EE376A/STATS376A Information Theory Lecture 12 - 02/15/2018

Lecture 12: Lossy Compression & Rate Distortion Theory

Lecturer: Tsachy Weissman *Scribe: Charles Chen, Jason Kriss*

In this lecture, we introduce basic definitions for lossy compression schemes and introduce rate distortion theory. We present without proof the main result equating the rate distortion function with the informational rate distortion function ($R(D) = R^{(I)}(D)$) and provide examples of rate distortion computations. Lossy compression is of interest even outside of information theory, with applications in statistics and machine learning (clustering).

1 Compression

Recall the following setup for the general compression problem:

- **source sequence**: $U_1, U_2, ...U_N \sim$ iid $U \in \mathcal{U}$.
- **encoder**: maps N source symbols to n bits
- **decoder**: maps the n bits to the reconstructed symbols $V_1, V_2, ...V_N$
- **rate**: the number of bits used per source symbol (n/N)

$$U_1, U_2, ..., U_N \rightarrow \boxed{\text{encoder}} \xrightarrow{n \text{ bits}} \boxed{\text{decoder}} \rightarrow V_1, V_2, ..., V_N \tag{1}$$

For **lossless compression**, we want to have the reconstructed sequence perfectly match the source sequence. In **lossy compression**, we are willing to accept some amount of reconstruction error (**distortion**) in exchange for a potentially lower compression rate than could be achieved losslessly.

2 Rate Distortion Theory

2.1 Definitions

Definition 1. *a* ***distortion function*** *is a mapping*

$$d : U \times V \rightarrow \mathbb{R}_{\geq 0} \tag{2}$$

Definition 2. *the* ***distortion between sequences*** U^N *and* V^N *is*

$$d(U^N, V^N) = \frac{1}{N}\sum_{i=1}^{N} d(U_i, V_i) \tag{3}$$

Note that the U_is are random variables, so this distortion is itself a random variable. Therefore, when talking about the distortion of a given scheme, we usually mean the *expected* per-symbol distortion

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} d(U_i, V_i)\right] \tag{4}$$

With lossy compression, there is a natural tradeoff between the rate and the distortion. The more distortion we are willing to accept, the lower the rate we can hope to achieve. However, we usually will want to constrain the distortion to some upper limit. At one extreme, if we constrain ourselves to have zero distortion, we end up back in the lossless compression setup where we know that the minimal rate is the entropy.

Definition 3. *a compression **scheme** is defined to be a tuple (N, n, encoder, decoder)*

Definition 4. *a rate/distortion pair* (R, D) *is said to be **achievable** if* $\forall \epsilon > 0$, $\exists$ *scheme such that*

$$\frac{n}{N} \leq R + \epsilon \tag{5}$$

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} d(U_i, V_i)\right] \leq D + \epsilon \tag{6}$$

Definition 5. ***rate-distortion function***

$$R(D) \triangleq \inf\{R' : (R', D) \text{ is achievable}\} \tag{7}$$

The rate-distortion function is a natural analog to the channel capacity in a communication context. It represents the "best" rate we can hope to achieve for a given level of distortion.

Definition 6. ***information rate-distortion function***

$$R^{(I)}(D) \triangleq \min_{\mathbb{E}[d(U,V)] \leq D} I(U; V) \tag{8}$$

The information rate-distortion function is a natural analog to the information channel capacity in a communication context. Note that the distribution of U is given to us, so this minimization problem is actually over all possible conditional distributions of $V|U$.

2.2 Main Result

We now have the necessary setup to state the main result of rate distortion theory.

Theorem 7.

$$R(D) = R^{(I)}(D) \tag{9}$$

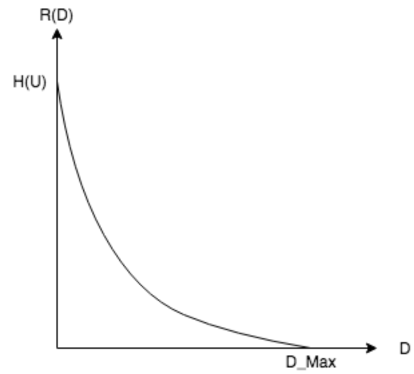
3 Intuition and Examples

Today, we give a sense of why perhaps this main result makes sense. In the next few lectures, we will prove the main result and present concrete lossy compression schemes.

3.1 What does $R(D)$ look like (qualitatively)?

Without focusing on a specific source distribution, we can understand some basic properties of the rate-distortion function. In Figure 1, we can see that $R(D)$ takes its maximal value at $D = 0$. This is the lossless compression case, and thus $R(0) = H(U)$. On the other end, we see that $R(D)$ reaches its minimal value of 0 at $D_{\max} \triangleq \min_v \mathbb{E}[d(U, v)]$. Intuitively, if we are willing to accept distortion of $D_{\max}$, then we can simply encode 0 bits and always decode as v. One other property we can see in the graph is that the rate-distortion function is monotonically decreasing. This makes sense as we can always compress at at least the same rate if given a higher distortion allowance. Further, we have the following convexity constraint on $R(D)$.

Figure 1: Sketch of a rate-distortion function



Claim: $R(D)$ is convex, i.e., for all $0 \leq \alpha \leq 1$, D_0 and D_1,

$$R(\alpha D_0 + (1-\alpha)D_1) \leq \alpha R(D_0) + (1-\alpha)R(D_1)$$

Proof idea: Consider the "time-sharing scheme" for encoding the source symbols $(U_1, U_2, \cdots, U_N)$. We take the first αN source symbols and encode them by employing a "good" scheme for $D = D_0$, and for the remaining $(1-\alpha)N$ source symbols, we employ a "good" scheme for $D = D_1$.

The expected distortion of this scheme is then $\alpha D_0 + (1-\alpha)D_1$. From this, it follows that the minimal number of bits needed across all schemes is at most the number of bits needed for this specific scheme, i.e.:

$$R(\alpha D_0 + (1-\alpha)D_1) \leq \alpha R(D_0) + (1-\alpha)R(D_1),$$

where the left hand side is the minimal number of bits required to achieve distortion $\alpha D_0 + (1-\alpha)D_1$, and the right hand side is the number of bits per symbol required by our "time-sharing" scheme to achieve distortion $\alpha D_0 + (1-\alpha)D_1$.

Homework exercise: Verify using properties of mutual information that the informational rate distortion function $R^{(I)}(D)$ is convex, without assuming the main result. Next week, when we prove the main result $R^{(I)}(D) = R(D)$, we will use this property.

3.2 Examples

Example I: Let $U \sim \text{Bern}(p)$ with $p \leq 1/2$ and define the Hamming distortion as:

$$d(u,v) = \begin{cases} 0, & u = v \\ 1, & u \neq v \end{cases}.$$

In this setting, U and V take values in $\mathcal{U}$ and $\mathcal{V}$, respectively, where $\mathcal{U} = \mathcal{V} = \{0,1\}$. Here, we claim that:

$$R(D) = \begin{cases} h_2(p) - h_2(D), & 0 \leq D \leq p \\ 0, & D > p \end{cases}.$$

Note that the function here is convex, since in the region $0 \leq D \leq p$, the function takes the value of a constant minus the binary entropy function, and the binary entropy function is concave. In the region $D > p$, the reconstruction can be chosen to be all zeros, with no bits to describe the source symbols, and still achieve the desired maximum distortion. Further, when $D = 0$ and no distortion is allowed, we need $h_2(p)$ bits, which is consistent with our previous work on entropy.

Proof: When the allowed distortion is $D > p$, we don't need any description. Thus, let us consider the case where $0 \le D \le p$. For any U, V such that $U \sim \text{Bern}(p)$ and $\mathbb{E}[d(U, v)] = P(U \neq V) \le D \le p \le 1/2$, consider

$$I(U;V) = H(U) - H(U|V) = H(U) - H(U \oplus_2 V|V)$$

We have that conditioning reduces entropy, so the above quantity satisfies the inequality below (with equality if $U \oplus_2 V$ and V are independent):

$$H(U) - H(U \oplus_2 V|V) \ge H(U) - H(U \oplus_2 V) = h_2(p) - h_2(P(U \neq V))$$

Next, note that the binary entropy function h_2 is monotonic on the interval $[0, 1/2]$, so the following holds (with equality if $P(U \neq V) = D$):

$$h_2(p) - h_2(P(U \neq V)) \ge h_2(p) - h_2(D)$$

Thus, $I(U;V) \ge h_2(p) - h_2(D)$ (so $R(D) = R^{(I)}(D) = \min_{\mathbb{E}[d(U,V) \le D]} \ge h_2(p) - h_2(D)$). We would like to show that equality here is achievable, which occurs if the two equality conditions above are satisfied. This is the case if we can find U, V in the feasible set such that: (1) $U \oplus_2 V$ is independent of V and (2) $U \oplus_2 V \sim \text{Bern}(D)$.

That is, can we find a binary variable $V \sim \text{Bern}(q)$ such that if you add (mod 2) $Z \sim \text{Bern}(D)$ with Z independent of V, it will yield our source $U \sim \text{Bern}(p)$ (note $V \oplus_2 Z = U \iff Z = U \oplus_2 V$)? If that is the case, we must have

$$p = P(U = 1) = q(1 - D) + (1 - q)D = q(1 - 2D) + D$$

$$\Rightarrow 0 \le q = \frac{p - D}{1 - 2D} \le \frac{\frac{1}{2} - D}{1 - 2D} = \frac{1}{2}.$$

Thus, we found a joint distribution U, V for which the equality $I(U;V) = h_2(p) - h_2(D)$ holds, and so in the region $0 \le D \le p$, $R(D) = h_2(p) - h_2(D)$, completing our proof.

Example II: Now, let's look at an analog source $U \sim \mathcal{N}(0, \sigma^2)$. We claim that

$$R(D) = \begin{cases} \frac{1}{2}\log(\sigma^2/D), & 0 < D \le \sigma^2 \\ 0, & D > \sigma^2 \end{cases}.$$

This function is convex, and for allowed distortion D greater than the variance σ^2, we don't need any bits to describe the reconstruction, since it can be taken to be always zero. Next, this is an analog source, so the entropy is infinite, so we can't expect to describe it and get zero distortion for a fixed number of bits per source symbol. Note further that if we invert $R(D)$, we obtain $D(R)$ (*the "distortion rate function"*) as $D(R) = \sigma^2 2^{-2R}$.

Next time, we will finish this computation, give a geometric interpretation of this example and start developing tools like the Method of Types and the notion of conditional typicality. Equipped with those notions, we will prove the Main Result. We will also discuss the construction of concrete schemes for lossy compression and their relation to clustering schemes.