Electrical Engineering 229A Lecture 25 Notes

Daniel Raban

November 23, 2021

1 Rate Distortion Theory

1.1 Shannon's rate distortion theorem

In rate distortion theory,, we have an iid $\mathscr{X}$-valued source $X_1, \dots, X_n$. At the compressor, we have

$$f_n : \mathscr{X}^n \to \{1, 2, \dots, 2^{nR}\}.$$

Assume $f_n(X^n)$ is perfectly received at the decompressor. The decompressor uses a map $g_n : \{1, \dots, 2^{nR}\} \to \widehat{\mathscr{X}^n}$, where $\widehat{\mathscr{X}}$ could be different from $\mathscr{X}$. Call (f_n, g_n) a $(2^{nR}, n)$ **distortion code**. We are also given a cost metric $d : \mathscr{X} \times \widehat{\mathscr{X}} \to \mathbb{R}_+$ called the **distortion measure**.

Definition 1.1. (R, D) is called **achievable** if there exists $((f_n, g_n), n \geq 1)$ of $(2^{nR}, n)$ distortion codes such that

$$\limsup_n \frac{1}{n} \mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D,$$

where $d(x^n, \widehat{x}^n)$ denotes $\sum_{i=1}^n d(x_i, \widehat{x}_i)$.

Theorem 1.1 (Shannon's rate distortion theorem)**.** *Let*

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

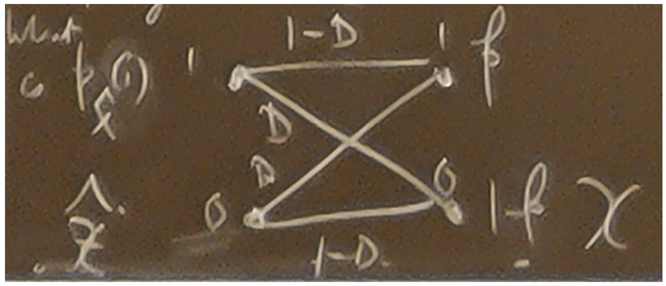
where $p(x)$ is the marginal distribution of the source. (R, D) is achievable if $R > R^{(I)}(D)$ and not achievable if $R < R^{(I)}(D)$.

We write $R^{(I)}(D)$ as $R(D)$, the **rate distortion function**.

Example 1.1 (Bernoulli source)**.** Let $\mathscr{X} = \{0, 1\}$ with $p(1) = p$ and $p(0) = 1 - p$, reproduction alphabet $\widehat{\mathscr{X}} = \{0, 1\}$, and distortion measure $d(0, 0) = 0 = d(1, 1)$, $d(0, 1) = 1 = d(1, 0)$. Here,

$$R(D) = \begin{cases} h(p) - h(D)Y & 0 \leq D \leq \min\{p, 1 - p\} \\ 0 & \text{otherwise} \end{cases}$$

To see this, if $D > \min\{p, 1-p\}$ then if $p < 1/2$, represent all binary sequences of length n by 0^n; if $p > 1/2$, represent all binary sequences of length n by 1^n. If $d \leq \min\{p, 1-p\}$, we can choose the optimizing $p(\widehat{x} \mid x)$ by defining the corresponding $p(x \mid \widehat{x})$ via a binary symmetric channel with crossover probability D ($p(\widehat{x})$ has to be chosen correctly to get the correct $p(x)$).



We must have

$$p_{\widehat{X}}(1)(1-D) + (1 - p_{\widehat{X}}(1))D = p$$

which gives

$$p_{\widehat{X}}(1) = \frac{p-D}{1-2D}.$$

This makes sense because $D \leq \min\{p, 1-p\}$ and $D \leq 1/2$.

If we made this choice, then

$$\begin{aligned} I(X;\widehat{X}) &= H(X) - H(X \mid \widehat{X}) \\ &= h(p) - h(D). \end{aligned}$$

To show that this is the best choice, we need to show that $I(X;\widehat{X}) \geq h(p) - h(D)$ for all other choices of $p(\widehat{x} \mid x)$. This holds because for any other choice of $p(\widehat{x} \mid x)$,

$$\begin{aligned} I(\widehat{X};X) &= H(X) - H(X \mid \widehat{X}) \\ &= h(p) - H(X \mid \widehat{X}) \\ &= h(p) - H(X \oplus \widehat{X} \mid \widehat{X}) \\ &\geq h(p) - H(X \oplus \widehat{X}) \\ &\geq h(p) - h(D). \end{aligned}$$

1.2 Proof of the rate distortion theorem

Let's prove the theorem.

Proof. Converse: We want to show that if $R < R^{(I)}(D)$¡ then (R, D) is not achievable. First, observe that $R^{(I)}(D)$ is a convex function of D (using $I(X;\widehat{X})$ is convex in $[p(\widehat{x} \mid x)]$

for fixed $(p(x), x \in \mathscr{X})$). Consider any sequence $((f_n, g_n), n \geq 1)$ of $(2^{nR}, n)$ rate distortion codes. Then

$$\begin{aligned} nR &\geq H(\widehat{X}^n) \\ &\geq I(X^n; \widehat{X}^n) \\ &= H(X^n) - H(X^n \mid \widehat{X}^n) \end{aligned}$$

Use the chain rule

$$\begin{aligned} &= H(X^n) - \sum_{i=1}^{n} H(X_i \mid X^{i-1}, \widehat{X}^n) \\ &= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid X^{i-1}, \widehat{X}^n) \end{aligned}$$

Conditioning on more decreases the entropy, so

$$\begin{aligned} &\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid \widehat{X}_i) \\ &= \sum_{i=1}^{n} I(X_i; \widehat{X}_i) \end{aligned}$$

If D_i denotes $\mathbb{E}[d(X_i, \widehat{X}_i)]$, then $I(X_i; \widehat{X}_i) \geq R^{(I)}(D_i)$.

$$\geq \sum_{i=1}^{n} R^{(I)}(D_i)$$

By convexity of $R^{(I)}$,

$$= nR^{(I)}(D).$$

Achievability: We use a random coding argument. Given $p(x, \widehat{x})$, define the set

$$A_{d,\varepsilon}^{(n)} := \left\{ (x^n, \widehat{x}^n) : (x^n, \widehat{x}^n) \in A_{\varepsilon}^{(n)}, \left| \frac{1}{n} \sum_{i} d(x_i, \widehat{x}_i) - \mathbb{E}[d(X, \widehat{X})] \right| < \varepsilon \right\}.$$

We can show that $\mathbb{P}((X^n, \widehat{X}^n) \in A_{d,\varepsilon}^{(n)}) \to 1$ as $n \to \infty$, where $(X_i, \widehat{X}_i)$ are iid $\sim p(x, \widehat{x})$. We can also show that

$$(1-\varepsilon) 2^{nH(X,\widehat{X})} 2^{-n\varepsilon} \leq |A_{d,\varepsilon}^{(n)}| \leq 2^{nH(X,\widehat{X})} 2^{n\varepsilon},$$

where the lower bound holds for all sufficiently large n. If $X^n \amalg \widetilde{\widehat{X}}^n$, where X^n is iid $\sim p(x)$ and $\widehat{X}^n$ is iid $\sim p(\widehat{x})$, then

$$(1-\varepsilon) 2^{-nI(X;\widehat{X})} 2^{-n3\varepsilon} \leq \mathbb{P}(X^n, \widetilde{\widehat{X}}^n \in A_{d,\varepsilon}^{(n)}) \leq 2^{-nI(X;\widehat{X})} 2^{n3\varepsilon}.$$

So generate 2^{nR} sequences

$$\begin{bmatrix} \widehat{X}_1(1) & \cdots & \widehat{X}_n(1) \\ \vdots & & \vdots \\ \widehat{X}_1(2^{nR}) & \cdots & \widehat{X}_n(2^{nR}) \end{bmatrix}$$

with entries iid over the coordinates and $\sim p(\widehat{x})$. To construct $f_n : \mathscr{X}^n \to \{1, \ldots, 2^{nR}\}$, on seeing x^n, find a row ℓ such that $(x^n, \widehat{X}^n(\ell)) \in A_{d,\varepsilon}^{(n)}$ if such exists. Then define $g_n : \{1, \ldots, 2^{nR}\} \to \widehat{\mathscr{X}^n}$ by $g_n(\ell)$ is row ℓ.

We claim that if $R > I(X; \widehat{X})$, then $\mathbb{P}(\text{row exists for } X^n) \to 1$ as $n \to \infty$. If $R > I(X; \widehat{X}) + 3\varepsilon$, then

$$\left(1 - (1-\varepsilon)2^{-nI(X;\widehat{X})}2^{-3n\varepsilon}\right)^{2^{nR}} \to 0.$$

This completes the proof. $\square$

1.3 The rate distortion function with a Gaussian source

For an iid Gaussian source $X_1, \ldots, X_n$ iid $\sim \mathcal{N}(0, \sigma^2)$, $\mathscr{X} = \widehat{\mathscr{X}} = \mathbb{R}$, and distortion $d(x, \widehat{x}) = (x - \widehat{x})^2$ with the goal of asymptotical per letter distortion at most D (i.e. $\frac{1}{n}\,\mathbb{E}[\sum_{i=1}^n (X_i - \widehat{X}_i)^2] \le D$ with $\widehat{X}^n = g_n(f_n(X^n)))$, we have

$$R^{(I)}(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & \text{if } D < \sigma^2 \\ 0 & \text{if } D > \sigma^2. \end{cases}$$

The first case is achieved via $Z \sim \mathcal{N}(0, \sigma^2 - D)$, where $\widehat{X} \amalg Z$. Here, $I(X; \widehat{X}) = h(X) - h(X \mid \widehat{X}) = \frac{1}{2}\log\frac{\sigma^2}{D}$.