Shannon's Lower Bound for Vector Quantization

Kevin Liu

May, 4, 2025

1 Introduction

1.1 Rate Distortion Theory

In class, we discussed clustering and embedding at length. With clustering, data points are grouped into clusters, each represented by some principal. This is analogous to quantization in rate-distortion theory, where a continuous signal is approximated by a finite set of reconstruction points. In fact, vector quantization, a practical approach derived from rate-distortion theory, is mathematically equivalent to k-means clustering when distortion is measured by squared error. Both aim to reduce information, or variability, while preserving the most relevant structure of the data under a distortion constraint.

Breaking down the phrase *rate distortion* into its terms, we have

- rate: the amount of information allocated to represent each individual data point
- distortion: an indication of the degree to which the reconstructed output deviates from the original data

In essence, distortion reflects the loss in fidelity when compressing data or how closely the output preserves the original content. This brings us to the central question of rate distortion theory: given a fixed rate R, how low can the distortion be?

We define this problem as follows: We formulate the problem as follows:

Given a signal x, which can take on any value from a continuous range (i.e., the set of real numbers $\mathbb{R}$), we must convert this signal into a discrete form suitable for digital storage or transmission. Since digital systems operate with finite precision, they can represent values using only a limited number of bits. This approximation process is known as *quantization*.

Before we continue, we lay out a few definitions.

1.2 Distortion Functions

We aim to map an input symbol $x \in \mathbb{R}^d$ to a representative output $x^* \in \mathcal{X}^* \subset \mathbb{R}^d$, where $\mathcal{X}^*$ is a finite/countable set. It is necessary to introduce a distortion function which is the cost of representing x as x^*.

We define a distortion function as a mapping:

$$d : \mathcal{X} \times \mathcal{X}^* \mapsto R^{\geq 0}$$

given input alphabet $\mathcal{X}$ and output $\mathcal{X}^*$ which must assign strictly non-negative value to each pair (x, x^*). Ideally, a distortion measure should be easily computable. We are already familiar with Hamming distance and MSE.

$$d_{\text{Hamming}} = \begin{cases} 0 & \text{if } x = x^* \\ 1 & \text{if } x \neq x^* \end{cases} \qquad d_{\text{MSE}} = (x - x^*)^2$$

Another widely used distortion measure, especially in audio and spectral signal processing, is the *Itakura-Saito distance*. This measure is asymmetric and particularly sensitive to perceptual differences in magnitude spectra. It is only defined for positive values $x, x^* > 0$. It is defined as:

$$d_{\text{IS}} = \frac{x}{x^*} - \log\left(\frac{x}{x^*}\right) - 1$$

While selecting an appropriate distortion measure can be difficult and often subjective, we assume that one has been chosen and focus instead on designing systems to minimize average distortion.

Extending this definition of distortion over a sequences of pairs (x, x^*), we have average distortion.

$$d(X^n, X^{*n}) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, x_i^*)$$

Since most lossy compression methods are designed for data intended for human perception like video streaming or MP3's, the distortion introduced by compression is often evaluated using perceptually-informed loss functions. These functions aim to model how humans perceive differences in quality, making them more aligned with subjective experience than purely mathematical metrics. However, for analytical tractability and consistency with traditional signal processing approaches, we will use the Mean Squared Error as our distortion measure.

1.3 Entropy and Information

Entropy is a fundamental concept in information theory that quantifies the uncertainty or randomness of a random variable. For a discrete random variable X with probability mass function $p(x)$, the *entropy* of X is defined as:

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

Entropy measures the average number of bits needed to represent the outcome of X. A higher entropy indicates more uncertainty in the variable. The *conditional entropy* of X given another random variable X^*, denoted $H(X \mid X^*)$, quantifies the remaining uncertainty about X when X^* is known:

$$H(X \mid X^*) = -\sum_{x\in\mathcal{X}}\sum_{x^*\in\mathcal{X}^*} p(x, x^*)\log p(x \mid x^*)$$

Mutual information, denoted $I(X; X^*)$, measures the reduction in uncertainty about X due to knowledge of X^*. It is defined as:

$$I(X; X^*) = \sum_{x\in\mathcal{X}}\sum_{x^*\in\mathcal{X}^*} p(x, x^*)\log\left(\frac{p(x, x^*)}{p(x)p(x^*)}\right)$$

This can also be expressed in terms of entropy:

$$I(X; X^*) = H(X) - H(X \mid X^*) = H(X^*) - H(X^* \mid X)$$

Mutual information is always non-negative and symmetric, i.e., $I(X; X^*) = I(X^*; X)$, and equals zero if and only if X and X^* are independent.

1.4 Rate Distortion Functions

We now define the relationship between rate and distortion more formally.

The *rate-distortion function* $R(D)$ of a source is the infimum of all rates R such that there exists an encoding scheme that achieves an expected distortion no greater than D. Formally,

$$R(D) = \min_{p(x^*|x):\mathbb{E}[d(X,X^*)]\leq D} I(X; X^*)$$

Alternatively, $R(D)$ can be defined as the minimum number of bits per symbol needed to represent the source while ensuring that the expected distortion does not exceed the target level D.

$$R(D) = \inf\left\{R : \exists \text{ a sequence of } (2^{nR}, n) \text{ codes such that } \lim_{n\to\infty}\mathbb{E}[d(X^n, X^{*n})] \leq D\right\}$$

The 2^{nR} comes from a code with block length n and 2^{nR} codewords, implying a rate of R bits per symbol.

1.5 Information Rate Distortion Function

Given a source X and a distortion measure $d(x, x^*)$, the *information rate-distortion function* $R_I(D)$ quantifies the minimum rate required to encode X such that the expected distortion between X and its reconstruction X^* does not exceed a threshold D. It is defined as the solution to the following optimization problem over all conditional distributions $p(x^* \mid x)$ satisfying the distortion constraint:

$$R_I(D) = \min_{p(x^*|x):\sum_{x,x^*} p(x)p(x^*|x)d(x,x^*)\leq D} I(X; X^*)$$

Here, $R_I(D)$ corresponds to the minimum mutual information between the source X and its encoded-decoded version X^* that satisfies the distortion limit.

At the zero-distortion limit $D = 0$, perfect reconstruction is required, and the rate reaches its maximum:

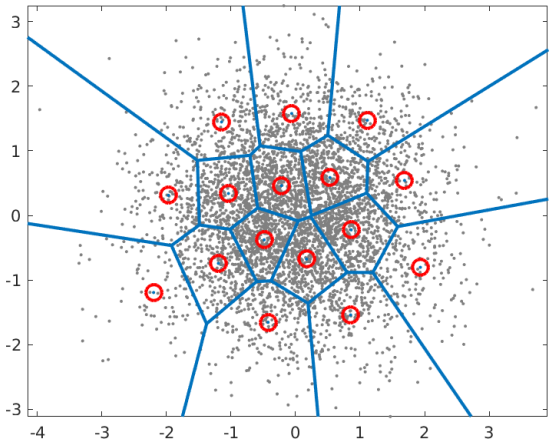
$$R_I(0) = H(X)$$

1.6 Vector Quantizers

A *vector quantizer* is a function that maps a block of source symbols to one of a finite set of reconstruction vectors. This finite set is called the *codebook*, and each element in the codebook is known as a *codeword*. The encoding process involves finding the codeword that is closest (under a chosen distortion measure, such as Euclidean distance) to the input vector.

Given a source producing symbols from an input alphabet $\mathcal{I}$, and assuming we have R bits per symbol to represent the source, a $(2^{nR}, n)$-rate distortion code consists of an encoding function f_n which maps an input block of length n to one of 2^{nR} digital indices, and a corresponding decoding function g_n, which maps each index back to an output sequence in the reproduction alphabet.

In a vector quantizer, the encoder searches for the codeword in the codebook that best approximates the input vector based on a distortion measure (e.g., mean squared error). It then transmits the corresponding index. The decoder receives this index and outputs the associated codeword, reconstructing an approximation of the original input. Together, the encoder and decoder form a lossy compression system that trades accuracy for reduced bit rate.



This image shows a vector quantization diagram where gray dots represent data points scattered in a 2D space. The blue lines form a Voronoi diagram that partitions the space into cells, and the red circles mark the codewords. Each data point is associated with its nearest centroid.

2 Rate Distortion Thereom

We now state the following thereom.

2.1 Rate-Distortion Theorem

The minimum rate R at which a source X can be encoded so that the expected distortion between the source and its reproduction X^* is no greater than D, is equal to the rate-distortion function $R(D)$. This means you cannot reliably compress below $R(D)$ bits per symbol without exceeding the distortion level D. Also, for lossless compression, where $D = 0$, the theorem reduces to Shannon's source coding theorem $R(0) = H(X)$.

Formally, let X^n be an i.i.d. source sequence. For any distortion level $D \geq 0$, the minimum rate R (in bits) required to compress the source such that the expected distortion does not exceed D, is given by the rate-distortion function as defined above:

$$R(D) = \inf_{p(x^*|x):\mathbb{E}[d(X,X^*)]\leq D} I(X; X^*)$$

2.2 Proof Sketch

We attack this proof by proving 2 subclaims.

- Firstly, **achievability**. That is, there exists a coding scheme that achieves distortion D using a rate $R \geq R(D)$. At a high level, we construct a specific random code and prove that it works with high probability, thus proving sufficiency of $R(D)$.
- Secondly, **necessity**. No coding scheme can achieve distortion D using a rate $R < R(D)$. Even if we find a code that works at rate $R = R(D)$, we still have to prove it's the best possible, i.e., no scheme could do better (i.e., use fewer bits). This establishes necessity of the rate $R(D)$.

This seems roundabout at first glance but proving that R cannot be smaller $R(D)$ and R can be greater or equal to $R(D)$ implies that the minimal rate $R = R(D)$.

2.3 Achievability

A few definitions before we proceed.

2.3.1 Conditional Typicality Bound

Let $(x^n, x^{*n}) \in A_d^{(n)}(\epsilon)$, the distortion-typical set under a joint distribution $p(x, x^*)$. Then for sufficiently small $\epsilon > 0$, we have:

$$p(x^{*n}) \geq p(x^{*n} \mid x^n) \cdot 2^{-n(I(X;X^*)+3\epsilon)}$$

To prove this, we start from the definition of conditional probability:

$$p(x^{*n} \mid x^n) = \frac{p(x^n x^{*n})}{p(x^n)}$$

Multiplying and dividing by $p(x^{*n})$, we get:

$$p(x^{*n} \mid x^n) = p(x^{*n}) \cdot \frac{p(x^n x^{*n})}{p(x^n)p(x^{*n})}$$

Now we use the properties of typical sequences. If $(x^n, x^{*n}) \in A_\epsilon^{(n)}(X, X^*)$ then:

$$\begin{aligned} p(x^n, x^{*n}) &\leq 2^{-n(H(X,X^*)-\epsilon)} \\ p(x^n) &\geq 2^{-n(H(X)+\epsilon)}, \\ p(x^{*n}) &\geq 2^{-n(H(X^*)+\epsilon)} \end{aligned}$$

Therefore:

$$\frac{p(x^n, x^{*n})}{p(x^n)p(x^{*n})} \leq \frac{2^{-n(H(X,X^*)-\epsilon)}}{2^{-n(H(X)+\epsilon)} \cdot 2^{-n(H(X^*)+\epsilon)}} = 2^{n(I(X;X^*)+3\epsilon)}$$

Substituting back, we get:

$$p(x^{*n} \mid x^n) \leq p(x^{*n}) \cdot 2^{n(I(X;X^*)+3\epsilon)}$$

Rearranging:

$$p(x^{*n}) \geq p(x^{*n} \mid x^n) \cdot 2^{-n(I(X;X^*)+3\epsilon)}$$

2.3.2 Bound on $(1-xy)^n$

We want to prove that for $0 \leq x, y \leq 1$ and $n > 0$: $(1-xy)^n \leq 1 - x + e^{-yn}$

We use the inequality $1 - z \leq e^{-z}$ which holds for all real z. Substituting $z = xy$, we get:

$$1 - xy \leq e^{-xy}$$

$$(1-xy)^n \leq (e^{-xy})^n = e^{-xyn}$$

Let's define a function $f(x) = 1 - x + e^{-yn} - e^{-xyn}$ for $0 \leq x \leq 1$. We'll show that $f(x) \geq 0$ in this range.

Note that

$$f(0) = 1 - 0 + e^{-yn} - e^0 = 1 + e^{-yn} - 1 = e^{-yn} \geq 0$$
$$f(1) = 1 - 1 + e^{-yn} - e^{-yn} = 0$$

Taking the first derivative of $f(x)$ and finding first order conditions: $f'(x) = -1 + yn \cdot e^{-xyn}$

Setting $f'(x) = 0$:

$$\begin{aligned} -1 + yn \cdot e^{-xyn} &= 0 \\ e^{-xyn} &= \frac{1}{yn} \\ -xyn &= \ln\left(\frac{1}{yn}\right) = -\ln(yn) \\ x &= \frac{\ln(yn)}{yn} \end{aligned}$$

Let's call this critical point

$$x_{cp} = \frac{\ln(yn)}{yn}$$

For all $yn > 1$, we have $0 < x_{cp} < 1$.

Now, taking the second derivative, we have:

$$f''(x) = -(yn)^2 \cdot e^{-xyn} < 0$$

Since $f''(x) < 0$, $f(x)$ is concave down for all x, so the critical point is a maximum. Therefore, $f(x)$ attains its minimum value at the endpoints, which we calculated as $f(0) = e^{-yn} \geq 0$ and $f(1) = 0$.

Since $f(x) \geq 0$ for all $x \in [0, 1]$, we have:

$$e^{-xyn} \leq 1 - x + e^{-yn}$$

Combining this with our result in the beginning:

$$(1-xy)^n \leq e^{-xyn} \leq 1 - x + e^{-yn}$$

2.3.3 Definition: Joint typicality

The set $A_d^{(n)}(\epsilon)$, often denoted simply as A, is the *jointly distortion-typical set*. It consists of all source-reconstruction sequence pairs (x^n, x^{*n}) that are jointly typical with respect to both the joint distribution $p(x, x^*)$ and the distortion constraint.

Formally, we define:

$$A_d^{(n)}(\epsilon) = \left\{ (x^n, x^{*n}) \in \mathcal{X}^n \times \mathcal{X}^{*n} : \left| -\frac{1}{n} \log p(x^n, x^{*n}) - H(X, X^*) \right| < \epsilon, \quad \frac{1}{n} \sum_{i=1}^{n} d(x_i, x_i^*) \leq D + \epsilon \right\}$$

The first condition enforces typicality under the joint distribution, while the second ensures that the empirical distortion between the source sequence and the codeword does not exceed the target distortion D by more than ϵ.

Given a codebook $\mathcal{C} = \{X^{*n}(1), X^{*n}(2), \ldots, X^{*n}(2^{nR})\}$, the encoder observes a source sequence X^n and selects the first codeword $X^{*n}(w)$ such that:

$$(X^n, X^{*n}(i)) \in A_d^{(n)}(\epsilon)$$

If no such codeword exists, the encoder assigns X^n to a default index (e.g., $i = 1$). Otherwise, the index i is transmitted using nR bits.

Proof

We want to prove achievability of the information-theoretic rate-distortion function, i.e., that for any distortion level D, we can build a code that compresses a source to a rate $R \geq R(D)$ and achieves expected distortion no greater than $D + \varepsilon$ for any small $\varepsilon > 0$.

Recall the rate-distortion function is:

$$R(D) = \min_{p(x^*|x):\mathbb{E}[d(X,X^*)]\leq D} I(X; X^*)$$

Let us choose some conditional distribution $p(x^* \mid x)$ that achieves this minimum. Then we compute the total probability of observing x^*, the marginal distribution:

$$p(x^*) = \sum_x p(x)p(x^* \mid x)$$

We construct a codebook $\mathcal{C}$ consisting of 2^{nR} codewords $X^{*n}(i)$, where each codeword is a length-n sequence generated i.i.d. according to $p(x^*)$. That is:

$$X^{*n}(i) \sim \prod_{i=1}^{n} p(x_i^*)$$

Given a source sequence X^n, the encoder chooses the first codeword $X^{*n}(i)$ such that the pair $(X^n, X^{*n}(i))$ is jointly typical with respect to distortion:

$$(X^n, X^{*n}(w)) \in A_d^{(n)}(\epsilon)$$

This set contains all pairs that are typical under the joint distribution $p(x, x^*)$ and whose empirical distortion is within ϵ of the expected value.

If no such codeword is found, we assign the sequence X^n to the first codeword which we reserve as a default/failure codeword. This fallback ensures the encoder is still well-defined even in the rare case of no match.

Now for some case work, we consider the two kinds of source sequences. For a fixed codebook $\mathcal{C}$, consider the two possibilities for a given source sequence x^n:

2.3.4 Case 1: we find a jointly typical codeword

The encoder finds a codeword $X^{*n}(w)$ such that

$$d(x^n, x^{*n}(i)) < D + \epsilon$$

Then the distortion contributed by this sequence is at most $D+\epsilon$. Since the total probability of all such sequences is at most 1, their total contribution to the average distortion is at most $D + \epsilon$.

2.3.5 Case 2: no typical codeword is found

The encoder fails to find any matching codeword, so it defaults to the first codeword. The worst-case distortion incurred here is $d_{\max}$, the largest possible distortion between any two pair (x, x^*). These sequences occur with probability P_f, so their total contribution is at most $P_f d_{\max}$.

Now, we take the weighted average of these 2 cases for total distortion. Combining both cases, the expected distortion over all X^n and codebooks is bounded by:

$$\mathbb{E}\left[d(X^n, X^{*n})\right] \leq D + \epsilon + P_f d_{\max}$$

Since $\varepsilon > 0$ is arbitrary, we can get distortion arbitrarily close to D as long as $P_f \to 0$ as $n \to \infty$, the total distortion is arbitrarily close to $D + \varepsilon$.

To ensure this, it suffices to show we can choose a small ε and show that the error probability P_f can be made small.

To analyze the performance of the random code construction, we study the probability that a source sequence $x^n \sim p(x)^n$ is not represented by any codeword in the codebook $\mathcal{C}$. Specifically, we consider the case where there is no codeword $x^{*n} \in \mathcal{C}$ such that $(x^n, x^{*n}) \in A_d^{(n)}(\epsilon)$, i.e., the pair is not jointly typical with respect to distortion.

Let us define an indicator function for distortion typicality:

$$\mathbb{1}_{\text{enc}}(x^n, x^{*n}) = \begin{cases} 1 & \text{if } (x^n, x^{*n}) \in A_d^{(n)}(\epsilon), \\ 0 & \text{otherwise} \end{cases}$$

By complementary counting, the probability that a single randomly drawn codeword does not represent x^n is the following :

$$\Pr((x^n, X^{*n}) \notin A_d^{(n)}(\epsilon)) = 1 - \mathbb{E}[\mathbb{1}_{\text{enc}}(x^n, x^{*n})] = 1 - \sum_{x^{*n}} p(x^{*n}) \mathbb{1}_{\text{enc}}(x^n, x^{*n})$$

Since the encoder draws 2^{nR} codewords i.i.d, the probability that none of them represent x^n is simply the previous probability raised to the 2^{nR} power.

$$\left[1 - \sum_{x^{*n}} p(x^{*n}) \mathbb{1}_{\text{enc}}(x^n, x^{*n})\right]^{2^{nR}}$$

Now averaging over all input sequences $x^n \sim p(x^n)$ we have:

$$P_f = \sum_{x^n} p(x^n) \left[1 - \sum_{x^{*n}} p(x^{*n}) \mathbb{1}_{\text{enc}}(x^n, x^{*n})\right]^{2^{nR}}$$

By the conditional typicality lemma, for typical x^n, we have $p(x^{*n}) \geq p(x^{*n} \mid x^n) \cdot 2^{-n(I(X;X^*)+3\epsilon)}$ on the typical set. So,

$$\sum_{x^{*n}} p(x^{*n}) \mathbb{1}_{\text{enc}}(x^n, x^{*n}) \geq \sum_{x^{*n}} p(x^{*n} \mid x^n) \cdot 2^{-n(I(X;X^*)+3\epsilon)} \cdot \mathbb{1}_{\text{enc}}(x^n, x^{*n})$$

We now substitute this hunk of an expression into our bound for P_f.

$$P_f \leq \sum_{x^n} p(x^n) \left[1 - \sum_{x^{*n}} p(x^{*n} \mid x^n) \cdot 2^{-n(I(X;X^*)+3\epsilon)} \cdot \mathbb{1}_{\text{enc}}(x^n, x^{*n})\right]^{2^{nR}}$$

Applying the $(1-xy)^n \leq 1 - x + e^{-yn}$ lemma with $x = \sum_{x^{*n}} p(x^{*n} \mid x^n) \mathbb{1}_{\text{enc}}(x^n, x^{*n})$ and $y = 2^{n(R-I(X;X^*)-3\epsilon)}$, we obtain:

$$\left[1 - \sum_{x^{*n}} p(x^{*n} \mid x^n) \cdot 2^{-n(I(X;X^*)+3\epsilon)} \cdot \mathbb{1}_{\text{enc}}(x^n, x^{*n})\right]^{2^{nR}} < 1 - \sum_{x^{*n}} p(x^{*n} \mid x^n) \mathbb{1}(x^n, x^{*n}) + e^{-\left(2^{-n(I(X;X^*)+3\epsilon)} 2^{nR}\right)}$$

Combining like terms and simplifying the exponents:

$$P_f \leq \sum_{x^n} p(x^n) \left[1 - \sum_{x^{*n}} p(x^{*n} \mid x^n) \mathbb{1}(x^n, x^{*n}) + e^{-2^{n(R-I(X;X^*)-3\epsilon)}}\right]$$

Next, we bound the term:

$$1 - \sum_{x^n, x^{*n}} p(x^n, x^{*n}) \mathbb{1}(x^n, x^{*n}) = \Pr\left((X^n, X^{*n}) \notin A_d^{(n)}(\epsilon)\right) < \epsilon$$

for sufficiently large n, by the distortion-typicality lemma. The remaining term

$$e^{-2^{n(R-I(X;X^*)-3\varepsilon)}}$$

vanishes as $R - I(X;X^*) - 3\varepsilon > 0$ or $R > I(X;X^*) + 3\varepsilon$. By construction, we chose $p(x^*|x)$ explicitly such that $I(X;X^*) = R(D)$. That is, we are constructing our code using a distribution where the mutual information is exactly equal to the rate-distortion function. So any $R = I(X;X^*) + O(\varepsilon)$ naturally also satisfies $R = R(D) + O(\varepsilon)$.

Finally, we return to our expression for P_f with $R = R(D) + O(\varepsilon)$:

$$\begin{aligned} P_f &\leq \sum_{x^n} p(x^n) \left[1 - \sum_{x^{*n}} p(x^{*n} \mid x^n) \mathbb{1}(x^n, x^{*n}) + 0 \right] \\ &\leq 1 - \sum_{x^n} p(x^n) \sum_{x^{*n}} p(x^{*n} \mid x^n) \mathbb{1}(x^n, x^{*n}) \qquad \text{by LOE} \end{aligned}$$

For sufficiently large n, the probability that x^n is successfully encoded $p(x^{*n} \mid x^n)\mathbb{1}(x^n, x^{*n})$ converges to 1 by law of large numbers thus for large enough n, we obtain:

$$\lim_{n\to\infty} P_f \leq \epsilon$$

Therefore, the expected distortion of the code satisfies:

$$\mathbb{E}[d(X^n, X^{*n})] \leq D + \epsilon + P_f d_{\max} < D + O(\epsilon)$$

for any $\varepsilon > 0$, $R > R(D)$.

2.4 Necessity

We present a few lemmas and definitions to make the proof more smooth.

2.4.1 Chain Rule of Conditional Entropy

We aim to derive the chain rule for entropy, which expresses the joint entropy of a sequence of random variables as a sum of conditional entropies. In equations, that is:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, X_2, \ldots, X_{i-1})$$

To begin, we start with the joint entropy over $X_1, X_2, \ldots, X_n$ is:

$$H(X_1, X_2, \ldots, X_n) = -\sum_{x_1, x_2, \ldots, x_n} P(x_1, x_2, \ldots, x_n) \log P(x_1, x_2, \ldots, x_n)$$

Directly from the chain rule of probability:

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x_2) \cdots P(x_n \mid x_1, \ldots, x_{n-1})$$

Taking the log:

$$\log P(x_1, \ldots, x_n) = \log P(x_1) + \log P(x_2 \mid x_1) + \cdots + \log P(x_n \mid x_1, \ldots, x_{n-1})$$

Substituting back in

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= -\sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n) \left[\log P(x_1) + \log P(x_2 \mid x_1) + \cdots + \log P(x_n \mid x_1, \ldots, x_{n-1}) \right] \\ &= -\sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n) \log P(x_1) \\ &\quad -\sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n) \log P(x_2 \mid x_1) \\ &\quad - \cdots \\ &\quad -\sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n) \log P(x_n \mid x_1, \ldots, x_{n-1}) \end{aligned}$$

Each of these terms corresponds to a conditional entropy. So:

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1, \ldots, X_{n-1}) \\ &= \sum_{i=1}^{n} H(X_i \mid X_1, X_2, \ldots, X_{i-1}) \end{aligned}$$

2.4.2 R(D) is convex

We now show that the rate-distortion function $R(D)$ is convex.

Let $p_1(x, x^*) = p(x)p_1(x^* \mid x)$ and $p_2(x, x^*) = p(x)p_2(x^* \mid x)$ be two joint distributions with distortions D_1 and D_2, and rates $R_1 = I_{p_1}(X; X^*)$, $R_2 = I_{p_2}(X; X^*)$. Define a convex combination:

$$p_\lambda(x, x^*) = \lambda p_1(x, x^*) + (1 - \lambda)p_2(x, x^*)$$

Then distortion is linear because it's a weighted average of distortions associated with the two individual distributions.

$$D(p_\lambda) = \lambda D_1 + (1 - \lambda)D_2$$

By the definition of $R(D)$ and the convexity of mutual information:

$$R(D(p_\lambda)) \leq I_{p_\lambda}(X; X^*) \leq \lambda I_{p_1}(X; X^*) + (1 - \lambda)I_{p_2}(X; X^*)$$

Hence,

$$R(\lambda D_1 + (1 - \lambda)D_2) \leq \lambda R(D_1) + (1 - \lambda)R(D_2)$$

2.4.3 Proof

Now we move on to proving the converse.

For any $(2^{nR}, n)$ rate-distortion code with expected distortion at most D, the rate must satisfy $R \geq R(D)$.

Let (f_n, g_n) be an encoding-decoding pair such that:

$$X^{*n} = g_n(f_n(X^n))$$

Let's say that X^* has M possible values. Entropy is maximized when X^* is completely unpredictable. This happens when X^* is uniformly distributed over all codewords.

$$p(X_i^*) = \frac{1}{M} \quad \text{for all } i$$

$$H(X) = -\sum_{i=1}^{M} \frac{1}{M} \log_2 \left(\frac{1}{M}\right) = -M \cdot \frac{1}{M} \log_2 \left(\frac{1}{M}\right) = \log_2 M$$

No distribution over the same support can have higher entropy. This is easy to see if we treat this as a constrained optimization problem.

$$\max H(p_1, \ldots, p_M) = -\sum_{i=1}^{M} p_i \log p_i \quad \text{s.t.} \quad \sum_{i=1}^{M} p_i = 1$$

$$\mathcal{L}(p_1, \ldots, p_M, \lambda) = -\sum_{i=1}^{M} p_i \log p_i + \lambda \left(\sum_{i=1}^{M} p_i - 1\right)$$

$$\frac{\partial \mathcal{L}}{\partial p_i} = -\log p_i - 1 + \lambda = 0 \quad \rightarrow \quad \log p_i = \lambda - 1$$

So $p_i =$ constant for all i, meaning:

$$p_i = \frac{1}{M}$$

Since the encoder's output takes at most 2^{nR} values:

$$H(X^{*n}) \leq \log_2(2^{nR}) = nR$$

Conditional entropy $H(A \mid B)$ measures the uncertainty remaining in A after knowing B.

X^{*n} is the output of a deterministic decoding function applied to the encoded representation of X^n. So, once X^n is known, the value of X^{*n} is fully determined, and thus:

$$H(X^{*n} \mid X^n) = 0$$

Therefore:

$$nR \geq H(X^{*n}) - H(X^{*n} \mid X^n)$$

Since we defined that X^n is an i.i.d source sequence and using the chain rule for conditional entropy from above we have:

$$nR \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid X^{*n}, X_{i-1}, X_{i-2}, \ldots, X_1)$$

By the fact that conditioning can only reduce entropy which follows from the fact that mutual information is non-negative, $I(A; B) = H(A) - H(A \mid B) \geq 0$:

$$\sum_{i=1}^{n} H(X_i \mid X^{*n}, X_{i-1}, X_{i-2}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i \mid X^{**}_i)$$

$$nR \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid X^{**}_i) = \sum_{i}^{n} I(X_i; X^*_i)$$

For each variable X_i, we have a corresponding rate-distortion function $R\left(\mathbb{E}[d(X_i, X_i^*)]\right)$, which gives the minimum rate required for encoding X_i such that the expected distortion is below D. Subsituting our definition of mutual information from above:

$$nR \geq \sum_{i=1}^{n} R\left(\mathbb{E}[d(X_i, X_i^*)]\right)$$

Applying Jensen's inequality since we proved above that $R(D)$ is convex in D:

$$nR \geq nR\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[d(X_i, X_i^*)]\right)$$

Since

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[d(x_i, x_i^*)] = \mathbb{E}[d(x^n, x^{*n})] \leq D$$

Thus,

$$R \geq R(D)$$

3 Shannon's Lower Bound

We now prove one of the most fundamental thereoms of information thereom, the Shannon lower bound. The Shannon lower bound states that the rate-distortion function $R(D)$ is the fundamental lower bound on an quantization scheme. No scalar or vector quantizer can do better, on average rate, than $R(D)$.

More formally, given random variable X drawn from a finite set $\{1, 2, \ldots, m\}$ and some distortion function $D(x, x^*)$ where all columns of the distortion matrix are fixed permutations of the same set $\{d_1, d_2, ...d_m\}$.

More formally, the bound says:

$$R(D) \geq H(X) - \max_{p:\sum_{i=1}^{m} p_i d_i \leq D} H(p)$$

where:

- $H(X)$ is the entropy of the source
- $\max_{p:\sum_{i=1}^{m} p_i d_i \leq D} H(p)$ is a function that gives the maximum entropy of any probability distribution p over the source symbols such that the average distortion is at most D.

So, you can't compress the source below this bound without exceeding the distortion limit. Again, as in the previous parts, we present some lemmas before proceeding.

3.1 Concavity of $\max_{p:\sum_{i=1}^{m} p_i d_i \leq D} H(p)$

Let $f(D) = \max_{p:\sum_{i=1}^{m} p_i d_i \leq D} H(p)$. To prove f is concave, take any D_1, D_2, and $\lambda \in [0, 1]$.

Let p_1 and p_2 be the distributions maximizing entropy at distortion bounds D_1 and D_2 respectively. Define $p_\lambda = \lambda p_1 + (1 - \lambda)p_2$.

Since distortion is linear in p:

$$\sum_{i=1}^{m} (p_\lambda)_i d_i = \lambda \sum_{i=1}^{m} (p_1)_i d_i + (1 - \lambda) \sum_{i=1}^{m} (p_2)_i d_i \leq \lambda D_1 + (1 - \lambda) D_2$$

Therefore, p_λ is feasible for distortion bound $\lambda D_1 + (1 - \lambda)D_2$.

Since entropy is concave:

$$H(p_\lambda) \geq \lambda H(p_1) + (1 - \lambda) H(p_2)$$

Combining these facts:

$$f(\lambda D_1 + (1 - \lambda)D_2) \geq H(p_\lambda) \geq \lambda H(p_1) + (1 - \lambda) H(p_2) = \lambda f(D_1) + (1 - \lambda) f(D_2)$$

Therefore, f is concave.

3.2 Lemma: $\sum_{x^*} p(x^*) D_{x^*} \leq D$

We begin with the distortion constraint from the definition of the rate-distortion function:

$$\mathbb{E}[d(X, X^*)] = \sum_{x, x^*} p(x, x^*) d(x, x^*) \leq D$$

Now define the conditional expected distortion for each x^* as:

$$D_{x^*} = \sum_{x} p(x|x^*) d(x, x^*)$$

Then, taking the expectation over x^*, we have:

$$\sum_{x^*} p(x^*) D_{x^*} = \sum_{x^*} p(x^*) \sum_{x} p(x|x^*) d(x, x^*) = \sum_{x, x^*} p(x, x^*) d(x, x^*)$$

Therefore, using the distortion constraint again:

$$\sum_{x^*} p(x^*) D_{x^*} = \mathbb{E}[d(X, X^*)] \leq D$$

3.3 Proof

This proof is actually less involved than the previous theorem which we will use to begin.

We have some $D \geq \mathbb{E}[d(X, X^*)]$. Directly from the rate-distortion theorem:

$$R(D) = \min_{p(x^*|x):\sum_{x,x^*} p(x)p(x^*|x)d(x,x^*) \leq D} I(X; X^*).$$

By definition of mutual information and unpacking $H(X^*)$ with the definition of conditional entropy we have:

$$I(X;X^*) = H(X) - H(X|X^*) = H(X) - \sum_{x^*} p(x^*) H(X \mid X^* = x^*)$$

Trivially since maximum entropy over all probability distributions satisfying max distortion D must be at least as big as entropy of any specific feasible distribution, we have:

$$\max_{p:\sum_{i=1}^m p_i d_i \leq D} H(p) \geq H(X|X^* = x^*) \quad \rightarrow \quad I(X;X^*) \geq H(X) - \sum_{x^*} p(x^*) \max_{p:\sum_{i=1}^m p_i d_i \leq D_{x^*}} H(p)$$

Since $\max_{p:\sum_{i=1}^m p_i d_i \leq D} H(p)$ is concave. Again applying Jensen's inequality:

$$\sum_{x^*} p(x^*) \max_{p:\sum_{i=1}^m p_i d_i \leq D_{x^*}} H(p) \leq \max_{p:\sum_{i=1}^m p_i d_i \leq D_{x^*}} \sum_{x^*} p(x^*) H(p)$$

Which directly leads to:

$$I(X;X^*) \geq H(X) - \max_{p:\sum_{i=1}^m p_i d_i \leq D_{x^*}} \sum_{x^*} p(x^*) H(p)$$

Now, since $\sum_{x^*} p(x^*) D_{x^*} = \mathbb{E}[D_{X^*}] \leq D$ by our lemma above, we have:

$$\sum_{x^*} p(x^*) \max_{p:\sum_{i=1}^m p_i d_i \leq D_{x^*}} H(p) \leq \max_{p:\sum_{i=1}^m p_i d_i \leq D} H(p)$$

This gives us:

$$H(X|X^*) = \sum_{x^*} p(x^*) H(X \mid X^* = x^*) \leq \sum_{x^*} p(x^*) \max_{p:\sum_{i=1}^m p_i d_i \leq D_{x^*}} H(p) \leq \max_{p:\sum_{i=1}^m p_i d_i \leq D} H(p)$$

Now substitute into the mutual information expression:

$$I(X;X^*) = H(X) - H(X|X^*) \geq H(X) - \max_{p:\sum_{i=1}^m p_i d_i \leq D} H(p)$$

Finally, since $R(D) = \min I(X;X^*)$ over all feasible $p(x^*|x)$, we obtain the lower bound:

$$R(D) \geq H(X) - \max_{p:\sum_{i=1}^m p_i d_i \leq D} H(p)$$

References

[1] L. Devroye, *A quick introduction to information theory*, McGill University, 2015.

[2] C. P. Chen, *Rate-Distortion Theory Notes*, National Sun Yat-sen University

[3] T. Linder and R. Zamir, *On the Asymptotic Tightness of the Shannon Lower Bound*

[4] Y. W. Liao, *Lecture 13: Rate-Distortion Theory*, Yale University

[5] S. K. B. Yeung, *Rate Distortion Theory*, Stanford University

[6] F. M. Zennaro, *Notes on Rate Distortion Theory*