Lower Bounds on the Redundancy of Universal Codes

We have seen that even when the measure generating a process is not known, we can achieve coding redundancies of the order of \( \log n \) for every iid measure. The normalized redundancy, \( \frac{1}{n} R_P(x_1^n) \) therefore goes to zero for every sequence \( x_1^n \), with respect to every iid measure \( P \). The next result shows that if a code is to have small redundancy over an entire class of measures, its expected redundancy with respect to most measures in the class can grow no slower than \( \log n \).

**Rissanen’s Theorem:** Let \( \{ P_\theta, \ \theta \in \Theta \} \) be any\(^1\) family of measures on \( \mathcal{X}^\infty \), where \( \Theta \subset \mathbb{R}^k \). Let \( p_\theta(\cdot) \) denote the \( n \)-dimensional probability density or mass function on \( \mathcal{X}^n \) corresponding to \( P_\theta \) for every \( \theta \in \Theta \). Suppose that for every \( n > n_0 \) there exists an estimator \( \hat{\theta}_n : \mathcal{X}^n \to \mathbb{R}^k \) with

\[
E_\theta \left[ \left\| \hat{\theta}_n(X_1^n) - \theta \right\|^2 \right] \leq \frac{c(\theta)}{n}, \quad \text{for every } \theta \in \Theta,
\]

for some \( \mathbb{R} \)-valued, Lebesgue measurable function \( c(\cdot) \). Then, for every \( \epsilon > 0 \) there exists a constant \( A > 0 \) such that for \( n > n_0 \) we have for every \( n \)-dimensional probability density or mass function \( g \)

\[
E_\theta \left[ \log \frac{p_\theta(X_1^n)}{g(X_1^n)} \right] \geq \frac{k}{2} \log n - A,
\]

for every \( \theta \in \Theta \), except possibly on a set whose Lebesgue measure is less than \( \epsilon \).

**Proof:** For a \( n \)-dimensional probability density or mass function \( g \), let \( \Theta_1 = \Theta_1(n, g, A) \) denote the set for which (2) is violated, and \( \lambda(\Theta_1) \) denote its Lebesgue measure. Clearly,

\(^1\)The measures are not required to necessarily be iid; not even stationary!
for fixed \( n \) and \( g \), \( \Theta_1 \) shrinks as \( A \) is increased, and thus \( \lambda(\Theta_1) \) is a decreasing function of \( A \). The theorem asserts that given \( \epsilon > 0 \), \( A \) can be selected \textit{a priori} in such a manner that no matter what \( n > n_0 \) and \( g \) we choose, \( \lambda(\Theta_1) < \epsilon \). To falsify the assertion we would have to have that for some \( \epsilon \), no matter how large \( A \) is chosen, we could find some \( n > n_0 \) and a \( n \)-dimensional \( g \) for which \( \lambda(\Theta_1) \geq \epsilon \). We shall show that for a suitable choice of \( A \), \( \lambda(\Theta_1) \geq \epsilon \) leads to a contradiction.

This is not satisfactory. The way things are set up, given \( \epsilon \), for every choice of \( A \), it would suffice, for contradicting the assertion, to find an \( n = n_A > n_0 \), and a \( n \)-dimensional probability density or mass function \( g = g_A \) for which \( \lambda(\theta_1(n, A)) \geq \epsilon \). What the current proof does is to show that if \( n \) and \( g \) were indeed found to satisfy \( \lambda(\theta_1(n, A)) \geq \epsilon \), then for another choice \( A' \), \( \lambda(\theta_1(n, A', A')) < \epsilon \). This is clearly not enough, because for \( A' \) we may be able to find some other \( n_{A'} \) and \( g_{A'} \) which offers a contradiction. Rissanen's original result does not assert an \( A \) independent of \( n \). In fact, his lower bound is with an \( A \) which also grows as \( \epsilon \log n \).

Assume that for any choice of \( A \) we can find \( n \) and \( g \) such that \( \lambda(\Theta_1) \geq \epsilon \). First observe that since every \( \theta \in \Theta_1 \) satisfies (1), we can choose a sufficiently large constant \( c \) such that the subset of \( \Theta_1 \) given by

\[
\Theta_2 = \Theta_2(n, g, A, c) = \{ \theta \in \Theta_1 : c(\theta) < c \}
\]

has Lebesgue measure \( \lambda(\Theta_2) \geq \frac{\epsilon}{2} \). Let \( \nu \) be a probability measure which assigns all its mass mass uniformly on \( \Theta_2 \). Let \( Z \) denote a random variable drawn from \( \Theta_2 \) according to \( \nu \), and given \( Z = \theta \), let \( X^n_1 \) be drawn according to \( p_0 \). Let \( \hat{Z} = \hat{\theta}_n(X^n_1) \) be the estimate of \( Z \) given by an estimator which satisfies (1). The mutual information between \( Z \) and \( X^n_1 \) and between \( Z \) and \( \hat{Z} \) is related, by the data processing inequality, as

\[
I(\nu) = I(Z \wedge X^n_1) \geq I(Z \wedge \hat{Z}).
\]

However, we can write the second mutual information term as

\[
I(Z \wedge \hat{Z}) = h(Z) - h(Z|\hat{Z}) \\
= h(Z) - h(Z - \hat{Z}|\hat{Z}) \\
\geq h(Z) - h(Z - \hat{Z}),
\]

where \( h(\cdot) \) denotes the differential entropy of a continuous-valued random variable. Now, \( Z - \hat{Z} \) is a \( \mathbb{R}^k \)-valued random variable whose second moment is bounded by \( \frac{c}{n} \), for every \( n > n_0 \). Therefore its differential entropy is no greater than that of a \( k \)-variate Gaussian distribution with independent components, each having zero mean and variance \( \frac{c}{nk} \). Also, \( Z \) is uniform over a set of Lebesgue measure \( \frac{c}{2} \). Therefore,

\[
I(\nu) \geq I(Z \wedge \hat{Z}) \geq \log \frac{\epsilon}{2} - \frac{k}{2} \log 2 \pi e \frac{c}{nk} = \frac{k}{2} \log n - \log \frac{4\pi e c}{\epsilon^k}. \tag{3}
\]

\(^2\)For definitions, and expressions for the differential entropy of some continuous random variables, see, e.g., T. M. Cover and J. A. Thomas, Elements of Information Theory, Chapt. 9, John Wiley & Sons, 1991.
Now, since (2) is violated for all \( \theta \in \Theta_2 \), i.e., for some \( n > n_0 \) and for some probability density or mass function \( g \) on \( \mathcal{X}^n \),

\[
E_\theta \left[ \log \frac{p_\theta(X^n_1)}{g(X^n_1)} \right] < \frac{k}{2} \log n - A, \quad \theta \in \Theta_2,
\]

it follows that

\[
\sup_{\theta \in \Theta_2} E_\theta \left[ \log \frac{p_\theta(X^n_1)}{g(X^n_1)} \right] \leq \frac{k}{2} \log n - A,
\]

for some \( n > n_0 \). Therefore

\[
\inf \sup_{g, \theta \in \Theta_2} E_\theta \left[ \log \frac{p_\theta(X^n_1)}{g(X^n_1)} \right] \leq \frac{k}{2} \log n - A, \quad \text{for some } n > n_0.
\]

However, we know that \( E_\theta \left[ \log \frac{p_\theta(X^n_1)}{g(X^n_1)} \right] = D(p_\theta||g) \), and the result of the previous lemma shows that

\[
\inf \sup_{g, \theta \in \Theta_2} D(p_\theta||g) = \sup_\nu I(\nu).
\]

Finally, as shown by a particular choice of \( \nu \) in the derivation of (3), we have for every \( n > n_0 \), that

\[
\sup_\nu I(\nu) \geq \frac{k}{2} \log n - \log \frac{4\pi ec}{\epsilon k} > \frac{k}{2} \log n - 2 \max \left\{ 0, \log \frac{4\pi ec}{\epsilon k} \right\}.
\]

Therefore, we have that no matter how we choose \( A \), we can find an \( n > n_0 \) such that

\[
\frac{k}{2} \log n - 2 \max \left\{ 0, \log \frac{4\pi ec}{\epsilon k} \right\} < \frac{k}{2} \log n - A,
\]

which is a contradiction! Therefore \( \Theta_1 \) cannot have Lebesgue measure greater than \( \epsilon \), and the assertion of the theorem is proved.

The proof is not quite done, because \( \epsilon \) is chosen to make \( \lambda(\Theta_2) \geq \epsilon \), and therefore \( \epsilon \) is a function of \( A \). If \( \epsilon \) grows sufficiently fast with \( A \), we do not get the contradiction we are seeking. Again, Rismanen's original result seems to suggest that \( \epsilon \) grows slower than \( n \), and thus choosing

\( A = \epsilon \log n \) provides the contradiction.

**Corollary:** If, for some subfamily \( \{P_\theta, \theta \in \Theta_0\} \) of a family of sources satisfying (1), there exist universal codes whose average redundancy grows slower than \( \frac{k}{2} \log n \), i.e.,

\[
\lim_{n \to \infty} E_\theta \left[ R(X^n_1) \right] - \frac{k}{2} \log n = -\infty, \quad \theta \in \Theta_0,
\]

then, necessarily, \( \lambda(\Theta_0) = 0 \).

**Proof:** Without loss of generality, we can assume that the universal code is a Shannon code with respect to some \( g \). Then, if \( \lambda(\Theta_0) \) were positive, then for sufficiently large \( n \), \( E_\theta \left[ \log \frac{p_\theta(X^n_1)}{g(X^n_1)} \right] \) would be smaller than every finite \( -A \), contradicting the assertion in (2).