*, then

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} f \left(\left\langle k\alpha + y - (k - np)^2 / (2pqn \ln 2) \right\rangle \right) = \int_0^1 f(t) dt,$$

where the convergence is uniform for all shifts $y \in \mathbf{R}$.

(ii) If $\alpha = \frac{N}{M}$ (*rational*) ($\gcd(N, M) = 1$), then uniformly $y \in \mathbf{R}$

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} f \left(\left\langle k\alpha + y - (k - np)^2 / (2pqn \ln 2) \right\rangle \right) = \int_0^1 f(t) dt + H_M(y)$$

where

$$H_M(y) := \frac{1}{M} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \left(\left\langle M \left(y - \frac{x^2}{2 \ln 2} \right) \right\rangle - \int_0^1 f(t) dt \right) dx$$

is a *periodic function* with period $\frac{1}{M}$.

General Memoryless Source: Main Results

We consider m -ary alphabet \mathcal{A} with probabilities of symbols:

$$p_1 \leq p_2 \leq \cdots \leq p_{m-1} \leq p_m.$$

Entropy: $H(X) = -\sum_i p_i \log p_i$. Define also $B_i = \log p_m/p_i$.

Our main results are:

Theorem 4. For a *memoryless source* with finite alphabet \mathcal{A} , the *minimum expected length* of a lossless binary encoding is

$$L_n = \lfloor n \log_2 |\mathcal{A}| \rfloor + o(1).$$

if the source is **equiprobable**, and

$$L_n = nH(X) - \frac{1}{2} \log_2 n + O(1)$$

if the source is **not equiprobable**.

Sketch of the Proof

1. Type: Type $\mathcal{T}_{n,m}$

$$\mathcal{T}_{n,m} = \{(k_1, \dots, k_m) \in \mathbb{N}^m, k_1 + \dots + k_m = n\}.$$

such that k_i is the number of symbols i in a sequence and probabilities are the same

$$p^{\mathbf{k}} = p_1^{k_1} \cdots p_m^{k_m}, \quad \mathbf{k} \in \mathcal{T}_{n,m}.$$

2. Order among types:

$$\mathbf{l} \preceq \mathbf{k} \quad \text{iff} \quad p^{\mathbf{l}} \geq p^{\mathbf{k}},$$

sort from the **smallest index** to the **largest**; or equivalently

$$l_1 B_1 + \dots + l_{m-1} B_{m-1} \leq k_1 B_1 + \dots + k_{m-1} B_{m-1}$$

where $B_i = \log p_m / p_i$.

3. There are

$$\binom{n}{\mathbf{k}} = \binom{n}{k_1, \dots, k_m}$$

sequences of **type** \mathbf{k} . Define

$$A_{\mathbf{k}} := \sum_{\mathbf{l} \preceq \mathbf{k}} \binom{n}{\mathbf{l}}.$$

Starting from position $A_{\mathbf{k}}$ the next $\binom{n}{\mathbf{k}+1}$ sequences have the **same probability** $p^{\mathbf{k}+1}$, where $\mathbf{k}+1$ is the “next” type.

Sketch of Proof

4. The average code length is

$$\begin{aligned}
 L_n &= \sum_{\mathbf{k} \in \mathcal{T}_{n,m}} p^{\mathbf{k}} \sum_{i=A_{\mathbf{k}-1}+1}^{A_{\mathbf{k}}} \lfloor \log i \rfloor = \sum_{\mathbf{k} \in \mathcal{T}_{n,m}} p^{\mathbf{k}} \sum_{i=1}^{\binom{n}{\mathbf{k}}} \lfloor \log(A_{\mathbf{k}} - i) \rfloor \\
 &= \sum_{\mathbf{k} \in \mathcal{T}_{n,m}} \binom{n}{\mathbf{k}} p^{\mathbf{k}} \log A_{\mathbf{k}} + O(1) \stackrel{\text{binomial sum}}{=} \log A_{n\mathbf{p}} + O(1)
 \end{aligned}$$

5. We need to estimate

$$A_{n\mathbf{p}} = \sum_{p^{\mathbf{l}} \geq p^{n\mathbf{p}}} \binom{n}{\mathbf{l}}$$

For $l_i = np_i + x_i$

$$A_{n\mathbf{p}} = \sum_{B_1 x_1 + \dots + B_{m-1} x_{m-1} \leq 0} \binom{n}{n\mathbf{p} + \mathbf{x}}.$$

Sketch of the Proof

6. By Stirling

$$\binom{n}{n\mathbf{p} + \mathbf{x}} \sim (1 + O(1/\sqrt{n})) C \frac{2^{nH(X)}}{n^{(m-1)/2}} \exp(B_1 x_1 + \dots + B_{m-1} x_{m-1}) \\ \cdot \exp\left(-\frac{1}{2n} \mathbf{x}^T \Sigma^{-1} \mathbf{x}\right)$$

where Σ is an invertible covariance matrix $(m-1) \times (m-1)$.

7. Observe that ($\mathbf{b} = (B_1, \dots, B_{m-1})$)

$$A_{n\mathbf{p}} = \frac{C 2^{nH(X)}}{n^{(m-1)/2}} \left(\sum_{\mathbf{b}^T \mathbf{x} = 0} \exp\left(-\frac{1}{2n} \mathbf{x}^T \Sigma^{-1} \mathbf{x}\right) + \sum_{\mathbf{b}^T \mathbf{x} < 0} \exp\left(\mathbf{b}^T \mathbf{x} - \frac{1}{2n} \mathbf{x}^T \Sigma^{-1} \mathbf{x}\right) \right).$$

which leads to

$$\log A_{n\mathbf{p}} = \log \left(C \frac{2^{nH(X)}}{n^{(m-1)/2}} n^{(m-2)/2} + O\left(\frac{2^{H(X)}}{n^{(m-1)/2}}\right) \right) = nH(X) - \frac{1}{2} \log n + O(1)$$

since

$$\int_{\mathcal{D}^{m-2}} \exp\left(-\frac{1}{2n} \mathbf{x}^T \Sigma^{-1} \mathbf{x}\right) = C n^{(m-2)/2}.$$

Example for $m = 3$

Assume now $m = 3$. Then

$$\begin{aligned}
 A_{np} &= \sum_{B_1 x + B_2 y \leq 0} \binom{n}{np_1 + x, np_2 + y} \\
 &\sim \frac{2^{nH(X)}}{n\sqrt{2\pi p_1 p_2 p_3}} \sum_{B_1 x + B_2 y = 0} \exp\left(-\frac{x^2}{2np_1} - \frac{y^2}{2np_2} - \frac{(x+y)^2}{2np_3}\right) \\
 &= O(\sqrt{n}) \frac{2^{nH(X)}}{n} = C \frac{2^{nH(X)}}{\sqrt{n}},
 \end{aligned}$$

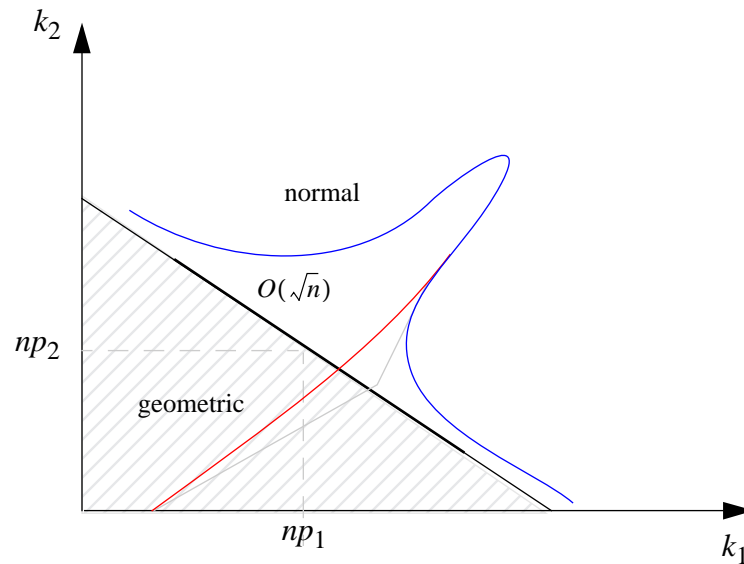


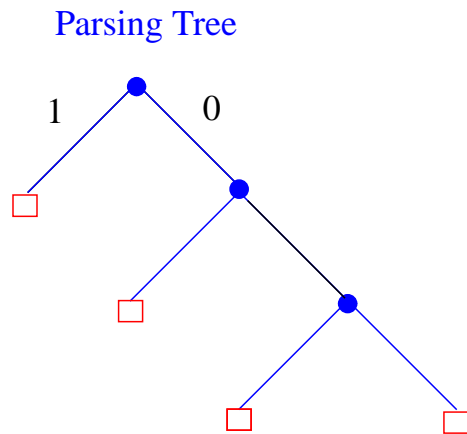
Figure 3: Illustration for $m = 3$

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
 - Shannon and Huffman Coding
 - Non-Prefix Codes
 - Tunstall Code
5. Minimax Redundancy: Universal (unknown) Sources

Variable-to-Fixed Codes

A VF coder consists of a **parser** and a **dictionary**.



Dictionary

1
01
001
000

1. A **variable-to-fixed** length encoder **partitions** the source string into a concatenation of **variable-length phrases**.

2. Each **phrase** belongs to a given **dictionary** \mathcal{D} of source strings.

3. A **dictionary** can be represented by a **complete parsing tree** \mathcal{T} .

The **dictionary** entries $d \in \mathcal{D}$ correspond to the **leaves** of the parsing tree.

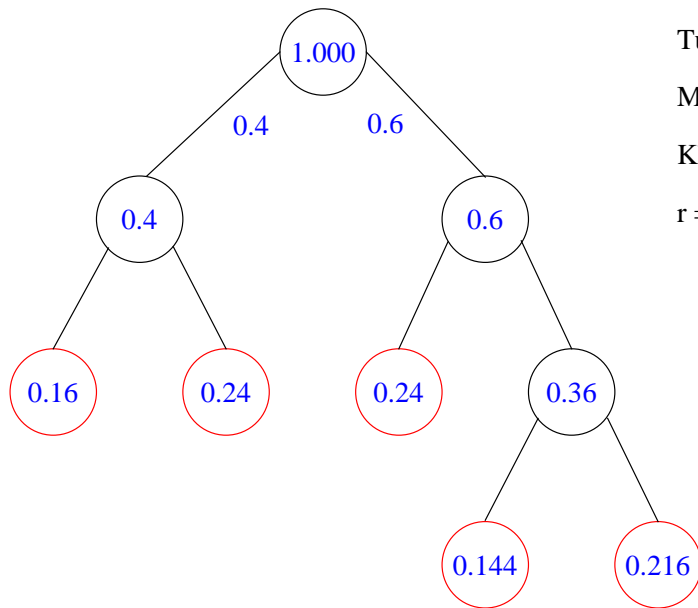
4. The **encoder** represents **phrases** by the **fixed length binary codewords**. i.e., a dictionary \mathcal{D} of M entries requires $\lceil \log_2 M \rceil$ bits to represent **entries**.

Average Redundancy Rate:

$$\bar{r} = \lim_{n \rightarrow \infty} \frac{\sum_{|x|=n} P_{\mathcal{S}}(x) (L(x) + \log P_{\mathcal{S}}(x))}{n} = \frac{\log M}{\mathbf{E}[D]} - h$$

where h is the **entropy rate** of the source.

Tunstall and Khodak Codes



$$p = 0.6 \quad q = 0.4$$

Tunstall's construction

$$M = 5$$

Khodak's construction

$$r = 0.25$$

Tunstall Code:

1. Start with a **root** and **leaves**.
2. In the J 's iteration select a leaf with the **highest probability** and **grow children** out it.
3. At J th step, the parsing tree has J **internal nodes** and $M = J + 1$ **leaves** corresponding to **dictionary entries**.

Khodak Construction:

1. Pick a real number r and grow a **complete parsing tree** satisfying

$$\min\{p, 1 - p\} \cdot r \leq P(d) < r, \quad d \in \mathcal{D}.$$

2. The resulting **parsing tree** is **exactly the same** as the **Tunstall tree**.
3. If y is a **proper prefix** of entries of \mathcal{D}_r , i.e., y is an **internal node** of \mathcal{T}_r , then

$$P(y) \geq r.$$

Phrase Length

We study the phrase length $D = |d|$, i.e., **path length** in the **parsing tree**.

Moment Generating Functions: Define

$$D(\mathbf{r}, z) := \mathbf{E}[z^D] = \sum_{d \in \mathcal{D}_r} P(d) z^{|d|}.$$

and its corresponding **internal nodes** generating function

$$S(\mathbf{r}, z) = \sum_{y: P(y) \geq r} P(y) z^{|y|}.$$

Simple Fact on Trees: Let $\tilde{\mathcal{D}}$ be a dictionary (**leaves of \mathcal{T}**) and $\tilde{\mathcal{Y}}$ be the collection of **proper prefixes** of dictionary entries (**internal nodes of \mathcal{T}**).

$$\sum_{d \in \tilde{\mathcal{D}}} P(d) \frac{z^{|d|} - 1}{z - 1} = \sum_{y \in \tilde{\mathcal{Y}}} P(y) z^{|y|}, \quad \text{Exercise.}$$

Thus

$$D(\mathbf{r}, z) = 1 + (z - 1)S(\mathbf{r}, z),$$

and

$$\mathbf{E}[D] = \tilde{S}(\mathbf{v}, 1) = \sum_{y \in \tilde{\mathcal{Y}}} P(y), \quad \mathbf{E}[D(D - 1)] = \tilde{S}'(\mathbf{v}, 1) = 2 \sum_{y \in \tilde{\mathcal{Y}}} P(y) |y|.$$

Recurrences

Define $v = 1/r$, z complex, and $\tilde{S}(v, z) = S(v^{-1}, z)$.

Let

$$A(v) = \sum_{y: P(y) \geq 1/v} 1$$

be the # of strings of probab. $\geq v^{-1}$ or the # of internal nodes. In fact:

$$M_r = A(v) + 1.$$

We have

$$A(v) = \begin{cases} 0 & v < 1, \\ 1 + A(vp) + A(vq) & v \geq 1 \end{cases}$$

and

$$\tilde{S}(v, z) = \begin{cases} 0 & v < 1, \\ 1 + zp\tilde{S}(vp, z) + zq\tilde{S}(vq, z) & v \geq 1, \end{cases}$$

since every binary string either is:

- – empty string,
- – string starting with the first symbol
- – string starting with second symbol.

Mellin Transform

The Mellin transform $F^*(s)$ of a function $F(v)$ is

$$F^*(s) = \int_0^{\infty} F(v)v^{s-1}dv.$$

From the recurrence on $S(v, z)$ we find

$$\tilde{D}^*(s, z) = \frac{1-z}{s(1-zp^{1-s}-zq^{1-s})} - \frac{1}{s}, \quad \Re(s) < s_0(z),$$

where $s_0(z)$ denotes the real solution of: $zp^{1-s} + zq^{1-s} = 1$.

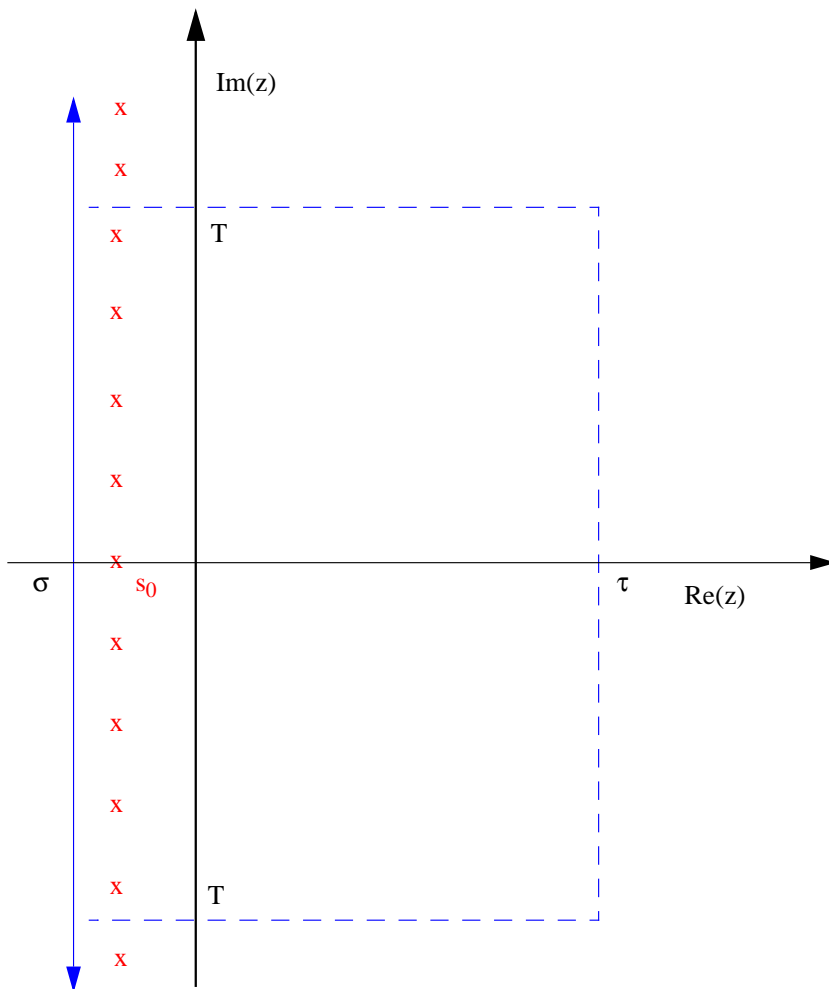
To find the asymptotics of $\tilde{D}(v, z)$ as $v \rightarrow \infty$ we compute the inverse transform of $\tilde{D}^*(s, z)$:

$$\tilde{D}(v, z) = \frac{1}{2\pi i} \lim_{T \rightarrow \infty} \int_{\sigma-iT}^{\sigma+iT} \tilde{D}^*(s, z)v^{-s} ds,$$

where $\sigma < s_0(z)$.

To determine the polar singularities of the meromorphic continuation of $\tilde{D}^*(s, z)$, we have to analyze the set

$$\mathbb{Z}(z) = \{s \in \mathbb{C} : zp^{1-s} + zq^{1-s} = 1\}.$$



From

$$\tilde{D}(v, s)$$

$$F_T(v, s)$$

$$= - \sum$$

$$+ \frac{1}{2\pi i} \int$$

$$= - \sum$$

$$+ \frac{1}{2\pi i} \int$$

provided that the series of residues converges
the last integral exists. **But they don't!**

Tauberian Rescue

Therefore, we analyze (as in [analytic number theory](#); cf. also [Vallee](#))

$$\tilde{D}_1(v, z) = \int_0^v \tilde{D}(w, z) dw.$$

whose Mellin transform is

$$\tilde{D}_1^*(s, z) = \frac{-\tilde{D}^*(s+1, z)}{s} = O(1/s^2).$$

Lemma 7 (Tauberian). Let $f(v, \lambda)$ be a *non-negative increasing* function such that

$$F(v, \lambda) = \int_0^v f(w, \lambda) dw$$

and has the *asymptotic expansion*

$$F(v, \lambda) = \frac{v^{\lambda+1}}{\lambda+1} (1 + \lambda \cdot o(1))$$

as $v \rightarrow \infty$ and *uniformly* in λ . Then as $v \rightarrow \infty$ *uniformly* in λ

$$f(v, \lambda) = v^\lambda (1 + |\lambda|^{\frac{1}{2}} \cdot o(1)).$$

Main Results

Theorem 5 (Central Limit Theorem). For large M_r

$$\frac{D_r - \frac{1}{H} \ln M_r}{\sqrt{\left(\frac{H_2}{H^3} - \frac{1}{H}\right) \ln M_r}} \rightarrow N(0, 1) \text{ standard normal distribution}$$

where H is *natural entropy* and $H_2 = p \ln^2 p + q \ln^2 q$.

If $\ln q / \lg p$ is *irrational*, then

$$\begin{aligned} M_r &= A(v) + 1 = \frac{v}{H} + o(v) \\ \mathbf{E}[D_r] &= \frac{\ln M_r}{H} + \frac{\ln H}{H} + \frac{H_2}{2H^2} + o(1); \end{aligned}$$

if $\ln q / \lg p$ is *rational*, then

$$\begin{aligned} M_r &= \frac{Q_1(\log v)}{H} v + O(v^{1-\eta}), \quad Q_1(x) = \frac{L}{1 - e^{-L}} e^{-L \langle \frac{x}{L} \rangle}, \\ \mathbf{E}[D_r] &= \frac{\ln M_r}{H} + \frac{\ln H}{H} + \frac{H_2}{2H^2} + \frac{-\ln L + \ln(1 - e^{-L}) + \frac{L}{2}}{H} + O(M_r^{-\eta}), \end{aligned}$$

L largest real number s.t. $\ln(1/p)$ and $\ln(1/q)$ are *integer multiples* of L .

Redundancy Rate

The average **redundancy rate** of Tunstall/Khodak's VF code is defined as

$$\bar{r}_{M_r} = \frac{\ln M_r}{\mathbf{E}[D]} - h.$$

Case 1: $\ln p / \ln q$ is **irrational**:

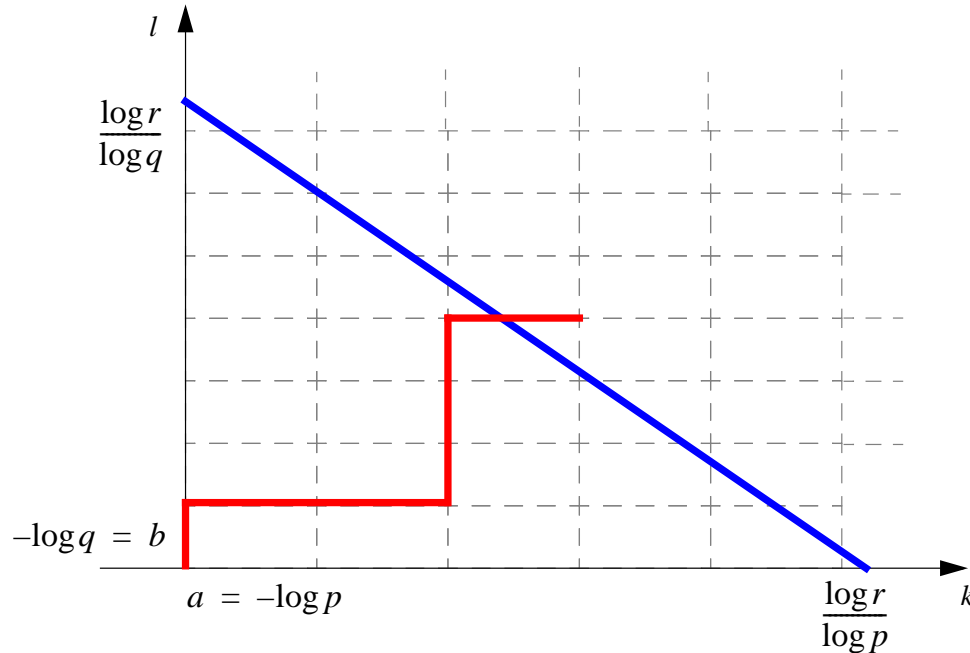
$$\bar{r}_{M_r} = \frac{H}{\ln M_r} \left(-\frac{H_2}{2H} - \ln H \right) + o\left(\frac{1}{\ln M_r}\right).$$

Case 2: $\ln p / \ln q$ is **rational**:

$$\bar{r}_{M_r} = \frac{H}{\ln M_r} \left(-\frac{H_2}{2H} - \ln H + \ln L - \ln(e^L - 1) + \frac{L}{2} \right) + O\left(M_r^{-\eta}\right),$$

for some $\eta > 0$, where $L > 0$ is the **largest real number** for which $\ln(1/p)$ and $\ln(1/q)$ are **integer multiples of L** . (**No oscillation!**)

Random Walk



Consider a **random walk** that corresponds to a **path** in the **associated parsing tree**.

For Khodak's code we studied

$$A(v) = \sum_{y: P(y) \geq 1/v} f(v)$$

for some function $f(v)$.

But $P(y) = p^k q^l$ ($k, l \geq 0$), and set $v = 2^V$ so that

$$\log P(v) = k \lg(1/p) + l \lg(1/q) \leq V.$$

This corresponds to a **random walk** in the **first quadrant** with the **linear boundary condition**

$$ax + by = V$$

where $a = \log(1/p)$ and $b = \log(1/q)$.

The **phrase length coincides** with the **exit time** of such a random walk.

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
5. **Minimax Redundancy: Universal (unknown) Sources**
 - **Universal Memoryless Sources**
 - Universal Markov Sources
 - Universal Renewal Sources

Minimax Redundancy

Unknown Source P

In practice, one can only hope to have **some knowledge** about a family of sources \mathcal{S} that generates real data.

Following Davisson we define the **average minimax redundancy** $\bar{R}_n(\mathcal{S})$ and **the worst case (maximal) minimax redundancy** $R_n^*(\mathcal{S})$ for a family of sources \mathcal{S} as

$$\begin{aligned}\bar{R}_n(\mathcal{S}) &= \min_{C_n} \sup_{P \in \mathcal{S}} \mathbf{E}[L(C_n, x_1^n) + \lg P(x_1^n)] \\ R_n^*(\mathcal{S}) &= \min_{C_n} \sup_{P \in \mathcal{S}} \max_{x_1^n} [L(C_n, x_1^n) + \lg P(x_1^n)].\end{aligned}$$

In the minimax scenario we look for the **best code** for the **the worst source**.

Source Coding Goal:

Find data compression algorithms that **match optimal redundancy rates** either on **average** or for **individual sequences**.

Maximal Minimax Redundancy

We consider the following **classes of sources** \mathcal{S} :

- **Memoryless sources** \mathcal{M}_0 over an m -ary (finite) alphabet, that is,

$$P(x_1^n) = p_1^{k_1} \cdots p_m^{k_m}$$

with $k_1 + \cdots + k_m = n$, where p_i are **unknown**!

- **Markov sources** \mathcal{M}_r over a **binary** alphabet of order r . Observe that for $r = 1$

$$P(x_1^n) = p_{x_1} p_{00}^{k_{00}} p_{01}^{k_{01}} p_{10}^{k_{10}} p_{11}^{k_{11}},$$

where k_{ij} is such that $(x_k, x_{k+1}) = (i, j) \in \{0, 1\}^2$ and

$$k_{00} + k_{01} + k_{10} + k_{11} = n - 1,$$

and that $k_{01} = k_{10}$ if $x_1 = x_n$ and $k_{01} = k_{10} \pm 1$ if $x_1 \neq x_n$.

- **Renewal Sources** \mathcal{R}_0 where an **1** is introduced after **a run of 0s** distributed according to some distribution.

Improved Shtarkov Bounds

For the **maximal minimax** redundancy define

$$Q^*(x_1^n) := \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{\sum_{y_1^n \in \mathcal{A}^n} \sup_{P \in \mathcal{S}} P(y_1^n)}.$$

the **maximum likelihood distribution**. Observe that

$$\begin{aligned} R_n^*(\mathcal{S}) &= \min_{C_n \in \mathcal{C}} \sup_{P \in \mathcal{S}} \max_{x_1^n} (L(C_n, x_1^n) + \lg P(x_1^n)) \\ &= \min_{C_n \in \mathcal{C}} \max_{x_1^n} \left(L(C_n, x_1^n) + \sup_{P \in \mathcal{S}} \lg P(x_1^n) \right) \\ &= \min_{C_n \in \mathcal{C}} \max_{x_1^n} [L(C_n, x_1^n) + \lg Q^*(x_1^n) + \lg \sum_{y_1^n \in \mathcal{A}^n} \sup_{P \in \mathcal{S}} P(y_1^n)] \\ &= R_n^{GS}(Q^*) + \lg \sum_{y_1^n \in \mathcal{A}^n} \sup_{P \in \mathcal{S}} P(y_1^n) \end{aligned}$$

where $R_n^{GS}(Q^*)$ is the **maximal redundancy** of a **generalized Shannon** code built for the (known) distribution Q^* . We also write

$$D_n(\mathcal{S}) = \lg \left(\sum_{x_1^n \in \mathcal{A}^n} \sup_{P \in \mathcal{S}} P(x_1^n) \right) := \lg d_n(\mathcal{S}).$$

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
5. **Minimax Redundancy: Universal (unknown) Sources**
 - **Universal Memoryless Sources**
 - Universal Markov Sources
 - Universal Renewal Sources

Maximal Minimax for Memoryless Sources

We first consider the **maximal minimax redundancy** $R_n^*(\mathcal{M}_0)$ for a class of **memoryless sources** over a **finite m -ary alphabet**. Observe that

$$\begin{aligned}d_n(\mathcal{M}_0) &= \sum_{x_1^n} \sup_{p_1, \dots, p_m} p_1^{k_1} \cdots p_m^{k_m} \\ &= \sum_{k_1 + \dots + k_m = n} \binom{n}{k_1, \dots, k_m} \sup_{p_1, \dots, p_m} p_1^{k_1} \cdots p_m^{k_m} \\ &= \sum_{k_1 + \dots + k_m = n} \binom{n}{k_1, \dots, k_m} \left(\frac{k_1}{n}\right)^{k_1} \cdots \left(\frac{k_m}{n}\right)^{k_m}.\end{aligned}$$

The **summation set** is

$$I(k_1, \dots, k_m) = \{(k_1, \dots, k_m) : k_1 + \dots + k_m = n\}.$$

The number $N_{\mathbf{k}}$ of **types** $\mathbf{k} = (k_1, \dots, k_m)$ is

$$N_{\mathbf{k}} = \binom{n}{k_1, \dots, k_m}$$

The (unnormalized) **likelihood distribution** is

$$\sup_{p_1, \dots, p_m} p_1^{k_1} \cdots p_m^{k_m} = \left(\frac{k_1}{n}\right)^{k_1} \cdots \left(\frac{k_m}{n}\right)^{k_m}$$

Generating Function for $d_n(\mathcal{M}_0)$

We write

$$d_n(\mathcal{M}_0) = \frac{n!}{n^n} \sum_{k_1 + \dots + k_m = n} \frac{k_1^{k_1}}{k_1!} \dots \frac{k_m^{k_m}}{k_m!}$$

Let us introduce a **tree-generating function**

$$B(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} z^k = \frac{1}{1 - T(z)},$$

where $T(z)$ satisfies $T(z) = ze^{T(z)}$ ($= -W(-z)$, **Lambert's** W -function) and also

$$T(z) = \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} z^k$$

enumerates all **rooted labeled trees**. Let now

$$D_m(z) = \sum_{n=0}^{\infty} \frac{n^n}{n!} d_n(\mathcal{M}_0) z^n.$$

Then by the **convolution formula**

$$D_m(z) = [B(z)]^m.$$

Asymptotics

The function $B(z)$ has an **algebraic singularity** at $z = e^{-1}$ and

$$B(z) = \frac{1}{\sqrt{2(1-ez)}} + \frac{1}{3} + O(\sqrt{(1-ez)}).$$

The **singularity analysis** yields

$$[z^n] \left(\frac{1}{\sqrt{1-ez}} \right) = \frac{e^n}{\sqrt{\pi n}} \left(1 - \frac{1}{8n} + O(1/n^2) \right),$$

$$[z^n] (\sqrt{1-ez}) = -\frac{e^n}{\sqrt{\pi n^3}} \left(\frac{1}{2} + \frac{3}{16n} \right)$$

$$[z^n] \left(\frac{1}{1-ez} \right) = e^n,$$

$$\frac{n!}{n^n} = e^{-n} \sqrt{2\pi n} \left(1 + \frac{1}{12n} + O(1/n^2) \right).$$

which leads to (cf. Clarke & Barron, 1990, W.S., 1998)

$$\begin{aligned} d_n(\mathcal{M}_0) &= \frac{m-1}{2} \log \left(\frac{n}{2} \right) + \log \left(\frac{\sqrt{\pi}}{\Gamma(\frac{m}{2})} \right) + \frac{\Gamma(\frac{m}{2})m}{3\Gamma(\frac{m}{2} - \frac{1}{2})} \cdot \frac{\sqrt{2}}{\sqrt{n}} \\ &+ \left(\frac{3 + m(m-2)(2m+1)}{36} - \frac{\Gamma^2(\frac{m}{2})m^2}{9\Gamma^2(\frac{m}{2} - \frac{1}{2})} \right) \cdot \frac{1}{n} + \dots \end{aligned}$$

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
5. **Minimax Redundancy: Universal (unknown) Sources**
 - Universal Memoryless Sources
 - **Universal Markov Sources**
 - Universal Renewal Sources

Maximal Minimax for Markov Sources

- (i) \mathcal{M}_1 is a Markov source of order $r = 1$,
- (ii) the transition matrix $P = \{p_{ij}\}_{i,j=1}^m$
- (iii) **circular** sequences (the first symbols follows the last).

$$\begin{aligned}
 d_n(\mathcal{M}_1) &= \sum_{x_1^n} \sup_P p_{11}^{k_{11}} \cdots p_{mm}^{k_{mm}} \\
 &= \sum_{\mathbf{k} \in \mathcal{F}_n} M_{\mathbf{k}} \left(\frac{k_{11}}{k_1} \right)^{k_{11}} \cdots \left(\frac{k_{mm}}{k_m} \right)^{k_{mm}},
 \end{aligned}$$

k_{ij} is the number of pairs $ij \in \mathcal{A}^2$ in x_1^n , $k_i = \sum_{j=1}^m k_{ij}$,
 $\mathbf{k} = \{k_{ij}\}_{i,j=1}^m$ is an integer matrix such that

$$\mathcal{F}_n : \sum_{i,j=1}^m k_{ij} = n, \quad \text{and} \quad \sum_{j=1}^m k_{ij} = \sum_{j=0}^{m-1} k_{ji},$$

Matrix \mathbf{k} satisfying the above conditions is called the **frequency matrix** or **Markov type**.

$M_{\mathbf{k}}$ represents the numbers of strings x_1^n of **type** \mathbf{k} .

We come back to Markov types soon.

Main Technical Tool

Let $g_{\mathbf{k}}$ be a sequence of scalars indexed by matrices \mathbf{k} and

$$g(\mathbf{z}) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}$$

be its regular generating function, and

$$\mathcal{F}g(\mathbf{z}) = \sum_{\mathbf{k} \in \mathcal{F}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} = \sum_{n \geq 0} \sum_{\mathbf{k} \in \mathcal{F}_n} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}$$

the \mathcal{F} -generating function of $g_{\mathbf{k}}$ for which $\mathbf{k} \in \mathcal{F}$.

Lemma 8. Let $g(\mathbf{z}) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}$. Then

$$\mathcal{F}g(\mathbf{z}) := \sum_{n \geq 0} \sum_{\mathbf{k} \in \mathcal{F}_n} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} = \left(\frac{1}{2^J \pi} \right)^m \oint \frac{dx_1}{x_1} \cdots \oint \frac{dx_m}{x_m} g\left(\left[z_{ij} \frac{x_j}{x_i}\right]\right)$$

with the ij -th coefficient of $\left[z_{ij} \frac{x_j}{x_i}\right]$ is $z_{ij} \frac{x_j}{x_i}$.

Proof. It suffices to observe

$$g\left(\left[z_{ij} \frac{x_j}{x_i}\right]\right) = \sum_{\mathbf{k}} g_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} \prod_{i=1}^m x_i^{\sum_j k_{ij} - \sum_j k_{ji}}$$

Thus $\mathcal{F}g(\mathbf{z})$ is the coefficient of $g\left(\left[z_{ij} \frac{x_j}{x_i}\right]\right)$ at $x_1^0 x_2^0 \cdots x_m^0$.

Main Results

Theorem 6. Let \mathcal{M}_1 be a *Markov source* over an m -ry alphabet. Then

$$d_n(\mathcal{M}_1) = \left(\frac{n}{2\pi}\right)^{m(m-1)/2} A_m \times \left(1 + O\left(\frac{1}{n}\right)\right)$$

with

$$A_m = \int_{\mathcal{K}(1)} m F_m(y_{ij}) \prod_i \frac{\sqrt{\sum_j y_{ij}}}{\prod_j \sqrt{y_{ij}}} d[y_{ij}]$$

where $\mathcal{K}(1) = \{y_{ij} : \sum_{ij} y_{ij} = 1\}$ and $F_m(\cdot)$ is a polynomial expression of degree $m - 1$.

In particular, for $m = 2$ $A_2 = 16 \times \text{Catalan}$ where Catalan is Catalan's constant $\sum_i \frac{(-1)^i}{(2i+1)^2} \approx 0.915965594$.

Theorem 7. Let \mathcal{M}_r be a *Markov source* of order r . Then

$$d_n(\mathcal{M}_r) = \left(\frac{n}{2\pi}\right)^{m^r(m-1)/2} A_m^r \times \left(1 + O\left(\frac{1}{n}\right)\right)$$

where A_m^r is a constant defined in a similar fashion as A_m above.

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
5. **Minimax Redundancy: Universal (unknown) Sources**
 - Universal Memoryless Sources
 - Universal Markov Sources
 - **Universal Renewal Sources**

Renewal Sources

The **renewal process** defined as follows:

- Let $T_1, T_2 \dots$ be a sequence of i.i.d. positive-valued random variables with distribution $Q(j) = \Pr\{T_i = j\}$.
- The process $T_0, T_0 + T_1, T_0 + T_1 + T_2, \dots$ is called the **renewal process**.
- With a renewal process we associate a **binary renewal sequence** in which **the positions of the 1's** are at the renewal epochs (**runs of zeros**) $T_0, T_0 + T_1, \dots$
- We start with $x_0 = 1$.

Csiszár and Shields (1996) proved that $R_n(\mathcal{R}_0) = \Theta(\sqrt{n})$.

We prove the following result.

Theorem 8 (Flajolet and WS, 1998). *Consider the class of renewal processes as defined above. Then*

$$R_n^*(\mathcal{R}_0) = \frac{2}{\log 2} \sqrt{cn} + O(\log n).$$

where $c = \frac{\pi^2}{6} - 1 \approx 0.645$.

Maximal Minimax Redundancy

For a sequence

$$x_0^n = 10^{\alpha_1} 10^{\alpha_2} 1 \dots 10^{\alpha_n} 1 \underbrace{0 \dots 0}_{k^*}$$

k_m is the number of i such that $\alpha_i = m$. Then

$$P(x_1^n) = Q^{k_0}(0) Q^{k_1}(1) \dots Q^{k_{n-1}}(n-1) \Pr\{T_1 > k^*\}.$$

It can be proved that

$$r_{n+1} - 1 \leq d_n(\mathcal{R}_0) \leq \sum_{m=0}^n r_m$$

where

$$r_n = \sum_{k=0}^n r_{n,k}$$

$$r_{n,k} = \sum_{\mathcal{P}(n,k)} \binom{k}{k_0 \dots k_{n-1}} \left(\frac{k_0}{k}\right)^{k_0} \left(\frac{k_1}{k}\right)^{k_1} \dots \left(\frac{k_{n-1}}{k}\right)^{k_{n-1}}$$

where $\mathcal{P}(n, k)$ is the summation set which happens to be the partition of n into k terms, i.e.,

$$n = k_0 + 2k_1 + \dots + nk_{n-1},$$

$$k = k_0 + \dots + k_{n-1}.$$

Main Results

Theorem 9 (Flajolet and WS, 1998). Consider the class of renewal processes as defined above. The quantity r_n attains the following asymptotics

$$r_n = \frac{2}{\log 2} \sqrt{cn} - \frac{5}{8} \lg n + \frac{1}{2} \lg \log n + O(1)$$

where $c = \frac{\pi^2}{6} - 1 \approx 0.645$.

Asymptotic analysis is sophisticated and follows these steps:

- **first**, we transform r_n into another quantity s_n that we know how to handle and (using a **probabilistic technique**) we know how to read back results for r_n from s_n ;
- use **combinatorial calculus** to find the generating function of s_n , which turns out to be an infinite product of tree-functions $B(z)$ defined above;
- transform this product into a **harmonic sum** that can be analyzed asymptotically by the **Mellin transform**;
- obtain an asymptotic expansion of the generating function around $z = 1$ which is the starting point for extracting the asymptotics of the coefficients;
- finally, estimate $R_n^*(\mathcal{R}_0)$ by the **saddle point method**.

Asymptotics: The Main Idea

The quantity r_n is too hard to analyze due to the factor $k!/k^k$, hence we define a new quantity s_n defined as

$$\begin{cases} s_n &= \sum_{k=0}^n s_{n,k} \\ s_{n,k} &= e^{-k} \sum_{\mathcal{P}(n,k)} \frac{k^{k_0}}{k_0!} \cdots \frac{k^{k_{n-1}}}{k_{n-1}!}. \end{cases}$$

To analyze it, we introduce the random variable K_n as follows

$$\Pr\{K_n = k\} = \frac{s_{n,k}}{s_n}.$$

Stirling's formula yields

$$\begin{aligned} \frac{r_n}{s_n} &= \sum_{k=0}^n \frac{r_{n,k} s_{n,k}}{s_{n,k} s_n} = \mathbf{E}[(K_n)! K_n^{-K_n} e^{-K_n}] \\ &= \mathbf{E}[\sqrt{2\pi K_n}] + O(\mathbf{E}[K_n^{-\frac{1}{2}}]). \end{aligned}$$

Fundamental Lemmas

Lemma 9. Let $\mu_n = \mathbf{E}[K_n]$ and $\sigma_n^2 = \mathbf{Var}(K_n)$.

$$s_n \sim \exp\left(2\sqrt{cn} - \frac{7}{8}\log n + d + o(1)\right)$$

$$\mu_n = \frac{1}{4}\sqrt{\frac{n}{c}}\log\frac{n}{c} + o(\sqrt{n})$$

$$\sigma_n^2 = O(n\log n) = o(\mu_n^2),$$

where $c = \pi^2/6 - 1$, $d = -\log 2 - \frac{3}{8}\log c - \frac{3}{4}\log \pi$.

Lemma 10. For large n

$$\mathbf{E}[\sqrt{K_n}] = \mu_n^{1/2}(1 + o(1))$$

$$\mathbf{E}[K_n^{-1/2}] = o(1).$$

where $\mu_n = \mathbf{E}[K_n]$.

Thus

$$\begin{aligned} r_n &= s_n \mathbf{E}[\sqrt{2\pi K_n}](1 + o(1)) \\ &= s_n \sqrt{2\pi \mu_n}(1 + o(1)). \end{aligned}$$

Sketch of a Proof: Generating Functions

1. Define the function $\beta(z)$ as

$$\beta(z) = \sum_{k=0}^{\infty} \frac{k^k}{k!} e^{-k} z^k.$$

One has (e.g., by **Lagrange inversion** or otherwise)

$$\beta(z) = \frac{1}{1 - T(ze^{-1})}.$$

2. Define

$$s_n(u) = \sum_{k=0}^{\infty} s_{n,k} u^k, \quad S(z, u) = \sum_{n=0}^{\infty} S_n(u) z^n.$$

Since $s_{n,k}$ involves **convolutions of sequences** of the form $k^k/k!$, we have

$$\begin{aligned} S(z, u) &= \sum_{\mathcal{P}_{n,k}} z^{1k_0+2k_1+\dots} \left(\frac{u}{e}\right)^{k_0+\dots+k_{n-1}} \frac{k^{k_0}}{k_0!} \cdots \frac{k^{k_{n-1}}}{k_{n-1}!} \\ &= \prod_{i=1}^{\infty} \beta(z^i u). \end{aligned}$$

We need to compute $s_n = [z^n]S(z, 1)$ where $[z^n]F(z)$ denotes the coefficient at z^n of $F(z)$.

Mellin Asymptotics

3. Let $L(z) = \log S(z, 1)$ and $z = e^{-t}$, so that

$$L(e^{-t}) = \sum_{k=1}^{\infty} \log \beta(e^{-kt}).$$

Mellin transform techniques provide an expansion of $L(e^{-t})$ around $t = 0$ (or equivalently $z = 1$) since the sum falls under the **harmonic sum** paradigm.

4. The **Mellin transform** $L^*(s) = \mathcal{M}(L(e^{-t}); s)$ of $L(e^{-t})$ is computed by the **harmonic sum property (M3)**. For $\Re(s) \in (1, \infty)$, the transform evaluates to

$$L^*(s) = \zeta(s)\Lambda(s)$$

where $\zeta(s) = \sum_{n \geq 1} n^{-s}$ is the Riemann zeta function, and

$$\Lambda(s) = \int_0^{\infty} \log \beta(e^{-t}) t^{s-1} dt.$$

This leads to

$$L^*(s) \asymp \left(\frac{\Lambda(1)}{s-1} \right)_{s=1} + \left(-\frac{1}{4s^2} - \frac{\log \pi}{4s} \right)_{s=0}.$$

What's Next?

5. An application of the **converse mapping property (M4)** allows us to come back to the original function,

$$L(e^{-t}) = \frac{\Lambda(1)}{t} + \frac{1}{4} \log t - \frac{1}{4} \log \pi + O(\sqrt{t}),$$

which translates in

$$L(z) = \frac{\Lambda(1)}{1-z} + \frac{1}{4} \log(1-z) - \frac{1}{4} \log \pi - \frac{1}{2} \Lambda(1) + O(\sqrt{1-z}).$$

where

$$\begin{aligned} c = \Lambda(1) &= - \int_0^1 \log(1 - T(x/e)) \frac{dx}{x} \\ &= \frac{\pi^2}{6} - 1. \end{aligned}$$

6. In summary, we just proved that, as $z \rightarrow 1^-$,

$$S(z, 1) = e^{L(z)} = a(1-z)^{\frac{1}{4}} \exp\left(\frac{c}{1-z}\right) (1 + o(1)),$$

where $a = \exp(-\frac{1}{4} \log \pi - \frac{1}{2}c)$.

7. To extract asymptotic we need to apply the **saddle point method**.

Saddle Point Method

Lemma 11. For positive $A > 0$, and reals B and C , define $f(z) = f_{A,B,C}(z)$ as

$$f(z) = \exp \left(\frac{A}{1-z} + B \log \frac{1}{1-z} + C \log \left(\frac{1}{z} \log \frac{1}{1-z} \right) \right).$$

Then,

$$\begin{aligned} [z^n] f_{A,B,C}(z) &= \exp \left(2\sqrt{An} + \frac{1}{2} \left(B - \frac{3}{2} \right) \log n + C \log \log \sqrt{\frac{n}{A}} \right. \\ &\quad \left. - \frac{1}{2} \log \left(4\pi e^{-A} / \sqrt{A} \right) + o(1) \right). \end{aligned}$$

Proof: We start with Cauchy's formula

$$[z^n] f(z) = \frac{1}{2\pi i} \oint e^{h(z)} dz$$

where $h(z) = \log f_{A,B,C}(z) - (n+1) \log z$. The saddle point r defined by $h'(r) = 0$ is asymptotically equal to

$$r = 1 - \sqrt{\frac{A}{n}} + \frac{B-A}{2n} + o(n^{-1}),$$

and

$$h(r) = 2A\sqrt{\frac{n}{A}} + B \log \left(\sqrt{\frac{n}{A}} \right) + C \log \log \left(\sqrt{\frac{n}{A}} \right) + \frac{1}{2}A + o(1).$$

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
5. Minimax Redundancy: Universal (unknown) Sources
6. Appendix A: Mellin Transform
7. Appendix B: Saddle Point Method

Appendix A: Mellin Properties

(M1) DIRECT AND INVERSE MELLIN TRANSFORMS. Let c belong to the *fundamental strip* defined below.

$$f^*(s) := \mathcal{M}(f(x); s) = \int_0^{\infty} f(x)x^{s-1}dx$$

then

$$f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} f^*(s)x^{-s}ds.$$

(M2) FUNDAMENTAL STRIP. The Mellin transform of $f(x)$ exists in the *fundamental strip* $\Re(s) \in (-\alpha, -\beta)$, where

$$f(x) = O(x^\alpha) \quad (x \rightarrow 0), \quad f(x) = O(x^\beta) \quad (x \rightarrow \infty).$$

(M3) HARMONIC SUM PROPERTY. By linearity and the scale rule $\mathcal{M}(f(ax); s) = a^{-s}\mathcal{M}(f(x); s)$,

$$f(x) = \sum_{k \geq 0} \lambda_k g(\mu_k x)$$

then

$$f^*(s) = g^*(s) \sum_{k \geq 0} \lambda_k \mu_k^{-s}.$$

(M4) MAPPING PROPERTIES (Asymptotic expansion of $f(x)$ and singularities of $f^*(s)$).

$$f(x) = \sum_{(\xi, k) \in A} c_{\xi, k} x^{\xi} (\log x)^k + O(x^M)$$

then

$$f^*(s) \asymp \sum_{(\xi, k) \in A} c_{\xi, k} \frac{(-1)^k k!}{(s + \xi)^{k+1}}.$$

(i) *Direct Mapping*. Assume that $f(x)$ admits as $x \rightarrow 0^+$ the asymptotic expansion of the above for some $-M < -\alpha$ and $k > 0$. Then for $\Re(s) \in (-M, -\beta)$, the transform $f^*(s)$ satisfies the singular expansion of above.

(ii) *Converse Mapping*. Assume that $f^*(s) = O(|s|^{-r})$ with $r > 1$, as $|s| \rightarrow \infty$ and that $f^*(s)$ admits the singular expansion above for $\Re(s) \in (-M, -\alpha)$. Then $f(x)$ satisfies the asymptotic expansion of above at $x = 0^+$.

Outline Update

1. Shannon Information Theory: Three Theorems of Shannon
2. Glance at Channel Coding Theorem
3. Source Coding
4. Redundancy: Known Sources
5. **Minimax Redundancy: Universal (unknown) Sources**
6. Appendix A: Mellin Transform
7. Appendix B: **Saddle Point Method**

Appendix B: Saddle Point Method

Input: A function $g(z)$ analytic in $|z| < R$ ($0 < R < +\infty$) with nonnegative Taylor coefficients and “fast growth” as $z \rightarrow R^-$. Let $h(z) := \log g(z) - (n+1) \log z$.

Output: The asymptotic formula for $g_n := [z^n]g(z)$ derived from the Cauchy coefficient integral

$$g_n = \frac{1}{2i\pi} \int_{\gamma} g(z) \frac{dz}{z^{n+1}} = \frac{1}{2i\pi} \int_{\gamma} e^{h(z)} dz$$

where γ is a loop around $z = 0$.

(S1). SADDLE POINT CONTOUR. *Require that $g'(z)/g(z) \rightarrow +\infty$ as $z \rightarrow R^-$. Let $r = r(n)$ be the unique positive root of the saddle point equation*

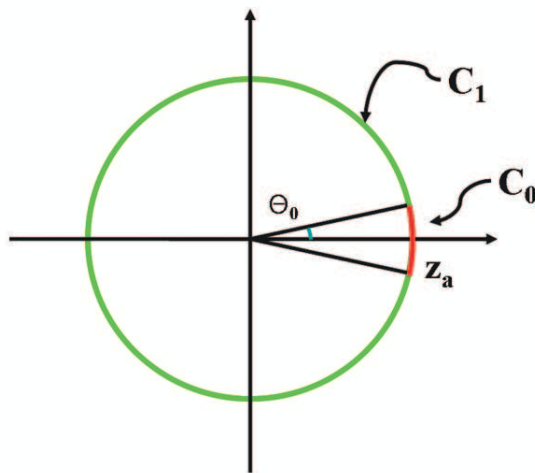
$$h'(r) = 0 \quad \text{or} \quad \frac{rg'(r)}{g(r)} = n + 1,$$

so that $r \rightarrow R$ as $n \rightarrow \infty$. The integral above is evaluated on $\gamma = \{z \mid |z| = r\}$.

(S2). BASIC SPLIT. Require that $h'''(r)^{1/3}h''(r)^{-1/2} \rightarrow 0$. Define $\varphi = \varphi(n)$ called the “range” of the saddle point by

$$\varphi = \left| h'''(r)^{-1/6} h''(r)^{-1/4} \right|,$$

so that $\varphi \rightarrow 0$, $h''(r)\varphi^2 \rightarrow \infty$, and $h'''(r)\varphi^3 \rightarrow 0$. Split $\gamma = \gamma_0 \cup \gamma_1$, where $\gamma_0 = \{z \in \gamma \mid |\arg(z)| \leq \varphi\}$, $\gamma_1 = \{z \in \gamma \mid |\arg(z)| \geq \varphi\}$.



(S3) ELIMINATION OF TAILS. Require that $|g(re^{i\theta})| \leq |g(re^{i\varphi})|$ on γ_1 . Then, the tail integral satisfies the trivial bound,

$$\left| \int_{\gamma_1} e^{h(z)} dz \right| = O \left(|e^{-h(re^{i\varphi})}| \right).$$

(S4) LOCAL APPROXIMATION. Require that $h(re^{i\theta}) - h(r) - \frac{1}{2}r^2\theta^2h''(r) = O(|h'''(r)\varphi^3|)$ on γ_0 . Then, the central integral is asymptotic to a complete Gaussian integral, and

$$\frac{1}{2i\pi} \int_{\gamma_0} e^{h(z)} dz = \frac{g(r)r^{-n}}{\sqrt{2\pi h''(r)}} \left(1 + O(|h'''(r)\varphi^3|)\right).$$

(S5) COLLECTION. Requirements (S1), (S2), (S3), (S4), imply the estimate:

$$[z^n]g(z) = \frac{g(r)r^{-n}}{\sqrt{2\pi h''(r)}} \left(1 + O(|h'''(r)\varphi^3|)\right) \sim \frac{g(r)r^{-n}}{\sqrt{2\pi h''(r)}}.$$