§ 9. Universal compression

In this lecture we will discuss how to produce compression schemes that do not require apriori knowledge of the distribution. Here, compressor is a map $\mathcal{X}^n \to \{0,1\}^*$. Now, however, there is no one fixed probability distribution P_{X^n} on $\mathcal{X}^n$. The plan for this lecture is as follows:

1. We will start by discussing the earliest example of a universal compression algorithm (of Fitingof). It does not talk about probability distributions at all. However, it turns out to be asymptotically optimal simulatenously for all i.i.d. distributions and with small modifications for all finite-order Markov chains.

2. Next class of universal compressors is based on assuming that a the true distribution P_{X^n} belongs to a given class. These methods proceed by choosing a good model distribution Q_{X^n} serving as the minimax approximation to each distribution in the class. The compression algorithm is designed to work for Q_{X^n} is made.

3. Finally, an entirely different idea are algorithms of Lempel-Ziv type. These automatically adapt to the distribution of the source, without any prior assumptions required.

Throughout this section instead of describing each compression algorithm, we will merely specify some distribution Q_{X^n} and apply one of the following constructions:

- Sort all x^n in the order of decreasing $Q_{X^n}(x^n)$ and assign values from $\{0,1\}^*$ as in Theorem 6.1, this compressor has lengths satisfying

$$\ell(f(x^n)) \le \log \frac{1}{Q_{X^n}(x^n)}\,.$$

- Set lengths to be

$$\ell(f(x^n)) \triangleq \lceil \log \frac{1}{Q_{X^n}(x^n)} \rceil$$

  and apply Kraft's inequality Theorem 6.5 to construct a prefix code.

- Use arithmetic coding (see next section).

The important conclusion is that in all these cases we have

$$\ell(f(x^n)) \le \log \frac{1}{Q_{X^n}(x^n)} + \text{const}\,,$$

and in this way we may and will always replace lengths with $\log \frac{1}{Q_{X^n}(x^n)}$. *In this way, the only job of a universal compression algorithm is to specify* Q_{X^n}.

Remark 9.1. Furthermore, if we only restrict attention to prefix codes, then any code $f : \mathcal{X}^n \to \{0,1\}^*$ defines a distribution $Q_{X^n}(x^n) = 2^{-\ell(f(x^n))}$ (we assume the code's tree is full). In this way, for prefix-free codes results on redundancy, stated in terms of optimizing the choice of Q_{X^n}, imply tight converses too. For one-shot codes without prefix constraints the optimal answers are slightly different, however. (For example, the optimal universal code for all i.i.d. sources satisfies $\mathbb{E}[\ell(f(X^n))] \approx H(X^n) + \frac{|\mathcal{X}|-3}{2}\log n$ in contrast with $\frac{|\mathcal{X}|-1}{2}\log n$ for prefix-free codes.)

9.1 Arithmetic coding

Constructing an encoder table from Q_{X^n} may require a lot of resources if n is large. Arithmetic coding provides a convenient workaround by allowing to output bits sequentially. *Notice that to do so, it requires that not only Q_{X^n} but also its marginalizations $Q_{X^1}, Q_{X^2}, \cdots$ be easily computable.* (This is not the case, for example, for Shtarkov distributions (9.8)-(9.9), which are not compatible for different n.)

Let us agree upon some ordering on the alphabet of $\mathcal{X}$ (e.g. $\mathsf{a} < \mathsf{b} < \cdots < \mathsf{z}$) and extend this order lexicographically to $\mathcal{X}^n$ (that is for $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, we say $x < y$ if $x_i < y_i$ for the first i such that $x_i \neq y_i$, e.g., $\mathsf{baba} < \mathsf{babb}$). Then let

$$F_n(x^n) = \sum_{y^n < x^n} Q_{X^n}(x^n)\,.$$

Associate to each x^n an interval $I_{x^n} = [F_n(x^n), F_n(x^n) + Q_{X^n}(x^n))$. These intervals are disjoint subintervals of $[0,1)$. Now encode

$$x^n \mapsto \text{largest dyadic interval contained in } I_{x^n}\,.$$

Recall that dyadic intervals are intervals of the type $[m2^{-k}, (m+1)2^{-k}]$ where a is an odd integer. Clearly each dyadic interval can be associated with a binary string in $\{0,1\}^*$. We set $f(x^n)$ to be that string. The resulting code is a prefix code satisfying

$$\ell(f(x^n)) \le \left\lceil \log_2 \frac{1}{Q_{X^n}(x^n)} \right\rceil + 1\,.$$

(This is an exercise.)

Observe that

$$F_n(x^n) = F_{n-1}(x^{n-1}) + Q_{X^{n-1}}(x^{n-1}) \sum_{y < x_n} Q_{X_n|X^{n-1}}(y|x^{n-1})$$

and thus $F_n(x^n)$ can be computed sequentially *if $Q_{X^{n-1}}$ and $Q_{X_n|X^{n-1}}$ are easy to compute.* This method is the method of choice in many modern compression algorithms because it allows to dynamically incorporate the learned information about the stream, in the form of updating $Q_{X_n|X^{n-1}}$ (e.g. if the algorithm detects that an executable file contains a long chunk of English text, it may temporarily switch to $Q_{X_n|X^{n-1}}$ modeling the English language).

9.2 Combinatorial construction of Fitingof

Fitingof suggested that a sequence $x^n \in \mathcal{X}^n$ should be prescribed information $\Phi_0(x^n)$ equal to the logarithm of the number of all possible permutations obtainable from x^n (i.e. log-size of the

type-class containing x^n). From Stirling's approximation this can be shown to be

$$\Phi_0(x^n) = nH(x_T) + O(\log n) \quad T \sim \mathrm{Unif}[n] \tag{9.1}$$

$$= nH(\hat{P}_{x^n}) + O(\log n)\,, \tag{9.2}$$

where $\hat{P}_{x^n}$ is the empirical distribution of the sequence x^n:

$$\hat{P}_{x^n}(a) \triangleq \frac{1}{n}\sum_{i=1}^{n} 1\{x_i = a\}\,. \tag{9.3}$$

Then Fitingof argues that it should be possible to produce a prefix code with

$$\ell(f(x^n)) = \Phi_0(x^n) + O(\log n)\,. \tag{9.4}$$

This can be done in many ways. In the spirit of what we will do next, let us define

$$Q_{X^n}(x^n) \triangleq \exp\{-\Phi_0(x^n)\}c_n\,,$$

where c_n is a normalization constant c_n. Counting the number of different possible empirical distributions (types), we get

$$c_n = O(n^{-(|\mathcal{X}|-1)})\,,$$

and thus, by Kraft inequality, there must exist a prefix code with lengths satisfying (9.4). Now taking expectation over $X^n \overset{\text{i.i.d.}}{\sim} P_X$ we get

$$\mathbb{E}[\ell(f(X^n))] = nH(P_X) + (|\mathcal{X}| - 1)\log n + O(1)\,,$$

for every i.i.d. source on $\mathcal{X}$.

9.2.1 Universal compressor for all finite-order Markov chains

Fitingof's idea can be extended as follows. Define now the 1-st order information content $\Phi_1(x^n)$ to be the log of the number of all sequences, obtainable by permuting x^n with extra restriction that the new sequence should have the same statistics on digrams. Asymptotically, Φ_1 is just the conditional entropy

$$\Phi_1(x^n) = nH(x_T|x_{T-1 \mod n}) + O(\log n), \quad T \sim \mathrm{Unif}[n]\,.$$

Again, it can be shown that there exists a code such that lengths

$$\ell(f(x^n)) = \Phi_1(x^n) + O(\log n)\,.$$

This implies that for every 1-st order stationary Markov chain $X_1 \to X_2 \to \cdots \to X_n$ we have

$$\mathbb{E}[\ell(f(X^n))] = nH(X_2|X_1) + O(\log n)\,.$$

This can be further continued to define $\Phi_2(x^n)$ and build a universal code, asymptotically optimal for all 2-nd order Markov chains etc.

9.3 Optimal compressors for a class of sources. Redundancy.

So we have seen that we can construct compressor $f : \mathcal{X}^n \to \{0,1\}^*$ that achieves

$$\mathbb{E}[\ell(f(X^n))] \le H(X^n) + o(n)\,,$$

simultaneously for all i.i.d. sources (or even all r-th order Markov chains). What should we do next? Krichevsky suggested that the next barrier should be to optimize regret, or *redundancy*:

$$\mathbb{E}[\ell(f(X^n))] - H(X^n) \to \min$$

simultaneously for a class of sources. We proceed to rigorous definitions.

Given a collection $\{P_{X^n|\theta}, \theta \in \Theta\}$ of sources, and a compressor $f : \mathcal{X}^n \to \{0,1\}^*$ we define its redundancy as

$$\sup_{\theta_0} \mathbb{E}[\ell(f(X^n))|\theta = \theta_0] - H(X^n|\theta = \theta_0)\,.$$

Replacing here lengths with $\log \frac{1}{Q_{X^n}}$ we define redundancy of the distribution Q_{X^n} as

$$\sup_{\theta_0} D(P_{X^n|\theta=\theta_0} \| Q_{X^n})\,.$$

Thus, the question of designing the best universal compressor (in the sense of optimizing worst-case deviation of the average length from the entropy) becomes the question of finding solution of:

$$Q^*_{X^n} = \operatorname*{argmin}_{Q_{X^n}} \sup_{\theta_0} D(P_{X^n|\theta=\theta_0} \| Q_{X^n})\,.$$

We therefore get to the following definition

Definition 9.1 (Redundancy in universal compression)**.** Given a class of sources $\{P_{X^n|\theta=\theta_0}, \theta_0 \in \Theta, n = 1, \ldots\}$ we define its minimax redundancy as

$$R^*_n \triangleq \min_{Q_{X^n}} \sup_{\theta_0} D(P_{X^n|\theta=\theta_0} \| Q_{X^n})\,. \tag{9.5}$$

Note that under condition of finiteness of R^*_n, Theorem 4.5 gives the maximin and capacity representation

$$R^*_n = \sup_{P_\theta} \min_{Q_{X^n}} D(P_{X^n|\theta} \| Q_{X^n}|P_\theta) \tag{9.6}$$

$$= \sup_{P_\theta} I(\theta; X^n)\,. \tag{9.7}$$

Thus redundancy is simply the capacity of the channel $\theta \to X^n$. This result, obvious in hindsight, was rather surprising in the early days of universal compression.

Finding exact Q_{X^n}-minimizer in (9.5) is a daunting task even for the simple class of all i.i.d. Bernoulli sources (i.e. $\Theta = [0,1]$, $P_{X^n|\theta} = \mathrm{Bern}^n(\theta)$). It turns out, however, that frequently the approximate minimizer has a rather nice structure: it matches the Jeffreys prior.

Remark 9.2. (Shtarkov and Fitingof) There is a connection between the combinatorial method of Fitingof and the method of optimality for a class. Indeed, following Shtarkov we may want to choose distribution $Q^{(S)}_{X^n}$ so as to minimize the worst-case redundancy *for each realization* x^n (not average!):

$$\min_{Q_{X^n}(x^n)} \sup_{\theta_0} \log \frac{P_{X^n|\theta}(x^n|\theta_0)}{Q_{X^n}(x^n)}$$

This leads to Shtarkov's distribution:

$$Q^{(S)}_{X^n}(x^n) = c \sup_{\theta_0} P_{X^n|\theta}(x^n|\theta_0)\,, \tag{9.8}$$

where c is the normalization constant. If class $\{P_{X^n|\theta}, \theta \in \Theta\}$ is chosen to be all i.i.d. distributions on $\mathcal{X}$ then

$$\text{i.i.d.} \quad Q^{(S)}_{X^n}(x^n) = c \exp\{-nH(\hat{P}_{x^n})\}\,, \tag{9.9}$$

and thus compressing w.r.t. $Q^{(S)}_{X^n}$ recovers Fitingof's construction Φ_0 up to $O(\log n)$ differences between $nH(\hat{P}_{x^n})$ and $\Phi_0(x^n)$. If we take $P_{X^n|\theta}$ to be all 1-st order Markov chains, then we get construction Φ_1 etc.

9.4* Approximate minimax solution: Jeffreys prior

In this section we will only consider the simple setting of a class of sources consisting of all i.i.d. distributions on a given finite alphabet. We will show that the prior, asymptoticall solving capacity question (9.7), is given by the Dirichlet-distribution with parameters set to 1/2, namely the pdf

$$P^*_\theta = \text{const} \frac{1}{\sqrt{\prod_{j=0}^d \theta_j}}\,.$$

First, we give the formal setting as follows:

- Fix $\mathcal{X}$ – finite alphabet of size $|\mathcal{X}| = d+1$, which we will enumerate as $\mathcal{X} = \{0, \ldots, d\}$.
- $\Theta = \{(\theta_j, j = 1, \ldots, d) : \sum_{j=1}^d \theta_j \le 1, \theta_j \ge 0\}$ – is the collection of all probability distributions on $\mathcal{X}$. Note that Θ is a d-dimensional simplex. We will also define

$$\theta_0 \triangleq 1 - \sum_{j=1}^d \theta_j\,.$$

- The source class is

$$P_{X^n|\theta}(x^n|\theta) \triangleq \prod_{j=1}^n \theta_{x_j} = \exp\left\{-n \sum_{a \in \mathcal{X}} \theta_a \log \frac{1}{\hat{P}_{x^n}(a)}\right\}\,,$$

  where as before $\hat{P}_{x^n}$ is the empirical distribution of x^n, cf. (9.3).

In order to derive the caod $Q^*_{X^n}$ we first propose a guess that the caid P_θ in (9.7) is some distribution with smooth density on Θ (this can only be justified by an apriori belief that the caid in such a natural problem should be something that employs all θ's). Then, we define

$$Q_{X^n}(x^n) \triangleq \int_\Theta P_{X^n|\theta}(x^n|\theta') P_\theta(\theta') d\theta'\,. \tag{9.10}$$

Before proceeding further, we recall the following method of approximating exponential integrals (called Laplace method). Suppose that $f(\theta)$ has a unique minimum at the interior point $\hat{\theta}$ of Θ

and that Hessian Hess$f$ is uniformly lower-bounded by a multiple of identity (in particular, $f(\theta)$ is strongly convex). Then taking Taylor expansion of π and f we get

$$\int_\Theta \pi(\theta)e^{-nf(\theta)}d\theta = \int (\pi(\hat\theta)+O(\|t\|))e^{-n(f(\hat\theta)-\frac{1}{2}t^T \mathrm{Hess} f(\hat\theta)t+o(\|t\|^2))}dt \tag{9.11}$$

$$= \pi(\hat\theta)e^{-nf(\hat\theta)}\int_{\mathbb{R}^d} e^{-x^T \mathrm{Hess} f(\hat\theta)x}\frac{dx}{\sqrt{n^d}}(1+O(n^{-1/2})) \tag{9.12}$$

$$= \pi(\hat\theta)e^{-nf(\hat\theta)}\left(\frac{2\pi}{n}\right)^{\frac{d}{2}}\frac{1}{\sqrt{\det \mathrm{Hess} f(\hat\theta)}}(1+O(n^{-1/2})) \tag{9.13}$$

where in the last step we computed Gaussian integral.

Next, we notice that

$$P_{X^n|\theta}(x^n|\theta') = e^{-n(D(\hat P_{x^n}\|P_{X|\theta=\theta'})+H(\hat P_{x^n}))\log e}\,,$$

and therefore, denoting

$$\hat\theta(x^n) \triangleq \hat P_{x^n}$$

we get from applying (9.13) to (9.10)

$$\log Q_{X^n}(x^n) = -nH(\hat\theta)+\frac{d}{2}\log\frac{2\pi}{n\log e}+\log\frac{P_\theta(\hat\theta)}{\sqrt{\det J_F(\hat\theta)}}+O(n^{-\frac{1}{2}})\,,$$

where we used the fact that $\mathrm{Hess}_{\theta'} D(\hat P\|P_{X|\theta=\theta'}) = \frac{1}{\log e}J_F(\theta')$ with J_F – Fisher information matrix, see (4.13). From here, using the fact that under $X^n \sim P_{X^n|\theta=\theta'}$ the random variable $\hat\theta = \theta' + O(n^{-1/2})$ we get by linearizing $J_F(\cdot)$ and $P_\theta(\cdot)$

$$D(P_{X^n|\theta=\theta'}\|Q_{X^n}) = n(\mathbb{E}[H(\hat\theta)]-H(X|\theta=\theta'))+\frac{d}{2}\log n-\log\frac{P_\theta(\theta')}{\sqrt{\det J_F(\theta')}}+\mathrm{const}+O(n^{-\frac{1}{2}})\,, \tag{9.14}$$

where const is some constant (independent of prior P_θ or θ'). The first term is handled by the next Lemma.

Lemma 9.1. *Let $X^n \overset{i.i.d.}{\sim} P$ on finite alphabet $\mathcal{X}$ and let $\hat P$ be the empirical type of X^n then*

$$\mathbb{E}[D(\hat P\|P)] = \frac{|\mathcal{X}|-1}{2n}\log e+o(\frac{1}{n})\,.$$

Proof. Notice that $\sqrt{n}(\hat P - P)$ converges in distribution to $\mathcal{N}(0,\Sigma)$, where $\Sigma = \mathrm{diag}(P)-PP^T$, where P is an $|\mathcal{X}|$-by-1 column vector. Thus, computing second-order Taylor expansion of $D(\cdot\|P)$, cf. (4.15), we get the result. □

Continuing (9.14) we get in the end

$$D(P_{X^n|\theta=\theta'}\|Q_{X^n}) = \frac{d}{2}\log n-\log\frac{P_\theta(\theta')}{\sqrt{\det J_F(\theta')}}+\mathrm{const}+O(n^{-\frac{1}{2}}) \tag{9.15}$$

under the assumption of smoothness of prior P_θ and that θ' is not too close to the boundary. Consequently, we can see that in order for the prior P_θ be the saddle point solution, we should have

$$P_\theta(\theta') \sim \sqrt{\det J_F(\theta')}\,,$$

provided that such density is normalizable. Prior proportional to square-root of the determinant of Fisher information matrix is known as *Jeffreys prior*. In our case, using the explicit expression for Fisher information (4.16) we get

$$P_\theta^* = \text{Beta}(1/2, 1/2, \cdots, 1/2) = c_d \frac{1}{\sqrt{\prod_{j=0}^d \theta_j}}, \tag{9.16}$$

where c_d is the normalization constant. The corresponding redundancy is then

$$R_n^* = \frac{d}{2} \log \frac{n}{2\pi e} - \log c_d + o(1). \tag{9.17}$$

Remark 9.3. In statistics Jeffreys prior is justified as being invariant to smooth reparametrization, as evidenced by (4.14). For example, in answering "will the sun rise tomorrow", Laplace proposed to estimate the probability by modeling sunrise as i.i.d. Bernoulli process with a uniform prior on $\theta \in [0,1]$. However, this is clearly not very logical, as one may equally well postulate uniformity of $\alpha = \theta^{10}$ or $\beta = \sqrt{\theta}$. Jeffreys prior $\theta \sim \frac{1}{\sqrt{\theta(1-\theta)}}$ is invariant to reparametrization in the sense that if one computed $\sqrt{\det J_F(\alpha)}$ under α-parametrization the result would be exactly the pushforward of the $\frac{1}{\sqrt{\theta(1-\theta)}}$ along the map $\theta \mapsto \theta^{10}$.

Making the arguments in this subsection rigorous is far from trivial, see [CB90, CB94] for details.

9.5 Sequential probability assignment: Krichevsky-Trofimov

From (9.16) it is not hard to derive the (asymptotically) optimal universal probability assignment Q_{X^n}. For simplicity we consider Bernoulli case, i.e. $d = 1$ and $\theta \in [0,1]$ is the 1-dimensional parameter. Then,[1]

$$P_\theta^* = \frac{1}{\pi\sqrt{\theta(1-\theta)}} \tag{9.18}$$

$$Q_{X^n}^*(x^n) = \frac{(2t_0-1)!! \cdot (2t_1-1)!!}{2^n n!}, \qquad t_a = \#\{j \le n : x_j = a\} \tag{9.19}$$

This assignment can now be used to create a universal compressor via one of the methods outlined in the beginning of this lecture. However, what is remarkable is that it has a very nice sequential interpretation (as does any assignment obtained via $Q_{X^n} = \int P_\theta P_{X^n|\theta}$ with P_θ not depending on n).

$$Q_{X_n|X^{n-1}}(1|x^{n-1}) = \frac{t_1 + \frac{1}{2}}{n}, \qquad t_1 = \#\{j \le n-1 : x_j = 1\} \tag{9.20}$$

$$Q_{X_n|X^{n-1}}(0|x^{n-1}) = \frac{t_0 + \frac{1}{2}}{n}, \qquad t_0 = \#\{j \le n-1 : x_j = 0\} \tag{9.21}$$

This is the famous "add 1/2" rule of Krichevsky and Trofimov. Note that this sequential assignment is very convenient for use in prediction as well as in implementing an arithmetic coder.

[1] This is obtained from identity $\int_0^1 \frac{\theta^a(1-\theta)^b}{\sqrt{\theta(1-\theta)}} d\theta = \pi \frac{1\cdot 3\cdots(2a-1)\cdot 1\cdot 3\cdots(2b-1)}{2^{a+b}(a+b)!}$ for integer $a, b \ge 0$. This identity can be derived by change of variable $z = \frac{\theta}{1-\theta}$ and using the standard keyhole contour on the complex plain.

Remark 9.4. Notice that attaining the first order term $\frac{d}{2}\log n$ in (9.17) is easy. For example, taking Q_{X^n} to be the result of uniform P_θ does achieve this redundancy. In the Bernoulli ($d = 1$) case, the corresponding successive probability is given by

$$Q_{X_n|X^{n-1}}(1|x^{n-1}) = \frac{t_1 + 1}{n+1}, \quad t_1 = \#\{j \le n-1 : x_j = 1\}.$$

This is known as Laplace's "add 1" rule.

9.6 Lempel-Ziv compressor

So given a class of sources $\{P_{X^n|\theta}, \theta \in \Theta\}$ we have shown how to produce an asymptotically optimal compressors by using Jeffreys' prior. Although we have done so only for i.i.d. class, it can be extended to handle a class of all r-th order Markov chains with minimal modifications. However, the resulting sequential probability becomes rather complex. Can we do something easier at the expense of losing optimal redundancy?

In principle, the problem is rather straightforward: as we observe a stationary process, we may estimate with better and better precision the conditional probability $\hat{P}_{X_n|X_{n-r}^{n-1}}$ and then use it as the basis for arithmetic coding. As long as $\hat{P}$ converges to the actual conditional probability, we will get to the entropy rate of $H(X_n|X_{n-r}^{n-1})$. Note that Krichevsky-Trofimov assignment (9.21) is clearly learning the distribution too: as n grows, the estimator $Q_{X_n|X^{n-1}}$ converges to the true P_X (provided sequence is i.i.d.). So in some sense the converse is also true: *any good universal compression scheme is inherently learning the true distribution.*

The main drawback of the learn-then-compress approach is the following. Once we extend the class of sources to include those with memory, we invariably are lead to the problem of learning the joint distribution $P_{X_0^{r-1}}$ of r-blocks. However, the number of samples required to obtain a good estimate of $P_{X_0^{r-1}}$ is exponential in r. Thus learning may proceed rather slowly. Lempel-Ziv family of algorithms works around this in an ingeniously elegant way:

- First, estimating probabilities of rare substrings takes longest, but it is also the least useful, as these substrings almost never appear at the input.
- Second, *and most crucial*, observation is that a great estimate of the $P_{X^r}(x^r)$ is given by the reciprocal of the distance to the last observation of x^r in the incoming stream.
- Third, there is a prefix code[2] mapping any integer n to binary string of length roughly $\log_2 n$:

$$f_{int} : \mathbb{Z}_+ \to \{0,1\}^+, \qquad \ell(f_{int}(n)) = \log_2 n + O(\log\log n). \tag{9.22}$$

  Thus, by encoding the pointer to the last observation of x^r via such a code we get a string of length roughly $\log P_{X^r}(x^r)$ automatically.

There are a number of variations of these basic ideas, so we will only attempt to give a rough explanation of why it works, without analyzing any particular algorithm.

We proceed to formal details. First, we need to establish a Kac's lemma.

[2]For this just notice that $\sum_{k\ge 1} 2^{-\log_2 k - 2\log_2\log(k+1)} < \infty$ and use Kraft's inequality.

Lemma 9.2 (Kac)**.** *Consider a finite-alphabet stationary ergodic process* $\ldots, X_{-1}, X_0, X_1 \ldots$. *Let* $L = \inf\{t > 0 : X_{-t} = X_0\}$ *be the last appearance of symbol* X_0 *in the sequence* $X_{-\infty}^{-1}$. *Then for any* u *such that* $\mathbb{P}[X_0 = u] > 0$ *we have*

$$\mathbb{E}[L|X_0 = u] = \frac{1}{\mathbb{P}[X_0 = u]}\,.$$

In particular, mean recurrence time $\mathbb{E}[L] = |\mathrm{supp} P_X|$.

Proof. Note that from stationarity the following probability

$$\mathbb{P}[\exists t \geq k : X_t = u]$$

does not depend on $k \in \mathbb{Z}$. Thus by continuity of probability we can take $k = -\infty$ to get

$$\mathbb{P}[\exists t \geq 0 : X_t = u] = \mathbb{P}[\exists t \in \mathbb{Z} : X_t = u]\,.$$

However, the last event is shift-invariant and thus must have probability zero or one by ergodic assumption. But since $\mathbb{P}[X_0 = u] > 0$ it cannot be zero. So we conclude

$$\mathbb{P}[\exists t \geq 0 : X_t = u] = 1\,. \tag{9.23}$$

Next, we have

$$\begin{aligned}
\mathbb{E}[L|X_0 = u] &= \sum_{t\geq 1} \mathbb{P}[L \geq t|X_0 = u] && (9.24)\\
&= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t\geq 1} \mathbb{P}[L \geq t, X_0 = u] && (9.25)\\
&= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t\geq 1} \mathbb{P}[X_{-t+1} \neq u, \ldots, X_{-1} \neq u, X_0 = u] && (9.26)\\
&= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t\geq 1} \mathbb{P}[X_0 \neq u, \ldots, Xt-2 \neq u, X_{t-1} = u] && (9.27)\\
&= \frac{1}{\mathbb{P}[X_0 = u]} \mathbb{P}[\exists t \geq 0 : X_t = u] && (9.28)\\
&= \frac{1}{\mathbb{P}[X_0 = u]}\,, && (9.29)
\end{aligned}$$

where (9.24) is the standard expression for the expectation of a $\mathbb{Z}_+$-valued random variable, (9.27) is from stationarity, (9.28) is because the events corresponding to different t are disjoint, and (9.29) is from (9.23). □

The following proposition serves to explain the basic principle behind operation of Lempel-Ziv:

Theorem 9.1. *Consider a finite-alphabet stationary ergodic process* $\ldots, X_{-1}, X_0, X_1 \ldots$ *with entropy rate* H. *Suppose that* $X_{-\infty}^{-1}$ *is known to the decoder. Then there exists a sequence of prefix-codes* $f_n(x_0^{n-1}, x_{-\infty}^{-1})$ *with expected length*

$$\frac{1}{n}\,\mathbb{E}[\ell(f_n(X_0^{n-1}, X_{\infty}^{-1}))] \to H\,,$$

Proof. Let L_n be the last occurence of the block x_0^{n-1} in the string $x_{-\infty}^{-1}$ (recall that the latter is known to decoder), namely

$$L_n = \inf\{t > 0 : x_{-t}^{-t+n-1} = x_0^{n-1}\}\,.$$

Then, by Kac's lemma applied to the process $Y_t^{(n)} = X_t^{t+n-1}$ we have

$$\mathbb{E}[L_n | X_0^{n-1} = x_0^{n-1}] = \frac{1}{\mathbb{P}[X_0^{n-1} = x_0^{n-1}]}\,.$$

We know encode L_n using the code (9.22). Note that there is crucial subtlety: even if $L_n < n$ and thus $[-t, -t+n-1]$ and $[0, n-1]$ overlap, the substring x_0^{n-1} can be decoded from the knowledge of L_n.

We have, by applying Jensen's inequality twice and noticing that $\frac{1}{n}H(X_0^{n-1}) \searrow H$ and $\frac{1}{n}\log H(X_0^{n-1}) \to 0$ that

$$\frac{1}{n}\,\mathbb{E}[\ell(f_{int}(L_n))] \le \frac{1}{n}\,\mathbb{E}[\log \frac{1}{P_{X_0^{n-1}}(X_0^{n-1})}] + o(1) \to H\,.$$

From Kraft's inequality we know that for any prefix code we must have

$$\frac{1}{n}\,\mathbb{E}[\ell(f_{int}(L_n))] \ge \frac{1}{n}H(X_0^{n-1}|X_{-\infty}^{-1}) = H\,.$$

□

6.441 Information Theory

Spring 2016