10-704: Information Processing and Learning **Spring 2015**

Lecture 17: March 24

Lecturer: Aarti Singh *Scribe: Shashank Singh*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

17.1 Overview

In the previous lecture we defined Sufficient Statistics, which capture all information in the data relevant to estimating a desired parameter (hopefully, in a concise way). Motivated by the fact that useful sufficient statistics (i.e., those with dimension independent of sample size) only exist for certain (i.e., exponential family) distributions, we defined two more general notions, the Rate Distortion Function and the Information Bottleneck Principle.

In this lecture, we begin by introducing Channel Coding, a fundamental problem in Information Theory, and presenting the Channel Coding Theorem. We then review the rate distortion function, including the Rate Distortion Theorem, and the Information Bottleneck Method and relate these to Channel Coding.

17.2 Channel Coding

17.2.1 Introduction and Setup

Recall that, in the source coding problem, we had the following model of infomation flow:

$$\text{Source} \xrightarrow{W_1^M} \text{Source Encoder} \xrightarrow{X_1^n} \text{Source Decoder} \xrightarrow{\widehat{W}_1^M} \text{Receiver.}$$

In particular, the decoder received exactly the stream emitted by the encoder, and, given an input distribution $p(X_1^n)$, we were interested in designing a conditional distribution $p(X_1^n|W_1^M)$ minimizing the expected length n of the code (or, more precisely, the limiting ratio $\frac{n}{m}$ as $m \to \infty$).

In the channel coding, we are again given an input string W_1^n which we may encode as X_1^n we wish. This time, however, noise in introduced into X_1^n (according to a known noise distribution), to create a new string $Y_1^{n'}$, and then $Y_1^{n'}$ is given to the decoder, as shown below:

$$\text{Source} \xrightarrow{W_1^M} \text{Channel Encoder} \xrightarrow{X_1^n} \text{Channel} \xrightarrow{Y_1^n} \text{Channel Decoder} \xrightarrow{\widehat{W}_1^M} \text{Receiver.}$$

We will focus on discrete, memoryless channels, where this noise can be encoded as a (known) conditional distribution $p(y|x)$ (i.e., each symbol X_i is mapped to Y_i according to a fixed distribution, so $n' = n$). The goal is then to design an encoding in the form of a distribution $p(x)$ that (asymptotically) achieves low probability of decoding errors $\frac{1}{M} \sum \mathbb{P}\left[\widehat{W}_i \neq W_i\right]$, while again minimizing the length n of the code (or, more precisely, maximizing the rate $\frac{M}{n}$ as $M \to \infty$). The main result, due to Shannon in 1948 [S48], is the Channel Coding Theorem (also known as Shannon's Theorem), which identifies the rates of codes that can achieve low decoding error in terms of the mutual information $I(X;Y)$.

17.2.2 The Channel Coding Theorem

Definition: The *capacity* C of a channel with noise distribution $p(y|x)$ is defined as $C := \max_{p(x)} I(X;Y)$

Example (Binary Symmetric Channel) The binary symmetric channel $BSC(p)$ parametrized by $p \in [0,1]$ has the noise distribution

$$P(y|x) = \left\{ \begin{array}{ll} p & y \neq x \\ 1-p & y = x \end{array} \right. , \quad \text{for all } x, y \in \{0,1\}.$$

That is, $BSC(p)$ flips the input bit with probability p, as illustrated in Figure 17.1.

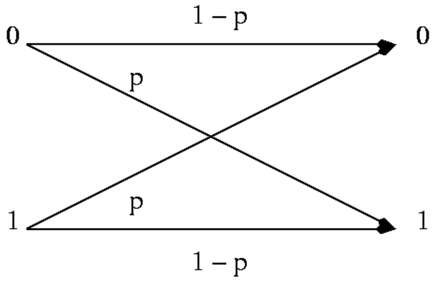

Figure 17.1: A binary symmetric channel with error probability p.

For any input distribution $P(x)$, for $X \sim P(x)$ and $Y \sim P(y|X)$,

$$I(X;Y) = H * Y) - H(Y|X) = H(Y) - h(p) \leq 1 - h(p). \tag{17.1}$$

where $h(p) = -p \log p - (1-p)\log(1-p)$ is the entropy of Bernoulli(p). Furthermore, if $X \sim$ Bernoulli$(\frac{1}{2})$, then by symmetry, then $Y \sim$ Bernoulli$(\frac{1}{2})$, and so $H(y) = 1$. Thus, equality holds in (17.1), and so the capacity of $BSC(p)$ is $C_{BSC(p)} = 1 - h(p)$. □

Definition: Consider a discrete, memoryless channel. A rate R is called *achievable* iff there exists a $(2^{nR}, n)$-channel code with asymptotically vanishing error, i.e.,

$$\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \mathbb{P}\left[\widehat{W_i} \neq W_i\right] = 0.$$

Theorem (Channel Coding): A rate R is achievable if $R < C := \max_{p(x)} I(X;Y)$, and unachievable if $R > C$.

Here, we will prove only the "if" statement (i.e., *achievability*). Later in the course, we will prove the "only if" statement (i.e., *necessity*), which will be useful for proving lower bounds in machine learning problems.

Like the proof of achievability for the source coding theorem, the proof here uses a simple (albeit, impractical) coding scheme based on the notion of *typicality*. Rather than just typicality of the input stream X_1^n, however, we require a stronger condition: *joint typicality* of the joint input and output stream (X_1^n, Y_1^n).

Definition: Given a joint probability density $p : \mathcal{X} \times \mathcal{Y} \to R$ with marginal densities p_X and p_Y, the *jointly typical set* $A_\varepsilon^{(n)} \subseteq (\mathcal{X} \times \mathcal{Y})^n$ is the set of sequences $\{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ such that

1. $\left|-\frac{1}{n}\sum_{i=1}^n \log p_X(x_i) - H(p_X)\right| < \varepsilon$,
2. $\left|-\frac{1}{n}\sum_{i=1}^n \log p_Y(y_i) - H(p_Y)\right| < \varepsilon$,
3. $\left|-\frac{1}{n}\sum_{i=1}^n \log p(x_i, y_i) - H(p)\right| < \varepsilon$.

Theorem (Joint AEP): The jointly typical set $A_\varepsilon^{(n)}$ satisfies

1) For $\{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from p,

$$\lim_{n\to\infty} \mathbb{P}[\{(x_i, y_i)\}_{i=1}^n \in A_\varepsilon^{(n)}] \to 1.$$

2) $|A_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)}$

3) for sequences $\tilde{x}_1, \ldots, \tilde{x}_n \sim p_X$ and $\tilde{y}_1, \ldots, \tilde{y}_n \sim p_Y$ drawn independently,

$$\mathbb{P}\left[\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n \in A_\varepsilon^{(n)}\right] \leq 2^{-n(I(X;Y)-3\varepsilon}.$$

Proof: Property 1) follows from the Weak Law of Large Numbers and a union bound. Property 2) follows from the basic AEP (which we proved with the source coding theorem). Property 3) follows from property 2) and parts 1. and 2. of the definition of $A_\varepsilon^{(n)}$ because

$$\begin{aligned}\mathbb{P}\left[\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n \in A_\varepsilon^{(n)}\right] &= \sum_{\{(x_i,y_i)\}_{i=1}^n \in A_\varepsilon^{(n)}} \prod_{i=1}^n p_X(x_i) p_Y(y_i) \\ &\leq 2^{n(H(X;Y)+\varepsilon)} 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)} = 2^{-n(I(X;Y)-3\varepsilon},\end{aligned}$$

Proof (Achievability of Channel Coding): Suppose $R < C$. *Coding Scheme:* The encoder and decoder operate as follows:

1. Generate 2^{nR} i.i.d. codewords of length n according to the distribution

   $$p(x_1^n) = \prod_{i=1}^n p(x_i).$$

   These strings form the (ordered) codebook $\mathcal{C}$, which is known to both the encoder and the decoder. For $W \in \{1, \ldots, 2^{nR}\}$, we write $X_1^n(W)$ to denote the W^{th} codeword in $\mathcal{C}$.
2. A length nR binary message W is generated uniformly at random (so each $\mathbb{P}[W = w] = 2^{-nR}$).
3. The encoder transmits the codeword $X_1^n(W)$.
4. The decoder receives a noisy codeword Y_1^n from the channel.
5. The decoder outputs an estimated message $\widehat{W} \in \{1, \ldots, 2^{nR}\}$ if $\widehat{W}$ is the *unique* message with $(X_1^n(\widehat{W}), Y_1^n) \in A_\varepsilon^{(n)}$. If $(X_1^n(\widehat{W}), Y_1^n) \notin A_\varepsilon^{(n)}$ or another $\widehat{W}' \in \{1, \ldots, 2^{nR}\}$ also satisfies $(X_1^n(\widehat{W}'), Y_1^n) \in A_\varepsilon^{(n)}$, the decoder outputs $\widehat{W} = 0$ (i.e., it reports failure).

Analysis: If we send M messages independently as described above, then

$$\frac{1}{nR}\sum_{i=1}^{nR}\mathbb{P}\left[W_i \neq \widehat{W_i}\right] = \mathbb{P}\left[W \neq \widehat{W}\right] = \mathbb{P}\left[(X_1^n(W), Y_1^n) \notin A_\varepsilon^{(n)}\right] + \mathbb{P}\left[\exists \widehat{W}' \neq \widehat{W} \text{ with } (X_1^n(\widehat{W}'), Y_1^n) \in A_\varepsilon^{(n)}\right].$$

Part 1) of the joint AEP implies that that the first probability vanishes as $n \to \infty$. For $\widehat{W'} \neq \widehat{W}$, $X_1^n(\widehat{W}')$ was generated independently of $X_1^n(\widehat{W})$ from p (and hence independently of Y_1^n). Thus, applying a union bound, part 3) of the joint AEP implies, for $\varepsilon = \frac{I(X;Y)-R}{6} > 0$,

$$\mathbb{P}\left[\exists \widehat{W}' \neq \widehat{W} \text{ with } (X_1^n(\widehat{W}'), Y_1^n) \in A_\varepsilon^{(n)}\right] \leq \sum_{\widehat{W}' \neq \widehat{W}} \mathbb{P}\left[(X_1^n(\widehat{W}'), Y_1^n) \in A_\varepsilon^{(n)}\right] \leq 2^{nR}2^{-n(I(X;Y)-3\varepsilon)} \to 0,$$

as $n \to \infty$, proving the theorem.

17.3 The Rate-Distortion Theorem

Recall that the Rate Distortion function is

$$R(D) := \inf_{p(t|x)} I(X;T) \quad \text{subject to} \quad \mathbb{E}\left[d(X,T)\right] \leq D. \tag{17.2}$$

In practice, while we can't typically compute the rate distortion function $R(D)$, we can approximate it via the Blahut-Arimoto algorithm ([A72] and [B72]). This doesn't, however, give the optimal code.

Definition: A rate-distortion pair (R, D) is achievable if and only if there exists a $(2^{nR}, n)$ code with

$$\lim_{n\to\infty} \mathbb{E}\left[d(X_1^n, \widehat{X}_1^n)\right] \leq D.$$

Theorem (Rate-Distortion): The rate-distortion function $R(D)$ defined in (17.2) gives the maximum achievable rate at distortion level D.

The proof of the rate-distortion theorem is similar to the proof of the channel coding theorem. The main addition is that the typical set needs one addtional property: *distortion typicality.* In particular, we add the condition $|d(x_1^n, \widehat{x}_1^n) - \mathbb{E}\left[d(X_1^n, \widehat{X}_1^n)\right] \leq \varepsilon$ in order for $(x_1^n, \widehat{x}_1^n)$ to be in $A_\varepsilon^{(n)}$.

Example (Compressing Gaussians): Suppose $X \sim \mathcal{N}(0, \sigma^2)$ (where σ is known) and $d(x,t) = (x-t)^2$. Then, the rate distortion function is

$$R(D) = \left\{ \begin{array}{ll} \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right) & \text{if } D \in (0,\sigma^2) \\ 0 & \text{else} \end{array} \right. .$$

When $D \in (0, \sigma^2)$, we can solve for the distortion in terms of the rate: $D(R) = \sigma^2 2^{-2R}$. Suppose, we use a simple statistic

$$T = \left\{ \begin{array}{ll} \mathbb{E}[X|X \geq 0] & \text{if } X \geq 0 \\ \mathbb{E}[X|X < 0] & \text{else} \end{array} \right. .$$

A straightforward computation gives $\mathbb{E}[X|X \geq 0] = \sqrt{\frac{2}{\pi}}\sigma$, and, symmetrically, $\mathbb{E}[X|X < 0] = -\sqrt{\frac{2}{\pi}}\sigma$, so $T = \text{sign}(X)\sqrt{\frac{2}{\pi}}\sigma$. T can be transmitted using a single bit, and so, according to the rate-distortion theorem, the optimal distortion is $\frac{\sigma^2}{4}$. On the other hand, while T appears to be an optimal 1-bit compression of X,

$$\mathbb{E}[(X-T)^2] = 2\int_0^\infty \left(x - \sqrt{\frac{2}{\pi}}\sigma\right)^2 \phi(x)\,dx = 2\int_0^\infty \left(x^2 - 2x\sqrt{\frac{2}{\pi}}\sigma + \frac{2}{\pi}\sigma^2\right)\phi(x)\,dx$$
$$= \sigma^2 + \frac{2\sigma^2}{\pi} - \frac{4\sigma^2}{\pi} = \sigma^2\left(\frac{\pi-2}{\pi}\right) \approx 0.36\sigma^2 > 0.25\sigma^2.$$

Thus, the distortion is significantly (44% per symbol) higher than optimal. The intuition is that the rate-distorition theorem is an asymptotic result; by transmitting T once per input X, we waste fractional bits that could be used to reduce the average distortion if we encoded long input sequences with a block code.

17.4 Information Bottleneck Method

The information bottleneck method tries to find a statistic T that shares minimal information with the data X, while still conveying information about a parameter Y. It does so by finding a distribution over T, depending on X, which minimizes the following criterion:

$$\min_{p(t|x)} I(X;T) - \beta I(T;Y),$$

where $\beta > 0$ is a Lagrange multiplier. It can be shown that the rate-distortion problem is a special case of the information bottleneck problem. This provides an intuitive link between the problems of channel coding and of finding sufficient statistics.

17.5 Continuous Channels

We now briefly consider the case of a channel operating over an uncountable alphabet. Suppose the channel takes input X outputs $Y = X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of X. In general, by choosing the distribution of X appropriately, we can make $I(X;Y)$ arbitrarily large. For example, if $X \sim \mathcal{N}(0, \eta^2)$, then

$$I(X;Y) = H(Y) - H(Y|X) = \frac{1}{2}\log\left(2\pi e(\sigma^2+\eta^2)\right) - \frac{1}{2}\log\left(2\pi e(\sigma^2)\right) = \frac{1}{2}\log\left(\frac{\sigma^2+\eta^2}{\sigma^2}\right) \to \infty$$

as $\eta \to \infty$. Thus, the usual definition of channel capacity is meaningless. A practical solution is to introduce a *power constraint* of the form $\mathbb{E}[X^2] \leq P$. From the maximum entropy property of the Gaussian distribution, it is easy to see that the capacity of the above channel is then $\frac{1}{2}\log\left(\frac{\sigma^2+P}{\sigma^2}\right)$.

Next time, we will discuss continuous channels further, including applications to areas such as privacy.

References

[S48] C.E. Shannon, "A Mathematical Theory of Communication." Bell System Technical Journal, 1948.

[A72] S. Arimoto, "An Algorithm for Computing the Capacity of Arbitrary DMCs", IEEE Trans. I.T., pp. 14-20, Jan. 1972.

[B72] R. Blahut, "Computation of Channel Capacity and Rate Distortion Functions", IEEE Trans. I.T., pp. 460-473, July 1972.