

Slide Set II: Lossless Source Coding and the S-M-B Theorem

- Lossless code
- Kraft inequality
- Entropy
- Entropy rate
- Universal lossless source code
- The Shannon-McMillan-Breiman theorem
- Pointwise universality

Kraft Inequality

Theorem 1. For any uniquely decodable code

$$\sum_{x \in A} 2^{-l(x)} \leq 1. \quad (1)$$

Conversely, given a set of codeword lengths satisfying (1), there exists a prefix code with these word lengths.

Lossless Code

We start from the very beginning. Assume X is a random variable taking values in the finite alphabet A .

Definition 1. A binary source code for X is a mapping $C : A \rightarrow \{0, 1\}^*$.

Denote:

Length function: $l(x) = |C(x)|$

Expected length: $L(C) = \sum_{x \in A} P(x)l(x)$.

Definition 2. A code for X is said to be lossless ([CT91]: “non-singular”) if $x \neq x' \Rightarrow C(x) \neq C(x')$.

Recall also:

Unique decodability (when extension is lossless) and

Prefix code or instantaneous code when no codeword is a prefix of any other.

Entropy

Equipped with Kraft’s inequality one can prove:

Theorem 2. The expected length L of any uniquely decodable code satisfies $L \geq H(X)$.

Proof idea: Define $R(x) = \frac{2^{-l(x)}}{\sum_{x'} 2^{-l(x')}}$ and use positivity of $D(P\|R)$. \square

The proof also hints that a good code should strive for $2^{-l(x)} \approx P(x)$.

Indeed, taking $l(x) = \lceil \log(1/P(x)) \rceil$ (Shannon code) gives

$\log(1/P(x)) \leq l(x) \leq \log(1/P(x)) + 1$, which satisfies Kraft’s inequality and attains $L \leq H(X) + 1$ implying

Theorem 3. The optimal UD code for X (in the sense of minimizing expected length), L^* , satisfies

$$H(X) \leq L^* \leq H(X) + 1.$$

Lower bound in Theorem 3 was based on Kraft's inequality which needed unique decodability. What about a general lossless code? In the HW ex. you will show that it does not quite hold, but almost:

Theorem 4. For any lossless code

$$L(C) \geq H(X) - \log[2 \log(|A| + 2)].$$

Universal Lossless Source Codes defined

Thus, optimum compression ratio is the entropy rate of the process. One example we saw of a (sequence of) code(s) attaining this optimum is the Shannon code

$$l(x^n) = \left\lceil \log \frac{1}{P(x^n)} \right\rceil.$$

Note the dependence on the distribution of the process.

Definition 3. Let C_n be a lossless code for source n -tuples and l_n denote the associated length function. $\{C_n\}$ will be said to be universal if

$$\lim_{n \rightarrow \infty} E \frac{1}{n} l_n(X^n) = \overline{H}(\mathbf{X})$$

for all stationary processes \mathbf{X} .

Assume $\mathbf{X} = \{X_t\}$ is emitted by a stationary source. Let $\overline{H}(\mathbf{X})$ denote its entropy rate defined by $\overline{H}(\mathbf{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X^n)$ (recall existence of limit).

Theorems 3 and 4 imply:

Theorem 5. Let L_n^* denote the minimum expected codeword length per symbol in lossless coding of $X^n = (X_1, \dots, X_n)$. Then

$$L_n^* \rightarrow \overline{H}(\mathbf{X}). \quad (2)$$

In words, the entropy rate of the process is the minimum expected number of bits per symbol required to losslessly describe the process, in the limit of large block length, *with no sequentiality (or any other type of) constraints*.

Universal Lossless Source Codes (cont.)

The first basic question about a universal (sequence of) source code(s) is whether such a creature exists.

The answer is: Yes.

1. Otherwise we wouldn't have bother defining them..
2. We will see two examples of a universal lossless source code.
 - (a) The first is a simplistic scheme for developing intuition as to how it is that such codes exist.
 - (b) The second is the celebrated Lempel-Ziv scheme.

One technical point before we describe the simplistic scheme:

Lemma 1. For any $X \in A$ there exists a UD code whose length function satisfies $\max_{x \in A} l(x) \leq \lceil \log_2 |A| \rceil + 1$ and

$$El(X) \leq H(X) + 2.$$

A Simplistic Universal Scheme (cont.)

So we got a universal (sequence of) scheme(s). Note, however,

- Need to look at the whole block before passing the first bit to the decoder.
- Need to perform an exhaustive search over family of possible schemes.

We will see a scheme which is

- Essentially sequential (delay negligible relative to block length).
- Easy to implement.

But before that we digress to develop another notion of optimality and universality, concerning the “sample” properties of the source.

-Divide block into n/k k -blocks

-Find, among all k -block schemes (with codelength bounded by $\lceil k \log_2 |A| \rceil + 1$), the one minimizing the overall code-length when applied successively on the sub-blocks

-Give index of that scheme, and then encode each block using it

Analysis: Letting \mathcal{N}_k denote the above family of lossless codes for block length k and N_k denote its size, bounding generously we will see that

$$\frac{1}{n} E[\text{code-length of scheme}] \leq \frac{1}{n} \log N_k + \frac{1}{k} [H(X^k) + 2] + \frac{1}{k} \log |A|. \quad (3)$$

So, if we let k grow with n but sufficiently slowly that $\frac{1}{n} \log N_{k_n} \rightarrow 0$, the right side of (3) $\rightarrow \overline{H}(\mathbf{X})$.

Towards the Shannon-McMillan-Breiman Theorem

While Definition 3 involved the expected performance of the code, it is often of interest to look at its actual performance.

In this context, we recall the AEP from 376A stating that when $\{X_i\}$ are i.i.d.

$$-\frac{1}{n} \log P(X^n) \rightarrow H(X_1) \quad \text{in probability.} \quad (4)$$

This was shown to imply that for

$$A_n^\epsilon \triangleq \left\{ \left| -\frac{1}{n} \log P(X^n) - H(X_1) \right| < \epsilon \right\}$$

and all sufficiently large n

$$\Pr(A_n^\epsilon) \geq 1 - \epsilon \quad \text{and} \quad 2^{n(H(X_1) - \epsilon)} \leq |A_n^\epsilon| \leq 2^{n(H(X_1) + \epsilon)}.$$

In words, there are $\approx 2^{nH(X_1)}$ source sequences capturing most of the probability. Can focus our coding efforts on them and have a code with length no more than the entropy with high probability.

Note that by the SLLN (4) actually holds with probability one.

The AEP turns out to hold much beyond i.i.d. sources, just as the SLLN holds much beyond i.i.d. sources. Processes for which the SLLN holds are known as *ergodic*. More precisely

Definition 4. Let $\mathbf{X} = (\dots X_{-1}, X_0, X_1, \dots)$ be stationary and $T(\mathbf{X})$ denote its shift (i.e., if $\mathbf{Y} = T(\mathbf{X})$ then $Y_t = X_{t-1}$). Let $T^k(\mathbf{X})$ denote the k -shift. \mathbf{X} is said to be ergodic if for every measurable $f : A^\infty \rightarrow \mathbb{R}$ with $E|f(\mathbf{X})| < \infty$

$$\frac{1}{n} \sum_{i=1}^n f(T^k(\mathbf{X})) \rightarrow Ef(\mathbf{X}) \quad \text{with probability 1.} \quad (5)$$

Qualitatively, a process is ergodic if it has the property that its statistical characterization can be inferred from an observation of its realization.

Ergodic theory is a rich domain. It is developed in a much more general framework of measure-preserving transformations (cf. [Petersen89] and MATH235). See also [Gray88] and:

Paul C. Shields, "The Interactions Between Ergodic Theory and Information Theory", IEEE Trans. Info. Theory, vol. IT-44, pp. 2079 - 2093, October 1998.

Remark on Definition 4 for those familiar with conditional expectation at the level of STAT310:

An event B is called invariant if $B = T(B)$ (here we consider two events to be equal if $P(B \Delta T(B)) = 0$).

You can convince yourself that the class of invariant events \mathcal{I} is a σ -field.

Birkhoff's (also known as the "pointwise") *ergodic theorem* states that for any stationary \mathbf{X} and f as in Definition (4),

$$\frac{1}{n} \sum_{i=1}^n f(T^k(\mathbf{X})) \rightarrow E[f(\mathbf{X})|\mathcal{I}] \quad \text{with probability 1.} \quad (6)$$

A process is then defined as being ergodic if \mathcal{I} is trivial. Note the equivalence with Definition 4.

Another intuition: \mathbf{X} is ergodic if it is not a mixture of two other stationary processes.

The Shannon-McMillan-Breiman Theorem

Equipped with the notion of ergodicity, we can now present the SMB theorem, also known as the "ergodic theorem of information theory" or the "AEP".

Theorem 6. If \mathbf{X} is stationary ergodic then

$$-\frac{1}{n} \log P(X^n) \rightarrow \overline{H}(\mathbf{X}) \quad w.p.1.$$

Actually:

S- did it for convergence in probability and Markov sources

M- did it for L_1 convergence

B- did it for a.s. convergence

Was later generalized in various directions, see [Gray88], Chapter 3.

We will outline the elegant proof of [CT91], some of the ideas of which will recur in other problems to be treated.

Define:

$$H^k = E[-\log P(X_0|X_{-k}^{-1})]$$

$$H^\infty = E[-\log P(X_0|X_{-\infty}^{-1})]$$

And let P^k denote the k -th order Markov approximation of P :

$$P^k(x^n) = P(x^k) \prod_{i=k+1}^n P(x_i|x_{i-k}^{i-1})$$

The proofs are easy when using martingales and the Borel-Cantelli lemmas (to be discussed at level of detail dependent on backgrounds).

Proof of SMB: With probability one:

$$\begin{aligned} \overline{H}(\mathbf{X}) &= H^\infty \\ &= \lim -\frac{1}{n} \log P(X^n|X_{-\infty}^0) \\ &\leq \liminf -\frac{1}{n} \log P(X^n) \\ &\leq \limsup -\frac{1}{n} \log P(X^n) \\ &\leq \limsup -\frac{1}{n} \log P^k(X^n) \\ &= H^k \rightarrow \overline{H}(\mathbf{X}). \square \end{aligned}$$

Lemma 2. *With probability one*

$$-\frac{1}{n} \log P^k(X^n) \rightarrow H^k \quad \text{and} \quad -\frac{1}{n} \log P(X^n|X_{-\infty}^0) \rightarrow H^\infty.$$

Lemma 3. H^k is decreasing and $\lim_{k \rightarrow \infty} H^k = \overline{H}(\mathbf{X}) = H^\infty$.

Lemma 4. *With probability one*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{P^k(X^n)}{P(X^n)} \leq 0 \tag{7}$$

and

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{P(X^n)}{P(X^n|X_{-\infty}^0)} \leq 0. \tag{8}$$

Consequences

Corollary 1. *For the Shannon code ($l_n(x^n) = \lceil \log 1/P(x^n) \rceil$), with probability one,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} l_n(X^n) = \overline{H}(\mathbf{X}),$$

provided the source is stationary and ergodic.

But is the entropy rate also a lower bound on achievable compression even in this “pointwise” setting ?

The answer, as you will prove in your HW exercise, is yes:

Theorem 7. *For any sequence of lossless codes and any stationary ergodic process*

$$\liminf_{n \rightarrow \infty} \frac{1}{n} l_n(X^n) \geq \overline{H}(\mathbf{X}) \quad \text{w.p. 1.} \tag{9}$$

A Stronger Notion of Universality

In the context of Corollary 1 and Theorem 7 we paraphrase Definition 3 as follows:

Definition 5. $\{C_n\}$ will be said to be pointwise universal if

$$\lim_{n \rightarrow \infty} \frac{1}{n} l_n(X^n) = \overline{H}(\mathbf{X}) \text{ w.p. } 1$$

for all stationary and ergodic \mathbf{X} .

Question: Do there exist pointwise universal source codes ?

Answer: Yes. The Lempel Ziv scheme we now turn to is one such example.