Universal Coding via Mixtures

The last theorem we discussed indicates that one way to construct a code whose normalized redundancy with respect to the underlying distribution $P$ goes to zero (without knowledge of $P$) is to use a mixture $Q$ of measures, some of which are guaranteed to be close to the unknown underlying measure $P$. This may sound like a convoluted argument at first glance: how is one to guarantee that members of a family $\mathcal{U}$ are arbitrarily close to $P$ without knowing $P$? Fortunately, this is possible for a large class of measures $P$. We shall now discuss the special case of when $P$ is any iid measure on sequences from a finite alphabet $\mathcal{X} = \{1, \ldots, d\}$.

If $P$ is an iid measure on $\mathcal{X}^\infty$, then $P$ is completely specified by the (marginal) probabilities $P(X_1 = i) = P_i$, $i \in \mathcal{X}$. For a sequence $x^n_1$,

$$P(x_1^n) = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{d} (P_i)^{n_{x_1^n}(i)},$$

where $n_{x_1^n}(i)$ is the number of times the letter $i$ appears in the sequence $x_1^n$. It is easy to show that among all iid measures, the measure which maximizes this likelihood of a sequence $x^n_1$ is given by

$$P_{ML}(x_1^n) = \max_P P(x^n_1) = \prod_{i=1}^{d} \left( \frac{n_{x_1^n}(i)}{n} \right)^{n_{x_1^n}(i)}.$$

Now, if we encode the sequence $x^n_1$ using a Shannon code with respect to a measure $Q$ on $\mathcal{X}^n$, then the redundancy of the code w.r.t. an underlying measure $P$ is bounded (within one bit) as

$$R(x^n_1) = \log \frac{P(x^n_1)}{Q(x^n_1)} \leq \log \frac{P_{ML}(x^n_1)}{Q(x^n_1)},$$

We will use this bound in the proof of our first result below.
Let $\mathcal{U}$ be the set of all iid measures and let $\nu$ be defined via the Dirichlet density on $\mathbb{P}^d$ with parameters $\alpha = (\alpha_1, \ldots, \alpha_d)$, i.e., we identify each iid measure $P$ on $\mathcal{X}^\infty$, specified by the probabilities $P_i$, $i \in \mathcal{X}$, with the corresponding point on the unit simplex $\mathbb{P}^d \subset \mathbb{R}^d$, and the measure $\nu$ on $\mathcal{U}$ is constructed using the density

$$
\nu_\alpha(P) = \frac{\Gamma(\alpha + d)}{\prod_{i=1}^d \Gamma(\alpha_i + 1)} \prod_{i=1}^d P_i^{\alpha_i}.
$$

where $\alpha = \sum_{i=1}^d \alpha_i$, and $\alpha_i > -1$ for each $i \in \mathcal{X}$. Using this density, the mixture $Q$ may be computed as

\begin{align*}
Q(x_1^n) &= \int_{\mathcal{U}} P(x_1^n) \nu(dP) \\
&= \int_{\mathbb{P}^d} \prod_{i=1}^d P_i^{n_{x_1^n}(i)} \nu_\alpha(P) dP \\
&= \int_{\mathbb{P}^d} \prod_{i=1}^d P_i^{\alpha_i + n_{x_1^n}(i)} \frac{\Gamma(\alpha + d)}{\prod_{i=1}^d \Gamma(\alpha_i + 1)} dP \\
&= \frac{\prod_{i=1}^d \Gamma(\alpha_i + n_{x_1^n}(i) + 1)}{\prod_{i=1}^d \Gamma(\alpha_i + 1)} \frac{\Gamma(\alpha + d)}{\prod_{i=1}^d \Gamma(\alpha_i + 1)}
\end{align*}

Now, recall that $\Gamma(x + 1) = x \Gamma(x)$, which may be used to rewrite the mixture probability as

\begin{align*}
Q(x_1^n) &= \frac{\Gamma(\alpha + d)}{\Gamma(\alpha + n + d)} \prod_{i=1}^d \frac{\Gamma(\alpha_i + n_{x_1^n}(i) + 1)}{\Gamma(\alpha_i + 1)} \\
&= \frac{\Gamma(\alpha + d)}{\Gamma(\alpha + n + d)} \prod_{i=1}^d (n_{x_1^n}(i) + \alpha_i)(n_{x_1^n}(i) - 1 + \alpha_i) \cdots (1 + \alpha_i) \Gamma(1 + \alpha_i) \\
&= \frac{\prod_{i=1}^d (n_{x_1^n}(i) + \alpha_i)(n_{x_1^n}(i) - 1 + \alpha_i) \cdots (1 + \alpha_i)}{(n - 1 + \alpha + d)(n - 2 + \alpha + d) \cdots (\alpha + d)} \\
&= \prod_{t=1}^n \frac{n_{x_1^{t-1}}(x_t) + 1 + \alpha_{xt}}{t - 1 + \alpha + d} \\
&= \prod_{t=1}^n Q(x_t|x_1^{t-1}),
\end{align*}

where, as before, $n_{x_1^{t-1}}(x_t)$ is the number of times the $t$-th symbol has occurred in the “past”, $x_1^{t-1}$, and the term

$$
Q(i|x_1^{t-1}) = \frac{n_{x_1^{t-1}}(i) + 1 + \alpha_i}{t - 1 + \alpha + d}, \quad i \in \mathcal{X},
$$

should be viewed as a smoothed version of the ML-probability $\frac{n_{x_1^{t-1}}(i)}{t-1}$ of the symbol $i$ based on $x_1^{t-1}$, where $1 + \alpha_i$ is added to the count $n_{x_1^{t-1}}(i)$ of the symbol $i$ to make sure that no symbol is assigned a zero probability.
**Theorem:** If $Q$ is the mixture of all iid measures with the Dirichlet prior as described above, with $\alpha_1 = \cdots = \alpha_d = -\frac{1}{2}$, then the pointwise redundancy of the Shannon code with respect to $Q$ for every $x_1^n$, with respect to every iid measure $P$, satisfies

$$R(x_1^n) \leq \log \frac{\Gamma \left( n + \frac{d}{2} \right) \Gamma \left( \frac{1}{2} \right)}{\Gamma \left( n + \frac{d}{2} \right) \Gamma \left( \frac{d}{2} \right)} \leq \frac{d - 1}{2} \log n - \log \frac{\Gamma \left( \frac{d}{2} \right)}{\Gamma \left( \frac{1}{2} \right)} + \epsilon_n,$$

where $\epsilon_n \to 0$ as $n \to \infty$.

**Proof:** For $\alpha_1 = \cdots = \alpha_d = -\frac{1}{2}$, we get $\alpha = -\frac{d}{2}$ and the expression for $Q$ in (2) becomes

$$Q(x_1^n) = \frac{\Gamma \left( \frac{d}{2} \right)}{\Gamma \left( n + \frac{d}{2} \right)} \prod_{i=1}^d \frac{\Gamma \left( n_x(i) + \frac{1}{2} \right)}{\Gamma \left( \frac{1}{2} \right)},$$

which may be further specialized for the case when $x_1^n$ is a sequence of identical symbols, say the all 1 sequence, to get

$$Q(1, 1, \ldots, 1) = \frac{\Gamma \left( \frac{d}{2} \right)}{\Gamma \left( n + \frac{d}{2} \right)} \frac{\Gamma \left( n + \frac{1}{2} \right)}{\Gamma \left( \frac{1}{2} \right)}.$$

Observe that

$$\log \frac{1}{Q(1, 1, \ldots, 1)} = \log \frac{\Gamma \left( n + \frac{d}{2} \right) \Gamma \left( \frac{1}{2} \right)}{\Gamma \left( n + \frac{1}{2} \right) \Gamma \left( \frac{d}{2} \right)},$$

and we may therefore prove the first inequality by equivalently showing that

$$R(x_1^n) \leq \log \frac{1}{Q(1, 1, \ldots, 1)},$$

for every $x_1^n$ and $P$. From (1), therefore, it suffices to show for every $x_1^n$ and $P$ that

$$\log \frac{P_{ML}(x_1^n)}{Q(x_1^n)} \leq \log \frac{1}{Q(1, 1, \ldots, 1)}, \quad \text{or} \quad P_{ML}(x_1^n) \leq \frac{Q(x_1^n)}{Q(1, 1, \ldots, 1)}.$$

If we specialize the expression for $Q$ in (3) to $\alpha_1 = \cdots = \alpha_d = -\frac{1}{2}$, we get

$$Q(x_1^n) = \frac{\prod_{i=1}^d (n_i - \frac{1}{2})(n_i - \frac{3}{2}) \cdots (\frac{1}{2})}{(n - 1 + \frac{d}{2})(n - 2 + \frac{d}{2}) \cdots (\frac{d}{2})},$$

where $n_i = n_x(i)$, and for the all 1 sequence we get

$$Q(1, 1, \ldots, 1) = \frac{(n - \frac{1}{2})(n - \frac{3}{2}) \cdots (\frac{1}{2})}{(n - 1 + \frac{d}{2})(n - 2 + \frac{d}{2}) \cdots (\frac{d}{2})}.$$
The first assertion of the theorem therefore follows if we show that
\[
\prod_{i=1}^{d} \left( \frac{n_i}{n} \right)^{n_i} \leq \prod_{i=1}^{d} \frac{(n_i - \frac{1}{2})(n_i - \frac{3}{2}) \cdots (\frac{1}{2})}{(n - \frac{1}{2})(n - \frac{3}{2}) \cdots (\frac{1}{2})}.
\]

At this point, note that
\[
\left( n - \frac{1}{2} \right) \left( n - \frac{3}{2} \right) \cdots \left( \frac{1}{2} \right) = \frac{n(n - \frac{1}{2})(n - \frac{3}{2}) \cdots (1)(\frac{1}{2})}{n!} = \frac{2n(2n - 1)(2n - 3) \cdots (2)(1)}{2n!} = \frac{2n(2n - 1) \cdots (n + 1)}{2^n},
\]
and it therefore suffices to show that
\[
\prod_{i=1}^{d} \left( \frac{n_i}{n} \right)^{n_i} \leq \prod_{i=1}^{d} \frac{2n_i(2n_i - 1) \cdots (n_i + 1)}{2n(2n - 1) \cdots (n + 1)} \frac{2^{2n}}{2^{2n_i}} = \frac{\prod_{i=1}^{d} 2n_i(2n_i - 1) \cdots (n_i + 1)}{2n(2n - 1) \cdots (n + 1)} \cdot \frac{2^{2n_i}}{2^n} = \prod_{i=1}^{d} \frac{(n_i + n_i)(n_i + (n_i - 1)) \cdots (n_i + 1)}{(n + n)(n + (n - 1)) \cdots (n + 1)}.
\]

Now, observe that right side of the inequality above is the product of \( n \) terms of the kind \( \frac{n_i + j_i}{n + t} \), \( t = 1, \ldots, n \), where \( t \to (i, j_i) \), \( 1 \leq i \leq d \) and \( 1 \leq j_i \leq n_i \), in such a manner that **no two \( t \)'s are mapped to the same pair \( (i, j) \).** The left side is a product of terms \( \frac{n_i}{n} \) for the corresponding indices \( i \). We shall now show that for every \( t \) it is possible to find a pair \( (i_t, j_t) \) in such a manner that the validity of the one-to-one relationship between the \( t \)'s and the \( (i, j) \)'s is maintained, and additionally
\[
\frac{n_i}{n} \leq \frac{n_i + j_i}{n + t}.
\]
(4)

This will prove the first assertion of the theorem. The inequality in (4) is satisfied if, for a given \( t \) and \( i \), we choose \( j \) such that
\[
\frac{n_i}{n} \leq \frac{n_i + j}{n + t} \quad \Rightarrow \quad \frac{t}{n} n_i \leq j \leq n_i.
\]

Thus, for a fixed \( t \) and \( i \), the number of acceptable \( j \)'s, disregarding the one-to-one mapping requirement, is **strictly greater than** \( n_i - \frac{t}{n} n_i \). Thus for a fixed \( t \), the number of available choices of \((i, j)\)-pairs is strictly greater than \( \sum_{i \in X} n_i - \frac{t}{n} n_i = n - t \). Therefore, if we start with \( t = n \), we have at least one pair which satisfies the inequality (4). Pick any such pair and declare it to be \((i_n, j_n)\). Then proceed to \( t = n - 1 \) etc., but select from the pairs which satisfy (4), a pair which has not already been selected. At a stage \( t = n - l \), the number
of available choices (without constraints) is strictly greater than \( n - (n - l) = l \), of which at most \( l \) have been already selected. Therefore we can always pick a previously unselected pair to satisfy (4)! This shows that our procedure never gets stuck, and completes the proof of the first inequality in the assertion of the theorem.

The second inequality follows from Sterling’s formula for the Gamma function.  

**Remark:** We therefore conclude that there exist codes for which the pointwise redundancy with respect to every iid measure is no greater than \( \sim \frac{d - 1}{2} \log n \). A natural question to ask is if we can do any better than this. It turns out, as we shall see shortly, that even if our objective is to ensure that only the *expected* normalized redundancy with respect to a “significant” number of iid measures goes to zero\(^1\), then (with \( P \)-probability 1) we cannot make the pointwise redundancy grow any slower than \( \sim \frac{d - 1}{2} \log n \). Thus *any universal code is unlikely to have a pointwise redundancy which is lower than that of the Shannon code for \( Q \) as described above.

As an intermediate step in proving this claim, we first establish a technical result which, as an aside, provides an information geometric view of channel capacity.

Let \( \Theta \) be the input alphabet for a channel, \( \mathcal{Y} \) the output alphabet, and for every \( \theta \in \Theta \) let the channel transition probabilities be given by a probability measure \( P_\theta \) on \( \mathcal{Y} \). If the input \( Z \) to the channel is randomly chosen according to a distribution \( \nu \) on \( \Theta \), the output \( Y \) of the channel has a distribution

\[
Q_\nu = \int_\Theta P_\theta \nu(d\theta),
\]

and the mutual information between the input and the output random variables of this channel is given by

\[
I(\nu) = H(Y) - H(Y|Z).
\]

The information capacity of this channel is

\[
C = \sup_\nu I(\nu).
\]

**Lemma:** If \( Q \) denotes a generic probability measure on \( \mathcal{Y} \), then

\[
\inf_Q \sup_{\theta \in \Theta} D(P_\theta \parallel Q) = \sup_{\nu} I(\nu) = C.
\]

Thus the capacity of a channel is the “radius” \( D(P \parallel Q) \) of the smallest I-divergence ball which “covers” all the distributions on \( \mathcal{Y} \) in the set of conditional output distributions which describe the channel.

\(^1\)Such codes are said to be universal with respect to expected redundancy over this set of iid measures.
Proof: We shall prove this only for the case when the channel input alphabet \( \Theta \) and output alphabet \( \mathcal{Y} \) are discrete and finite sets and the extrema are achieved. The more general proof is similar in spirit, though a little more technical. Let \( \Theta = \{1, \ldots, l\} \) and \( \mathcal{Y} = \{1, \ldots, m\} \). Then

\[
Q_\pi(j) = \sum_{i=1}^{l} \pi(i)P_i(j), \quad j \in \mathcal{Y}, \text{ and}
\]

\[
I(\pi) = \sum_{i=1}^{l} \sum_{j=1}^{m} \pi(i)P_i(j) \log \frac{P_i(j)}{Q_\pi(j)}
\]

\[
= \sum_{i=1}^{l} \sum_{j=1}^{m} \pi(i)P_i(j) \left[ \log \frac{P_i(j)}{Q(j)} + \log \frac{Q(j)}{Q_\pi(j)} \right]
\]

\[
= \sum_{i=1}^{l} \pi(i)D(P_i||Q) + \sum_{j=1}^{m} \sum_{i=1}^{l} \pi(i)P_i(j) \log \frac{Q(j)}{Q_\pi(j)}
\]

and thus

\[
I(\pi) + D(Q_\pi||Q) = \sum_{i=1}^{l} \pi(i)D(P_i||Q).
\]

If we minimize both sides with respect to \( Q \), we get that

\[
I(\pi) = \min_Q \sum_{i=1}^{l} \pi(i)D(P_i||Q),
\]

and therefore

\[
C = \max_{\pi} I(\pi) = \max_{\pi} \min_Q \sum_{i=1}^{l} \pi(i)D(P_i||Q).
\]

Now, \( \sum_{i=1}^{l} \pi(i)D(P_i||Q) \) is a linear (and hence concave) function of \( \pi \) for a fixed \( Q \), and for a fixed \( \pi \), since each \( D(P_i||Q) \) is a convex function of \( Q \), \( \sum_{i=1}^{l} \pi(i)D(P_i||Q) \) is a convex function of \( Q \). It is a continuous function of the variables \( \pi \) and \( Q \), which are allowed to vary on compact sets (a simplex in the appropriate space). Therefore the minimax theorem applies, and we can interchange the order of evaluation of the minimum and maximum and write

\[
C = \max_{\pi} I(\pi) = \min_Q \max_{\pi} \sum_{i=1}^{l} \pi(i)D(P_i||Q).
\]

Note that the maximum on the right is obtained by putting all the mass of \( \pi \) on the \( i \) which maximizes the I-divergence, and thus

\[
C = \max_{\pi} I(\pi) = \min_Q \max_{i \in \Theta} D(P_i||Q).
\]

The assertion of the theorem for the general \( \Theta \) and \( \mathcal{Y} \) follows similarly.