EE376A/STATS376A Information Theory Lecture 19 - 03/13/2018

Lecture 19: Joint source-channel coding 2

Lecturer: Tsachy Weissman Scribe: Yunfan Wu, Shivam Garg, Pulkit Tandon, Shengjia Zhao, Alex Bertrand

In this lecture, we will review the concepts of joint source-channel coding and give an example of Gaussian source and Gaussian channel. We will also discuss an application of information theory to machine learning.

1 Review of Joint Source-Channel Coding (JSCC)

A quick summary of the concepts

1. **The model:**

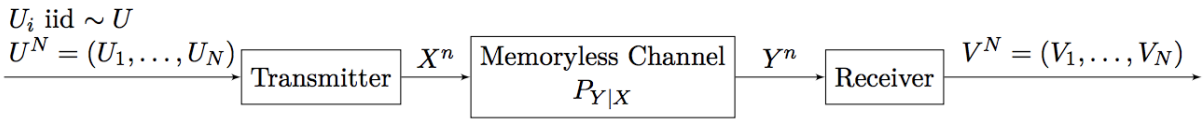

Figure 1: JSCC Problem Schematic

2. **Rate:** rate $= \frac{N}{n}$ $\frac{\text{source symbols}}{\text{channel use}}$
3. **Distortion:** $\mathbb{E}[d(U^N, V^N)]$
4. **Achievability:** (ρ, D) is achievable if $\forall \epsilon > 0$, $\exists$schemes with $\frac{N}{n} \geq \rho - \epsilon$ and $\mathbb{E}[d(U^N, V^N)] \leq D + \epsilon$
5. **"Source-channel Separation" theorem:** (ρ, D) is achievable if and only if $\rho R(D) \leq C$.

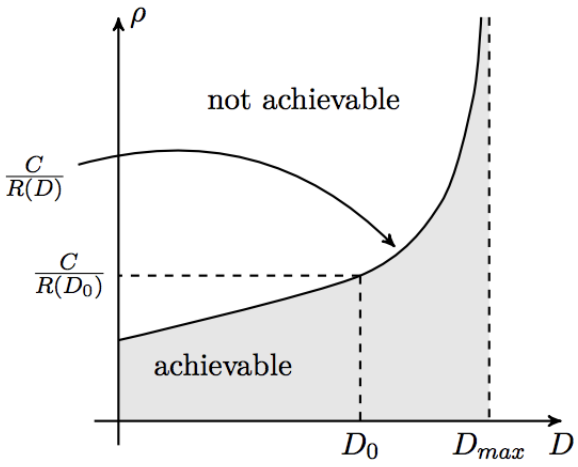

Figure 2: Example Rate Distortion Curve

2 Example: Gaussian source & Gaussian channel

Last class we gave an example of binary source & binary channel, this class we will introduce an example of Gaussian source: $U \sim \mathcal{N}(0, \sigma^2)$ and AWGN channel with transmission power constraint P, the distortion of which is defined as squared error.

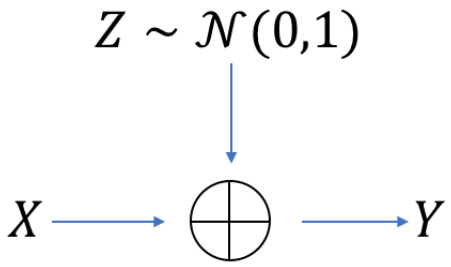

Figure 3: AWGN Channel

Recall: $R(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$, $0 < D \leq \sigma^2$. $C = \frac{1}{2}\log(1+P)$.
Then, by JSCC we get

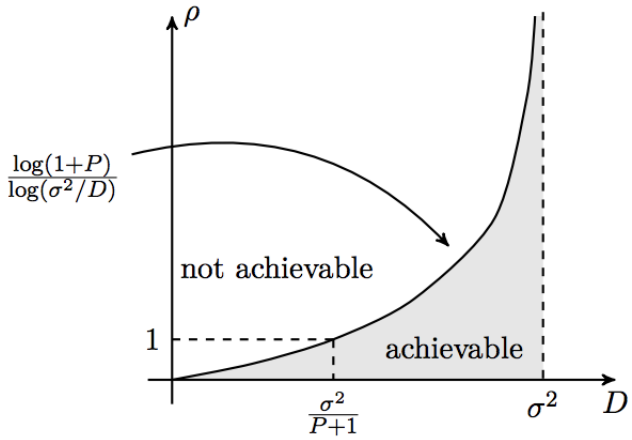

Figure 4: Rate-Distortion Curve for AWGN Channel

Observe that zero distortion is not possible for any positive rate since the source is continuous valued.

Consider the following scheme:
Rate: $\rho = 1$;

Transmit: $X_i = \sqrt{\frac{P}{\sigma^2}}U_i$ (here we rescale U_i because the power of X_i is constrained by P; however, $\text{var}(U) = \sigma^2$. In order to satisfy the power constraint, we rescale U_i to get X_i);

Receive: $Y_i = X_i + Z_i = \sqrt{\frac{P}{\sigma^2}}U_i + Z_i$;

Reconstruction: $V_i = \mathbb{E}[U_i|Y_i] = \frac{\sqrt{P/\sigma^2}\sigma^2}{(\sqrt{P/\sigma^2})^2\sigma^2+1}Y_i$.

Expected distortion achieved: $\mathbb{E}[(U_i - V_i)^2] = \frac{\sigma^2}{(\sqrt{P/\sigma^2})^2\sigma^2+1} = \frac{\sigma^2}{1+P}$

As can be seen, this simple scheme gives the optimum solution. This 'simple scheme' is rather an exception and typically we need non-trivial coding effort when $\rho \neq 1$ in this case and even for $\rho = 1$ with general sources and channels. This is illustrated by the following exercise:

Consider the following "symbol-by-symbol" scheme for $\rho = 1$:
Transmit: $X_i = f(U_i)$;
Reconstruction: $V_i = g(Y_i)$.
Exercise: for a memoryless source & channel, the symbol-by-symbol scheme is optimal **if** $g(\cdot)$ is one-to-one (injective) and $I(U;V)$ achieves $\min_{E[d(U,V)]\leq D} I(U;V)$ and $I(X;Y)$ achieves $\max_{P(X\mid Y)} I(X;Y)$ under the joint distribution of (U, X, Y, V) when $X = f(U)$ and $V = g(Y)$.
Proof Sketch:
Here we have a Markov Chain: $U - X - Y - Z$
Then, by the properties of Markov Chain $I(U;Y) \leq I(X;Y)$, but since $X = f(U)$, we also have by data-processing inequality that $I(X;Y) = I(f(U);Y) \leq I(U;Y)$ and hence we have $I(X;Y) = I(U;Y)$. Now, by using that $g(\cdot)$ is a one-to-one function, we also have $I(U;V) = I(U;g(Y)) = I(U;Y)$. Thus, we have shown that given conditions imply that $I(U;V) = I(X;Y)$. But, based on the conditions in the optimization problem given above, and by our formulations of Channel Coding Theorem and Rate Distortion, we can identify $I(X;Y) = C$ and $I(U;V) = R(D)$. Thus, for these set of conditions, we know by JSSC that there exist a scheme with $\rho = \frac{C}{R(D)} = \frac{I(X;Y)}{I(U;V)} = 1$ which is optimal. Thus, these conditions are sufficient for an optimal "symbol-by-symbol" scheme.

You can apply this exercise and see for yourself that the these conditions exist for both (Binary Source, Binary Channel) example as well as (Gaussian Source, Gaussian Channel) example. For instance, in the second case, we saw above that the "simple scheme" is optimal. For this simple scheme, $X = f(U) = \sqrt{\frac{P}{\sigma^2}}U$ and $X \sim \mathcal{N}(0, P)$, but remember that we have already shown that for Gaussian channel C is achieved when $X \sim \mathcal{N}(0, P)$, that is, $X = f(U)$ in our scheme indeed maximizes $I(X;Y)$. Similarly, all other relations can be established in the two examples.

3 Application of information theory to machine learning

We briefly discuss the application of information theory to machine learning here. For the details of this part, one can refer to the slides: information theory, graphical models and decision trees on the website.

In machine learning problems, usually we are given training data $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, where $X_i \in \mathbb{R}^d$ is the feature vector, Y_i is some label. Our task is to predict Y_i for X_i that we encounter in future. There are two general approaches for this:

- Decision theoretic approach (a.k.a. generative models): learn a probabilistic model of the joint distribution P_{XY}, and then output the most probable label Y_i given X_i under our learned model.
- Learning theoretic approach (a.k.a discriminative model): directly learn a prediction function $f(X)$ with the aim that $\mathbb{E}[L(f(X), Y)]$ is small. Here L denotes the loss function which quantifies how far off is our prediction $f(X)$ from Y.

Now, we discuss a decision theoretic approach that relies on mutual information. For simplicity, let's consider the problem where one gets n i.i.d. samples $x_1, x_2, \cdots, x_n$ from distribution P_X for $x_i \in \mathbb{R}^d$. Given these samples, we want to estimate P_X. One way to do this is by finding a probability distribution Q that maximizes the probability of observing $(x_1, x_2, \cdots, x_n)$. In a previous lecture on method of types, we observed that $Q(x_1, x_2, \cdots, x_n) = 2^{-n(H(\hat{P})+D(\hat{P}||Q))}$ where $\hat{P}$ represents the empirical distribution of the

observed sequence. The Q that maximizes this is the empirical distribution $\hat{P}$ itself. But this is the case when we allow the set of possible Q's to be all the distributions.

Generally, we assume a model (restricted set of distributions) from where Q comes. In this case, our model assumes that Q either factorizes as a tree graphical model, or is reasonably well approximated by one (note that there are plenty of results to suggest that this should be true in most cases). Graphical models are a way to represent the dependencies between various random variables, allowing us to specify one initial distribution and then only conditional distributions for the other (dependent) variables. For more details about them, refer to the slides. Our model doesn't put any restriction on which set of edges are present in the tree. Some calculations (present in the slides) show that the tree that maximizes $Q(x_1, x_2, \cdots, x_n)$ is the maximum weight spanning tree of the complete graph (all edges present) on d vertices (as $x_i \in \mathbb{R}^d$) where the edge weight between two vertices is equal to the empirical mutual information between them. Although the total number of spanning trees of a complete graph is huge, there exist fast (quadratic time) algorithms for finding the maximum weight spanning trees (for example, Kruskal's algorithm or Prim's algorithm). In the next lecture, we will discuss this in more detail.