EE276 Information Theory Lecture 14 - 02/22/2024

Lecture 14: Lossy Compression & Rate Distortion Theory Continued

Lecturer: Tsachy Weissman

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 $V_1, V_2, ...V_N/sim$
- **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 $V|U$.

3 Example: Gaussian Source

In the case where $U \sim N(0, \sigma^2)$, $d(u, v) = (u - v)^2$.
Claim:

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

Which is equivalent to:

$$D(R) = \sigma^2 2^{-2R}$$

In particular, the best distortion that we can achieve if we are willing to dedicate only 1 bit per source symbol is $D(1) = \frac{\sigma^2}{4}$. This can be compared to the result at the start of class: $D = \frac{\pi-2}{\pi}\sigma^2 \approx 0.363\sigma^2$ (distortion that we can achieve if we work symbol by symbol and represent every symbol with one bit (see lecture note 2)).

Proof of claim:

For any U, V such that $U \sim N(0, \sigma^2)$ and $\mathbb{E}[(U-V)^2] \leq D$:

$$\begin{aligned}
I(U;V) &= h(U) - h(U|V) \\
&= h(U) - h(U-V|V) \quad \text{differential entropy is invariant under a constant,and V is a constant given V} \\
&\geq h(U) - h(U-V) \quad \text{because conditioning reduces entropy} \\
&\geq h(U) - h(N(0,D)) \quad \text{gaussians maximize differential entropy among distributions} \\
&\qquad\qquad\qquad\qquad \text{with bounded second moment} \\
&= \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log(2\pi eD) \\
&= \frac{1}{2}\log\frac{\sigma^2}{D}
\end{aligned}$$

$\Rightarrow R(D) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$

We can achieve an equality if and only if:

1. $h(U-V|V) = h(U-V)$, i.e. $U-V$ independent from V
2. $h(U-V) = h(N(0,D))$, i.e. $U-V \sim N(0,D)$

Can we find a distribution V that satisfies these two conditions?

The answer is yes. If we take $V \sim N(0, \sigma^2 - D)$ and add an independent Gaussian $G \sim N(0,D)$, we reconstruct U by $U = V + G$.

1. $U - V = G$ is independent of V
2. $U - V = G \sim N(0,D)$

Conclusion: $R(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$ □

4 Interpretation of the Gaussian example

We have found that the minimum achievable rate is $R(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$ when $U \sim N(0,\sigma^2)$.

If we want to visualize this, we can consider the values $U_1, ..., U_n$ as a vector in $\mathbb{R}^n$. Using the law of large numbers and the fact that $\mathbb{E}[U^2] = \sigma^2$, we know that:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} U_i^2 &\approx \sigma^2 \\
\sum_{i=1}^{n} U_i^2 &\approx n\sigma^2 \\
\sqrt{\sum_{i=1}^{n} U_i^2} &\approx \sqrt{n\sigma^2} \\
||U||_2 &\lessapprox \sqrt{n\sigma^2}
\end{aligned}$$

So $(U_i) \in \mathbb{R}^n$ is in a ball centered on 0 and of radius $\sqrt{n\sigma^2}$.

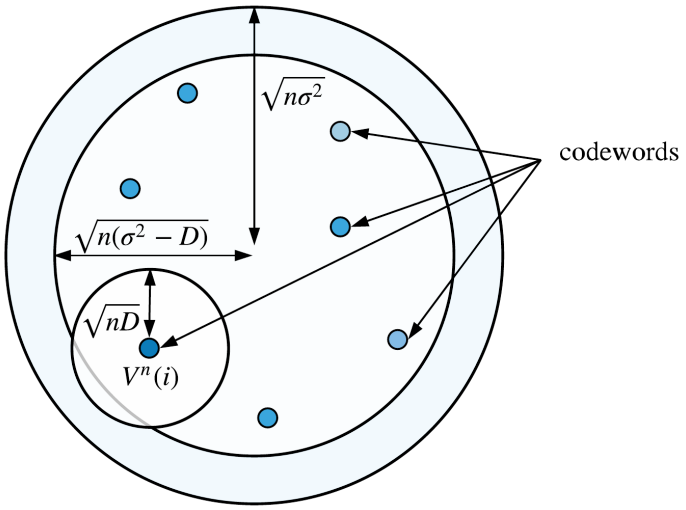

Figure 1: Gaussian example interpretation

Furthermore, we can represent each reconstructed element $V^n(i)$ for $i \in [1, M]$ of the codebook also in the space $\mathbb{R}^n$. As we need to achieve a distortion lower than D, each point $V^n(i)$ can represent points in a ball of radius $\sqrt{nD}$ centered on $V^n(i)$. This is because we have:

$$\begin{aligned} d(U^n, V^n) &\leq D \\ \frac{1}{n}\sum_{i=1}^{n}(U_i - V_i)^2 &\leq D \\ \frac{1}{n}||U - V||_2^2 &\leq D \\ ||U - V||_2 &\leq \sqrt{nD} \end{aligned}$$

Therefore if we want to cover the whole ball of radius $\sqrt{n\sigma^2}$ with these small balls of radius $\sqrt{nD}$, we need the number of points in the codebook M to be:

$$\begin{aligned} M &\geq \frac{Vol(\text{ball of radius } \sqrt{n\sigma^2})}{Vol(\text{ball of radius } \sqrt{nD})} \\ &= \frac{c_n(\sqrt{n\sigma^2})^n}{c_n(\sqrt{nD})^n} \\ &= \left(\frac{\sigma^2}{D}\right)^{n/2} \end{aligned}$$

Because the rate is $R = \frac{m}{n} = \frac{\log M}{n}$, we obtain:

$$\begin{aligned} R &= \frac{\log M}{n} \\ &\geq \frac{1}{2} \log \frac{\sigma^2}{D} \end{aligned}$$

To rephrase, we need at least these M smaller balls to cover the full ball of radius $\sqrt{n\sigma^2}$. In lower dimensions, it looks like there is a lot of overlap between the smaller balls. But in higher dimensions, it is easy to cover the whole space in a very efficient way and achieve the optimal rate $R(D) = \frac{1}{2} \log \frac{\sigma^2}{D}$.

The optimal V we found before is $V \sim N(0, \sigma^2 - D)$, which means we will take the reconstructed codewords $V^n(i)$ iid. on the sphere of radius $\sqrt{n(\sigma^2 - D)}$ to cover the whole space.

Even in low dimensions like $n = 5$ or $n = 6$ we can see that choosing these random codewords is already a very effective scheme.