Average Redundancy of the Shannon Code for Markov Sources

Neri Merhav and Wojciech Szpankowski
Technion and Purdue University

May 27, 2013

NSF STC Center for Science of Information

AofA, Menorca 2013

Dedicated to PHILIPPE FLAJOLET

Outline

1. Source Coding
2. Redundancy: Known Sources
3. Shannon and Huffman Redundancy for Memoryless Sources
4. Shannon Coding Redundancy for Markov Sources

Source Coding

A source code is a bijective mapping

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

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:

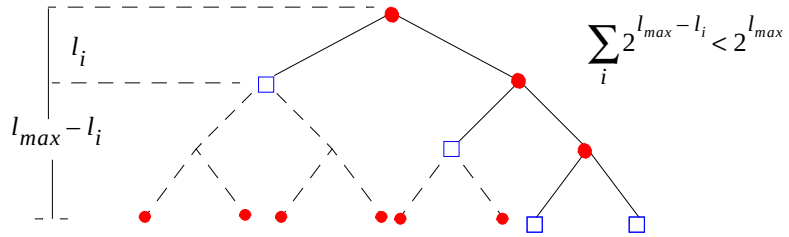
- Fixed-to-Variable (**FV**) length codes (e.g., Huffman and Shannon codes).
- Variable-to-Fixed (**VF**) length codes (e.g., Tunstall and Khodak codes).
- Variable-to-Variable (**VV**) length codes (e.g., Khodak VV code).

Prefix Codes

Prefix code is such that no codeword is a prefix of another codeword.
We write:
$P(x)$ be the probability of $x \in \mathcal{A}^*$;
$L(C, x)$ be the code length for the source sequence $x \in \mathcal{A}^*$;
$H(P) = -\sum_{x \in \mathcal{A}^*} P(x) \lg P(x)$ for entropy.

Kraft's Inequality

A binary code is a prefix code iff the code lengths $\ell_1, \ell_2, \ldots, \ell_N$ satisfy



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

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,

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

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. Source Coding
2. Redundancy: Known Sources
3. Shannon and Huffman Redundancy for Memoryless Sources
4. Shannon Coding Redundancy for Markov Sources

Redundancy for Huffman's Code

We consider **fixed-to-variable length codes**: Shannon & Huffman codes.

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

In 1994 Stubley 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 $n \to \infty$*

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

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

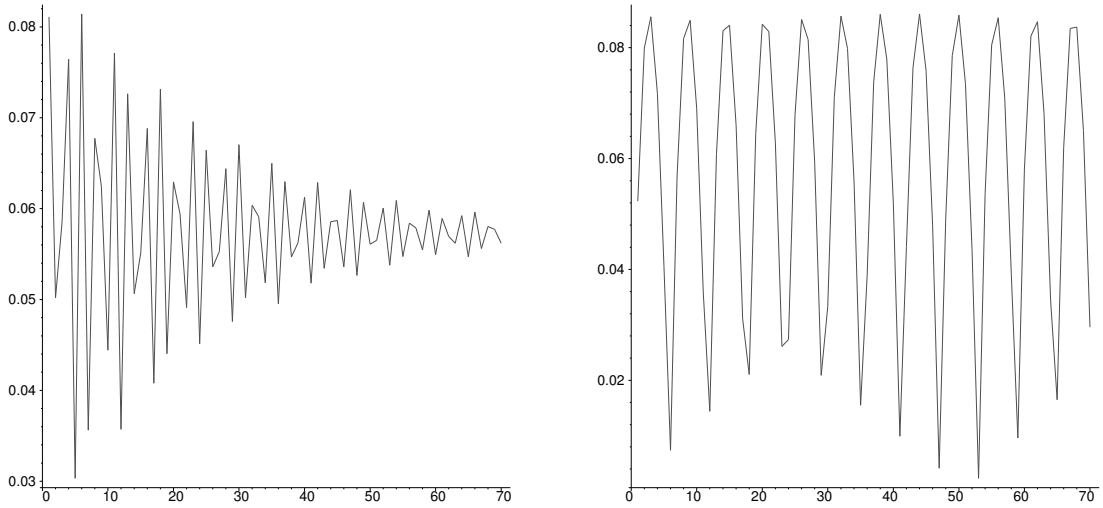


Figure 1: The average redundancy of Huffman codes versus block size n for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

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

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

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 $f : [0,1] \to \mathbf{R}$.

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

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

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 $f : [0,1] \to \mathbb{R}$*

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

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

Shannon Redundancy – Rational Case

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

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

$$= \sum_{\ell=0}^{M-1} \left\langle \frac{\ell}{M} + \beta n \right\rangle \sum_{m:\ k=\ell+mM\le n} p_{n,k}.$$

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

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

Proof. Let $\omega_k = e^{2\pi i k/M}$ for $k = 0, 1, \ldots, M-1$ be the Mth 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

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

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

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

Continuing the derivation and using the above lemma we obtain

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

Outline Update

1. Source Coding
2. Redundancy: Known Sources
3. Shannon and Huffman Redundancy for Memoryless Sources
4. Shannon Coding Redundancy for Markov Sources

Markov Sources

Source sequence $X_1, X_2, \ldots,$ over alphabet $\mathcal{A} = \{1, 2, \ldots, r\}$ is generated by a first–order Markov chain with a given matrix

$$P = \{p(j|k)\}_{j,k=1}^{r}.$$

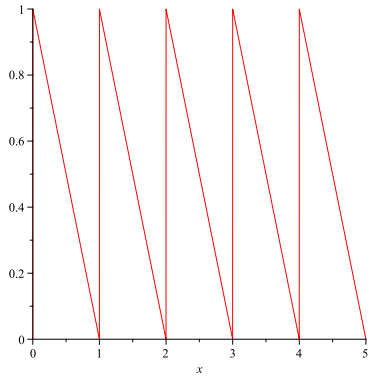
with initial state probabilities p_k, $k = 1, 2, \ldots, r$;
stationary state probabilities π_k, $k = 1, 2, \ldots, r$.

For $x^n = (x_1, \ldots, x_n) \in \mathcal{A}^n$ under the given Markov source, is

$$P(x^n) = p_{x_1} \prod_{t=2}^{n} p(x_t|x_{t-1}).$$

The average redundancy of the Shannon code is defined as

$$R_n = \boldsymbol{E}[\lceil -\log P(X^n) \rceil + \log P(X^n)] = \mathbf{E}[\varrho(-\log P(X^n))].$$



$$\varrho(u) = \lceil u \rceil - u.$$

Main Result for Markov Sources

Theorem 2 (Merhav & W.S.). *Consider the Shannon code of length n for an aperiodic and irreducible Markov source. Define*

$$\alpha_{jk} = \log\left[\frac{p(j|1)p(j|j)}{p(k|1)p(j|k)}\right], \quad j,k \in \{1,2,\ldots,r\}.$$

(a) If not all $\{\alpha_{jk}\}$ are rational, then

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

(b) If all $\{\alpha_{jk}\}$ are rational, then let

$$\zeta_{jk}(n) = M[-(n-1)\log p(1|1) + \log p(j|1) - \log p(k|1) - \log p_j],$$

and

$$\Omega_n = \frac{1}{2}\left(1 - \frac{1}{M}\right) + \frac{1}{M}\sum_{j=1}^{r}\sum_{k=1}^{r} p_j \pi_k \varrho[\zeta_{jk}(n)],$$

M is the smallest common multiple of the denominators of $\{\alpha_{jk}\}$. Then, there exists a positive sequence $\xi_n \to 0$

$$R_n \leq \Omega_n + \frac{1}{M}\sum_{j=1}^{r}\sum_{k=1}^{r} p_j \pi_k \mathcal{I}\{\varrho[\zeta_{jk}(n)] \notin (\xi_n, 1-\xi_n)\} + o(1),$$

$$R_n \geq \Omega_n - \frac{1}{M}\sum_{j=1}^{r}\sum_{k=1}^{r} p_j \pi_k \mathcal{I}\{\varrho[\zeta_{jk}(n)] \notin (\xi_n, 1-\xi_n)\} - o(1).$$

Sketch of Proof

1. We note that $\rho(u)$ has the following Fourier series expansion

$$\varrho(u) = \frac{1}{2} + \sum_{m \neq 0} a_m e^{2\pi i m u}, \qquad a_m = \frac{1}{2\pi m}$$

where $a_{m \cdot k} = a_m / k$ for integers m, k.

2. Since $R_n = \mathbf{E}[\varrho(\log P(X^n))]$ (for aperiodic irreducible MC) we have

$$R_n = \frac{1}{2} + \sum_{m \neq 0} a_m \boldsymbol{E}[e^{-2\pi i m \log P(X^n)}].$$

which we can re-write as

$$R_n = \frac{1}{2} + \sum_{m \neq 0} a_m \sum_{\mathbf{x} \in \mathcal{A}^n} \prod_{t=1}^{n} p(x_t | x_{t-1}) \exp\left[-2\pi i m \log p(x_t | x_{t-1})\right]$$

since $P(x^n) = p_{x_1} \prod_{t=2}^{n} p(x_t | x_{t-1})$.

Continuation

3. Define for $j, k = 1, \ldots, r$

$$a_{jk}(m) = p(k|j) \exp\left[-2\pi i m \log p(k|j)\right], \quad A_m = [a_{jk}(m)]_{j,k=1}^{r},$$

and r–dimensional column vectors:

$$\boldsymbol{c}_m = (p_1 \exp[-2\pi i m \log p_1)], \ldots, p_r \exp[-2\pi i m \log p_r])^T,$$

and $\mathbf{1} = (1, 1, \ldots, 1)^T$. Then

$$R_n = \frac{1}{2} + \sum_{m \neq 0} a_m \cdot \boldsymbol{c}_m^T A_m^{n-1} \mathbf{1}.$$

4. Let $\mathbf{l}_{i,m}$ and $\boldsymbol{r}_{i,m}$ be, respectively, the left eigenvector and the right eigenvector pertaining to the eigenvalue $\lambda_{i,m}$ of the matrix A_m such that $\rho(A_m) := |\lambda_{1,m}| \geq |\lambda_{2,m}| \geq \cdots |\lambda_{r,m}|$. By the spectral representation of matrices

$$A_m^{n-1} \mathbf{1} = \sum_{i=1}^{r} \lambda_{i,m}^{n-1} \cdot (\mathbf{l}_{i,m}^T, \mathbf{1}) \cdot \boldsymbol{r}_{i,m},$$

leading to

$$R_n = \frac{1}{2} + \sum_{m \neq 0} a_m \cdot \sum_{i=1}^{r} \lambda_{i,m}^{n-1} \cdot (\mathbf{l}_{i,m}^T, \mathbf{1}) \cdot (\boldsymbol{c}_m^T, \boldsymbol{r}_{i,m}).$$

Conclusions

5. If (a) all eigenvalues $\lambda_{i,m} < 1$, then by Fejer's theorem

$$R_n \to \frac{1}{2}.$$

(b) If some (largest) $\lambda_{i,m} = 1$, we have oscillatory provided

$$\rho(A_m) = \rho(P) = 1.$$

Lemma 4. *Let $F = \{f_{kj}\}$ and $G = \{g_{kj}\}$ be two $r \times r$ matrices. Assume that F is a real, non–negative and irreducible matrix, G is a complex matrix, and*

$$f_{kj} = |g_{kj}|, \quad k, j \in \{1, 2, \ldots, r\}.$$

Then

$$\rho(G) = \rho(F)$$

if and only if there exist real numbers s, and $w_1, \ldots, w_r$ such that

$$G = e^{2\pi i s} D F D^{-1},$$

where $D = diag\{e^{2\pi i w_1}, \ldots, e^{2\pi i w_r}\}$.

Thus $\rho(A_m) = \rho(P) = 1$ if and only if there exist s and $w_1, \ldots w_r$:

$$-m \log p(j|k) = (s + w_k - w_j) \bmod 1, \quad j, k = 1, \ldots, r,$$

where $x = y \bmod 1$ means $\langle x \rangle = \langle y \rangle$.

Extensions

We can extend our Theorem to irreducible and periodic Markov chains, but not further than this as example below shows.

Example 2. *Reducible Markov source*.
Consider the case $r = 2$ with the following transition matrix

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ 0 & 1 \end{pmatrix}.$$

Assume also that $p_1 = 1$ and $p_2 = 0$. Direct computation shows that

$$R_n = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k \varrho[-\log \alpha - k \log(1-\alpha)] + o(1).$$

We see then that there is *no oscillatory mode* in this case, as R_n always tends to a constant that depends on α.

That's It

THANK YOU