1 Review - Coding schemes for compression

- Huffman Code (optimal prefix-free code)
- Shannon-Fano code
- Shannon-Fano-Elias code
- Arithmetic code (can handle a sequence of symbols)

In general, the first three codes do not achieve the optimal rate $H(X)$, and there are no immediate extensions of these codes to rate-optimal codes for a sequence of symbols. On the other hand, arithmetic coding is rate-optimal. However, all these schemes assume knowledge of the distribution. We will now study universal source codes.

2 Introduction

In the case of universal source codes, the distribution of the source $P$ is unknown. We look at coding schemes for compression in several regimes:

1. Fixed length codes attaining the optimal error exponent - We have already seen this as an application of the Method of Types.

2. Variable length codes which attain:
   - Minimax optimality over a family of distributions
   - Minimax optimality for worst case individual sequence, when compared with a class of expert algorithms (for instance, all iid distributions over the alphabet)

For the rest of the lecture, $Q_n$ will be an element of $\mathcal{P}(\mathcal{X}^n)$, i.e., a pmf on $\mathcal{X}^n$, $\Theta$ will refer to a family of distributions indexed by the parameter $\theta$, and $x$ will denote an element of $\mathcal{X}^n$, i.e., a sequence of length $n$.

Definition. Average redundancy:

$$\bar{R}(n, \mathcal{X}) := \min_{Q_n} \max_{P \in \mathcal{P}(\mathcal{X})} D(P^n||Q_n)$$
Definition. Let $\mathcal{X}$ be a finite set and $P,Q$ be pmfs on $\mathcal{X}$.

$$D_{\text{max}}(P||Q) := \max_{x \in \mathcal{X}} \log \frac{P(x)}{Q(x)}$$

Definition. Worst-case redundancy:

$$R(n,\mathcal{X}) := \min_{Q_n} \max_{\theta \in \Theta} \max_{x \in \mathcal{X}^n} \log \frac{1}{Q_n(x)} - \frac{1}{P_\theta(x)} = \min_{Q_n} \max_{\theta \in \Theta} D_{\text{max}}(P_\theta||Q_n)$$

We would ideally like to optimize, for a given code $f$, the expression $l(f(x)) - \frac{1}{P_\theta(x)}$. However, restricting ourselves to uniquely decodable codes, it suffices to consider probability distributions $Q_n$ given by $Q_n(x) = 2^{-l(f(x))}$.

Given an encoder $f$, consider

$$\tilde{R}(f, X^n) := \mathbb{E}[l(f(X^n))] - nH(X)$$

Clearly, $\tilde{R}(f, X^n) \geq 0$ if $f$ corresponds to a uniquely decodable code.

Definition. $\tilde{R}(n) := \min_f \tilde{R}(f, X^n)$

Here, the minimum (strictly, infimum) is taken over all $f$ such that $f$ is uniquely decodable.

Note that $\tilde{R}(n)$ is an operational quantity which we would like to study, whereas $\tilde{R}(n, \mathcal{X})$ is easier to handle mathematically. The following claim shows that the two quantities differ by at most 1 bit, and hence for estimating the rate, it suffices to study $\tilde{R}(n, \mathcal{X})$.

Claim. $\tilde{R}(n) - 1 \leq \tilde{R}(n, \mathcal{X}) \leq \tilde{R}(n)$

Proof. Let $f$ be a uniquely decodable code. Then by Kraft’s inequality, $\sum_{x} 2^{-l(x)} \leq 1$. Then the function $q_n$ on $\mathcal{X}^n$ defined by $q_n(x) = \frac{2^{-l(x)}}{\sum_{x} 2^{-l(x)}}$ is a pmf on $\mathcal{X}^n$.

$$D(P^n||q_n) = \sum_{x} P^n(x) \log \frac{P^n(x)}{q_n(x)} = \sum_{x} P^n(x) \log \frac{1}{q_n(x)} - nH(X)$$

$$\mathbb{E}[l(f(X^n))] = \sum_{x} P^n(x) \log \frac{1}{2^{-l(x)}} \geq \sum_{x} P^n(x) \log \frac{\sum_{x} 2^{-l(x)}}{2^{-l(x)}} = \sum_{x} P^n(x) \log \frac{1}{q_x}$$

So, $\tilde{R}(f, X^n) \geq D(P^n||q_n) \geq \tilde{R}(n, \mathcal{X})$. Taking the minimum over all uniquely decodable codes $f$, we get $\tilde{R}(n, \mathcal{X}) \leq \tilde{R}(n)$.

On the other hand, given a pmf $q_n$ on $\mathcal{X}^n$, there exists a uniquely decodable Shannon code of length $l(x) = \lceil - \log q_n(x) \rceil \leq - \log q_n(x) + 1$. Hence, we have $\tilde{R}(n) \leq \tilde{R}(n, \mathcal{X}) + 1$.

Now, we describe two coding schemes. For analysis, we restrict the alphabet to be binary.

### 3 Scheme 1 - Encoding types

The encoder $f$ has two components, i.e. $f = (f_1, f_2)$ where:
- $f_1$ is a non-singular code for the type of $x$
- $f_2$, given the type of $x$, stores $x$ in $\lceil \log |T_{f_1}| \rceil$ bits.
Example. Consider $X = \{0, 1\}$, $x = 0110100$.

$f_1(x)$ stores the number of 1’s in the sequence (3 in this case). This requires $\lceil \log |T| \rceil = \lceil \log 8 \rceil = 3$ bits.

$f_2(x)$ stores the enumeration of $x$ in the type. This requires $\lceil \log \binom{n}{k_x} \rceil$ bits.

Note that as the number of types is polynomial in $n$, the length of $f_1$ is insignificant for (asymptotic) calculation of rate.

Worst-case performance of Scheme 1

For ease of calculation, let $X = \{0, 1\}$ and let $f_1$ be a fixed length code for types (this can be improved on, but the analysis is more involved and this is order-optimal, in the sense that it can only be improved by a constant factor).

$l(x) = \lceil \log(n + 1) \rceil + \lceil \log \binom{n}{k_x} \rceil$

where $k_x = N(1|x)$

Suppose our class of experts consists of all iid distributions, i.e., we compete with

$$\min_{P \in \mathcal{F}(X)} -\log P^n(x) = \min_{P \in \mathcal{F}(X)} -\log \exp[-n(D(P_x||P) + H(P_x))] = \min_{P \in \mathcal{F}(X)} n[D(P_x||P) + H(P_x)] = nH(P_x) = nh \left( \frac{k_x}{n} \right)$$

Thus,

$$l(x) - \min_{P \in \mathcal{F}(X)} -\log P^n(x) \leq \log(n + 1) + \log \left( \frac{n}{k_x} \right) + 2 - nh \left( \frac{k_x}{n} \right) \leq \log(n + 1) + nh \left( \frac{k_x}{n} \right) - nh \left( \frac{k_x}{n} \right) + 2 = \log(n + 1) + 2 = o(n)$$

For the second inequality, we use the fact that $\binom{n}{k_x} \leq 2^{nh(\frac{k_x}{n})}$. This is true for all $x \in X^n$, and hence for the worst-case $x$ too. Thus, we have

$$R(n, \{0, 1\}) \leq \log(n + 1) + 2 = O(\log n)$$

Result: $R(n, \{0, 1\}) \geq \frac{1}{2} \log n + c$ for some constant $c$.

This can be proved by choosing $f_1$ more carefully (we do not do this here).

This result, together with the analysis above, implies that this scheme is order-optimal, for all pmfs $P \in \mathcal{F}(X)$, even in the absence of knowledge of $P$.

Result: $\frac{1}{2} \log n \leq R(n, \{0, 1\}) \leq O(\log n)$

The problem with this scheme is that it is not online - the encoder needs the entire message at once as input. Also, updating the code once a new symbol is not efficient since the entire encoding of types and sequences within the types has to be done all over again. This problem is
4 Scheme 2 - Bayesian Scheme

\[ R(n, \mathcal{X}) = \min_{Q_n} \max_{P \in \mathcal{P}(\mathcal{X})} D(P^n || Q_n) \]
\[ = \min_{Q_n} \max_{\Pi \in \mathcal{P}(\mathcal{X})} \Exp_{P \sim \Pi} D(P^n || Q_n) \]

We have used a technique called convexification for the second equality. The basic idea is the following:

Let \( k \in \mathbb{N}, [k] = \{1, 2, ..., k\} \) and let \( f : [k] \to \mathbb{R} \) be a real-valued function. Then

\[ \max_{i \in [k]} f(i) = \max_{P \in \mathcal{P}([k])} \sum_{i=1}^{k} P(i) f(i) \]

In fact, the maximum is attained for that pmf \( P \) which puts the entire mass on the element \( i \) at which \( f \) attains its maximum.

Since the expression on the RHS is linear in \( P \), it is both convex and concave. Hence, the maximum is attained at a vertex on the boundary of the probability simplex (since the probability simplex is bounded and convex), and so is the minimum.

Now, let \( g(\Pi, Q_n) = \Exp_{P \sim \Pi} D(P^n || Q_n) \).

- \( g(\Pi, Q_n) \) is linear in \( \Pi \).
- \( g(\Pi, Q_n) \) is convex in \( Q_n \), since \( D(P||Q) \) is convex in \( Q_n \). Thus, by Sion’s minimax theorem, the order of min and max can be interchanged, and we have

\[ R(n, \mathcal{X}) = \max_{\Pi} \min_{Q_n} \Exp_{P \sim \Pi} D(P^n || Q_n) \]

\( \Pi \) is known as a prior. It represents some knowledge about the distribution of \( P \). In this case, since we have no information about \( P \), we choose the uniform prior. We will restrict the analysis to the binary alphabet \( \mathcal{X} = \{0, 1\} \).

Choosing a pmf \( P \) on \{0, 1\} uniformly amounts to choosing \( p = P(1) \) uniformly at random from the interval \([0, 1]\). So, we choose \( p \sim Unif[0, 1] \).

Fix some sequence \( x \in \mathcal{X}^n \), and let \( N(1|x) = K \). Then for fixed \( p \), the probability of \( x \) is \( p^K (1-p)^{n-K} \).

As \( p \sim Unif[0, 1] \), the probability of \( x \) is

\[ Q_n(x) = \frac{1}{(n+1) \binom{n}{k}} \]

The above integration can be done by setting \( I(n, k) = \int_0^1 p^k (1-p)^{n-k} dp \) and deriving a recursion for \( I(n, k) \) (using integration by parts) in terms of \( I(n, k-1) \) with the base case \( I(n, 0) = \frac{1}{n+1} \).
Performance

Let $X \sim \text{Bin}(n,p)$. Then $E[X] = np$.

$$D(P^n||Q_n) = \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} \log \frac{p^k(1-p)^{n-k}}{(n+1)\binom{n}{k}}$$

$$= \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} \log(n+1) \binom{n}{k} + E[X] \log(1-p) + \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} \log(1-p)$$

$$\leq nE[h\left(\frac{X}{n}\right)] - nh(p) + \log(n+1)$$

We have used the following facts:
- $\binom{n}{k} \leq 2^n h\left(\frac{k}{n}\right)$, $0 \leq k \leq n$.
- $h$ is concave in $P$, hence $E[h(X)] \leq h(E[X])$.

This is true for all pmfs $P$.

Thus, $\hat{R}(n,\mathcal{X}) = \min_Q \max_{P \in \mathcal{P}(\mathcal{X})} D(P^n||Q_n) \leq \log(n+1)$.

Hence this scheme is order-optimal.

Update Rule

The advantage of this scheme is that it is online, i.e. the entire sequence does not have to be read at once. Moreover, updating is easy, as described below.

$$Q(X_{n+1} = 1|N(1)|X^n) = k) = \frac{Q(N(1)|X^{n+1}) = k + 1}{Q(N(1)|X^n) = k}$$

$$= \frac{\frac{1}{n+1} \frac{1}{(k+1)}}{\frac{1}{n+1} \frac{1}{k}}$$

$$= \frac{k+1}{n+2}$$

$$= \frac{k+1}{(k+1) + (n-k+1)}$$

This is known as the Add-1 estimator, in which the probability of seeing a symbol $x \in \mathcal{X}$ is estimated to be the number of times $x$ has been seen so far, incremented by 1, divided by the frequencies (incremented by 1) of all symbols seen so far. This also takes care of the situation when some symbol is being seen for the first time, since that probability will be non-zero by this update rule. This estimator is also known as the Laplace estimator.

Example. $x = 001$

$Q(0) = \frac{1}{1+1} = \frac{1}{2}$
\[
Q(00) = Q(0) \cdot Q(0|0) = \frac{1}{2} \cdot \frac{1+1}{1+2} = \frac{1}{2} \cdot \frac{2}{3}
\]
\[
Q(001) = Q(00) \cdot Q(1|00) = \frac{1}{2} \cdot \frac{3}{2} \cdot \frac{1+1}{2+2} = \frac{1}{2} \cdot \frac{3}{2} \cdot \frac{1}{3}
\]

**Fact:** A better prior than the uniform prior is known as Jeffrey’s Prior in which the probabilities
\[p(x)\alpha^\frac{1}{\sqrt{x(1-x)}}\] for \(x \in (0, 1)\). It turns out that this gives rise to the Add-\(\frac{1}{2}\) estimator, also known as the Krichevsky–Trofimov estimator, which achieves a rate of \(\frac{1}{2} \log n\) (upto a constant), hence it is optimal even w.r.t the constant.

In fact, any Add-\(\alpha\) estimator, for \(\alpha\) constant, is order-optimal. Also, the Add-\(\alpha\) estimator arises from the Beta\((\alpha, \alpha)\) prior. The Beta priors are known as the conjugate priors for iid distributions.