Remarks on universal coding and channel capacity

Here the signals are coming from an unknown process $\mathbb{P}$, we only know that it belongs to the class $\mathcal{P}$ (for example i.i.d., stationary, or Markov). The alphabet A of the process is finite, $|A| = k$, and the marginal distribution $\mathbb{P}_n$ of $\mathbb{P}$ is defined by

$$\mathbb{P}_n(a_1^n) = \mathbb{P}(a_1^n), \quad a_1^n \in A^n.$$

With binary coding, the ideal codelength of a message $x_1^n \in A^n$ coming from a process $\mathbb{P}$ is $-\log \mathbb{P}(x_1^n)$ (see Shannon code). But we do not know $\mathbb{P}$, therefore the actual code is selected according to some distribution $\mathbb{Q}$, and the actual codelenght, $L(x_1^n)$ is $-\log \mathbb{Q}(x_1^n)$. The difference between the two is the *redundancy*:

$$R(x_1^n) = L(x_1^n) - (-\log \mathbb{P}(x_1^n)) = \log \frac{\mathbb{P}(x_1^n)}{\mathbb{Q}(x_1^n)}.$$

In fact, we should adapt $\mathbb{Q}_n$ to $\mathbb{P}_n$, based on our knowledge of the class $\mathcal{P}$, so we want to find a universally "good" code $\mathbb{Q}$ for any $\mathbb{P} \in \mathcal{P}$, e.g., by arithmetic coding. While doing this, in some sense, we want to minimize the redundancy. For this purpose the following two criteria are introduced (they refer to the worst case selection of $\mathbb{P}$, but to the optimal selection of $\mathbb{Q}$):
Expected (or average) redundancy:

$$\bar{R}_n = \min_{Q_n} \sup_{\mathbb{P} \in \mathcal{P}} \sum_{\mathbf{x}_1^n \in A^n} \mathbb{P}(x_1^n) \log \frac{\mathbb{P}(x_1^n)}{\mathbb{Q}(x_1^n)} = \min_{Q_n} \sup_{\mathbb{P} \in \mathcal{P}} D(\mathbb{P}_n \| Q_n). \tag{1}$$

Maximal redundancy:

$$R_n^* = \min_{Q_n} \sup_{\mathbb{P} \in \mathcal{P}} \max_{\mathbf{x}_1^n \in A^n} \log \frac{\mathbb{P}(x_1^n)}{\mathbb{Q}(x_1^n)}. \tag{2}$$

Of course, $\bar{R}_n \leq R_n^*$, and for i.i.d. processes with $|A| = k$, Theorem 7.5 states:

$$\frac{k-1}{2} \log n - K_1 \leq \bar{R}_n \leq R_n^* \leq \frac{k-1}{2} \log n + K_2$$

where K_1, K_2 are positive constants. Among the prefix codes, the arithmetic code, defined by the coding process $\mathbb{Q}$ of (6.5), is the best possible. Similar statement holds for m-th order Markov chains, see the second part of Theorem 7.5.

As for R_n^*, in Chapter 6 (pp. 479-480) it is proved that

$$R_n^* = \log \sum_{x_1^n \in A^n} \mathbb{P}_{ML}(x_1^n)$$

and the minimum in (2) is attained at $\mathbb{Q}_n = NML_n$. However, these are usually not marginals of a process, so they are not of practical use.

Luckily, the expected redundancy is closely related to the channel capacity, therefore, it is more suitable to construct universal codes.

First consider a parametric family of processes: $\Pi = \{\mathbb{P}_\theta : \theta \in \Theta\}$, where Θ is finite dimensional parameter space. Sometimes, Θ is called input, and A is called output alphabet.

In a Bayesian setup, θ is selected according to a probability measure ν on Θ (in other words, ν is the prior distribution of θ). Then $\mathbb{Q}_\nu$ is constructed as a mixture of the $\mathbb{P}_\theta$ distributions with respect to ν, by the Bayes rule: $\mathbb{Q}_\nu = \int_\Theta \mathbb{P}_\theta \nu(d\theta)$.

The mutual information between the input ($\mathbb{P}_\nu$) and output ($\mathbb{Q}_\nu$) distributions is

$$I(\nu) = I(\mathbb{Q}_\nu \wedge \mathbb{P}_\nu) = H(\mathbb{Q}_\nu) - H(\mathbb{Q}_\nu | \mathbb{P}_\nu).$$

The *channel capacity* is defined as

$$\sup_\nu I(\nu) = I(\nu_0), \tag{3}$$

where ν_0 is called capacity achieving distribution.

It will be more convenient to consider the following parametrization. For any θ, $\mathbb{P}_\theta$ is a distribution on the same finite alphabet $A = \{a_1, \dots, a_k\}$. Therefore, it can be identified with the point $(P(a_1), \dots, P(a_k))$ of nonnegative coordinates, summing to 1 (these points are in a $(k-1)$-dimensional hyperplane within the positive orthant of $\mathbb{R}^k$). It makes sense to consider the natural parametrization $\Theta = \Pi$, and accordingly, instead of the measure ν on Θ, we consider the measure μ on Π. We also use $I(\mu)$ instead of $I(\nu)$.

Lemma 7.2. states that for any closed set Π of distributions on A, there exists a probability measure μ_0 that maximizes $I(\mu)$. This μ_0 is a finite mixture, that is concentrated on a finite subset of Π of size less than $|A|$. Further, if Π is a closed subset of parametric distributions, then this max is in fact a sup in (3), so here exists a capacity achieving distribution ν_0 that is concentrated on a finite subset of Θ of cardinality at most $|A|$.

To relate $\bar{R}_n$ of (1) to the channel capacity, we introduce the following notion: the *I-radius* of Π is

$$\min_{\mathbb{Q}} \sup_{\mathbb{P} \in \Pi} D(\mathbb{P} \| \mathbb{Q}) = \sup_{\mathbb{P} \in \Pi} D(\mathbb{P} \| \mathbb{Q}^*).$$

The *I-centroid* $\mathbb{Q}^*$ is unique if all the distributions are defined on the same finite alphabet A.

Theorem 7.2 states that for any parametric set of distributions, the I-radius is equal to the channel capacity:

$$I(\nu_0) = \sup_{\theta \in \Theta} D(\mathbb{P}_\theta \| \mathbb{Q}^*).$$

Applying the above results to A^n, we can get lower bounds for $\bar{R}_n$ via lower bounding $I(\nu)$. Theorems 7.4 and 7.5 are such. (Note that there not necessarily exists a process such that for all n the $\mathbb{Q}^*$s are marginals of it.)