10-704: Information Processing and Learning **Fall 2016**

Lecture 18: Nov 2

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. 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.*

18.1 Review

In the previous lecture, we defined three related notions - Minimal sufficient statistics originating in Statistics, Information Bottleneck Principle used in Machine learning, and the Rate Distortion Function used in Information Theory. Lets review these quickly.

Minimal sufficient statistics - A transformation $T(X)$ of the data X is a minimal sufficient statistic if

$$T(X) \in \arg\min_{S} I(X, S(X)) \quad \text{where } S(X) \text{ is s.t. } I(\theta, S(X)) = \max_{T'} I(\theta, T'(X)),$$

i.e., it is a transformation of the data X that has smallest mutual information with the data X while having largest mutual information with θ.

Information Bottleneck Principle - generalizes the notion of minimal sufficient statistics and suggests using a summary of the data $T(X)$ that has smallest mutual information with the data X while preserving some amount of information about an auxiliary variable Y:

$$T(X) \in \arg\min_{S(X)} I(X, S(X)) \quad \text{where } S(X) \text{ is s.t. } I(Y, S(X)) \geq \gamma$$

or equivalently,

$$T(X) \in \arg\min_{S(X)} \left[I(X, S(X)) - \beta \; I(Y, S(X)) \right]$$

for some $\beta \geq 0$. Thus, $T(X)$ serves as an information bottleneck between X and Y, extracting information from X that is relevant to Y.

Rate-distortion function - gives the limit of lossy compression, i.e. the smallest number of bits to which we can compress X while allowing distortion no more than D in some distortion metric $d(X, T)$, where T is the compressed version of data X.

$$R(D) = \min_{T} I(X, T) \quad \text{where } T \text{ is s.t. } \mathbb{E}[d(X, T)] \leq D. \tag{18.1}$$

Remark: The information bottleneck principle is essentially equivalent to lossy compression under the distortion function $d(X, T) = KL(p(Y|X)||p(Y|T))$.

We also saw how to compute the rate-distortion function or the value of the information bottleneck objective using an iterative procedure known as Blhaut-Arimoto Algorithm. For Rate-distortion function, we assume we are given $p(x)$ and we optimize over $p(t|x)$. The objective is convex in $p(t|x)$ and we initialize $p(t|x)$ to some value and then iterate over the following two updates:

$$p(t) = \sum_{x} p(t|x)p(x)$$

and

$$p(t|x) \propto p(t)e^{-\beta d(x,t)}.$$

For information bottleneck, we assume we are given $p(x)$ and $p(y|x)$. The objective is convex in $p(t|x)$ and we initialize $p(t|x)$ to some value and then iterate over the following three updates:

$$p(t) = \sum_x p(t|x)p(x)$$

and

$$p(y|t) = \frac{p(y,t)}{p(t)} = \frac{1}{p(t)} \sum_x p(y|x,t)p(t|x)p(x) = \frac{1}{p(t)} \sum_x p(y|x)p(t|x)p(x)$$

and

$$p(t|x) \propto p(t)e^{-\beta KL(p(y|x)||p(y||t))}.$$

Remark: Blhaut-Arimoto algorithm converges to a global minimum since the objectives are convex.

Remark: Blhaut-Arimoto algorithm enables us to calculate the optimal value of the information bottleneck objective and the rate-distortion function, but does not tell us the optimal statistic (though one could sample from $p(t|x)$) or do the optimal encoding.

Remark: If $p(x)$ and $p(y|x)$ are not known, then these could be treated as latent variables and we could optimize over these similar to EM (Expectation-Maximization) algorithm but the solution is only guaranteed to converge to local optima in that case, since the problem is no longer convex.

18.2 Rate-Distortion Theorem

Next, we establish the rate-distortion theorem which establishes that the above definition of rate distortion function indeed characterizes the limit of lossy compression.

Before, we state the theorem, we define a length n block code with rate R.

Definition: A length n block code is a mapping from a stream of symbols $x_1^n = x_1 \dots x_n$ to a codeword $c(x_1^n)$ and the rate of the code is $R = \mathbb{E}[\ell(X_1^n)/n]$, the average number of bits per symbol.

Definition: A rate-distortion pair (R, D) is achievable if and only if there exists a rate R code such that the reconstruction $\widehat{X}_1^n$ based on the code $c(X_1^n)$ satisfies

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

Theorem (Rate-Distortion): The rate-distortion function $R(D)$ defined in (18.1) gives the minimum achievable rate at distortion level D.

The achievability proof of the rate-distortion theorem is similar to the proof of the source coding theorem. The main difference is that we define a distortion-typical set A_ε^n which consists of pairs of sequences that are each typical, and moreover satisfy two additional properties: (1) $|d(x_1^n, \widehat{x}_1^n) - \mathbb{E}\left[d(X_1^n, \widehat{X}_1^n)\right]| \leq \varepsilon$ in order for $(x_1^n, \widehat{x}_1^n)$ to be in $A_\varepsilon^{(n)}$. (2) $|-\frac{1}{n}\log p(x_1^n, \widehat{x}_1^n) - H(X, \widehat{X})| \leq \varepsilon$. We refer to sec 10.5 of Cover-Thomas for the details.

The codebook in the proof is again random and based on an exponential lookup into 2^{nR} codewords and hence is not practical. Design of practical lossy compression methods that achieve close to rate-distortion performance with finite block-lengths is the focus of research into coding schemes, specifically vector quantization methods we briefly referred to in last lecture.

18.3 Channel Coding

Recall that, in the source coding problem, we had the following model of information flow (switching notation a bit):

$$\text{Source} \xrightarrow{W_1^m} \text{Source Encoder} \xrightarrow{c(W_1^m)} \text{Source Decoder} \xrightarrow{\widehat{W}_1^m} \text{Receiver}.$$

In particular, the decoder received exactly the stream transmitted by the encoder, and, given an input distribution $p(W_1^m)$, we were interested in designing a conditional distribution $p(\widehat{W}_1^m | W_1^m)$ minimizing the expected length of the code.

In the channel coding problem, we are again given an input string W_1^m which we may encode as $X_1^n = c(W_1^m)$. This time, however, noise in introduced into X_1^n, 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 first 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 $n' = n$), i.e. Y_i only depends on current channel input X_i and not on past (or future) inputs. The goal is then to design an encoding in the form of a distribution $p(x)$ that (asymptotically) achieves low probability of decoding errors $\mathbb{P}\left[\widehat{W} \neq W\right]$, 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, which identifies the rates of codes that can achieve low decoding error in terms of the mutual information $I(X;Y)$.

18.3.1 The Channel Coding Theorem

Definition: Consider a discrete, memoryless channel. A rate R is called *achievable* iff there exists a rate R channel code with asymptotically vanishing error, i.e.,

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

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

Theorem (Channel Coding): The capacity of a channel is the maximum achievable rate.

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 n 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

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

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.

Lets now try to understand the capacity of a channel via some examples.

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

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

That is, $BSC(p)$ flips the input bit with probability p.

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

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - h(p) \leq 1 - h(p). \tag{18.2}$$

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 (18.2), and so the capacity of $BSC(p)$ is $C_{BSC(p)} = 1 - h(p)$. □

18.3.2 Continuous Channels

We now consider the case of continuous channels using Gaussian channel as an example.

Example (Single Gaussian Channel) 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 $\frac{1}{2}\log\left(\frac{\sigma^2+P}{\sigma^2}\right)$.

Example (Parallel Independent Gaussian Channels) Now consider the multi-dimensional case. Suppose our input, output and noise now are in $\mathbb{R}^n$ space, and Z to is draw from $N(0, \sigma^2 I_{n\times n})$. We consider three settings:

- **Individual power constraint, same noise variance.** We have independent power constraint for each sub channel as $E[X_i^2] \leq P$ for $i = 1, \ldots, n$. Then the capacity is directly n times the capacity of one channel.

  Theorem 18.1 *Suppose $Y = X + Z$, $X, Y, Z \in \mathbb{R}^n$ is a multivariate Gaussian channel and $Z \sim N(0, \sigma^2 I)$. Given independent power constraint on each X_i, i.e. $E[X_i^2] \leq P$. Then the capacity of this channel is given by*

  $$\textit{Capacity} = \frac{n}{2}\log(1 + \frac{P}{\sigma^2})$$

- **Global power constraint, same noise variance.** Now consider the case if a global power constraint like $E[||X||^2] \leq P$ is used. With same condition as above, we can prove the capacity is maximized when we equally distribute the power constraint over all channels, i.e. $P_i = \frac{P}{n}$ and the capacity is given as follows.

Theorem 18.2 *Suppose $Y = X + Z$, $X, Y, Z \in \mathbb{R}^n$ is a multivariate Gaussian channel and $Z \sim N(0, \sigma^2 I)$. Given universal power constraint over X, i.e. $E[||X||^2] \leq P$. Then the capacity of this channel is given by*

$$Capacity = \frac{n}{2} \log(1 + \frac{P}{n\sigma^2})$$

- **Global power constraint, different noise variance.** In above examples, our input channels have independent noise with same variance. If our noise is still independent on each channel, but with different variance, the maximum capacity is given through a "water filling" way.

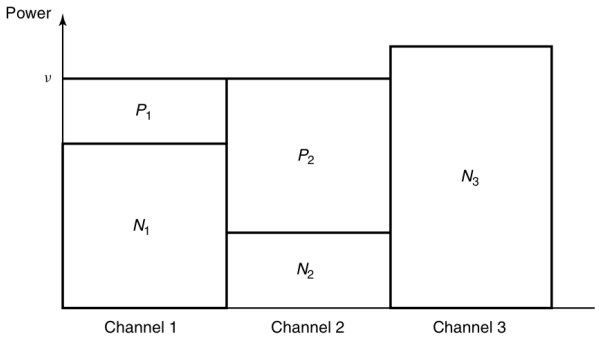

Figure 18.1: Water filling, figure from [Cover2012]

Theorem 18.3 *Suppose $Y = X+Z$, $X, Y, Z \in \mathbb{R}^n$ is a multivariate Gaussian channel, $Z \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2))$. Given universal power constraint, i.e. $E[||X||^2] \leq P$, then the capacity of this channel is given through a "water-filling" way. That is, the power allocated for each channel P_i is $(constant - \sigma_i^2)^+$, where the constant is chosen so that the total power $\sum_i P_i$ is P.*

Proof: We know capacity C is defined as

$$\begin{aligned}
C &= \max_{p(x)} I(X^n, Y^n) \\
&= \max_{p(x)} (H(Y^n) - H(Y^n|X^n)) \\
&= \max_{p(x)} (H(Y^n)) - H(Z^n), \quad \text{Since } Y = X + Z \text{ and } X \perp Z \\
&= \max_{p(x)} \sum_{i=1}^{n} H(Y_i) - H(Z_i), \quad \text{since } Y_i\text{s are independent and so are } Z_i\text{s}
\end{aligned}$$

We know $Y_i = X_i + Z_i$, so $E[Y_i^2] = E[(X_i + Z_i)^2] = E[X_i^2] + E[Z_i^2] = P_i + \sigma^2$. We know for a given variance, normal distribution maximize the entropy, thus $H(Y_i) \leq \frac{1}{2} \log 2\pi e(P_i + \sigma_i)$.

$$\begin{aligned}
&\leq \max_{\{P_i\}_{i=1}^n} \frac{1}{2} \sum_{i=1}^{n} \log 2\pi e(P_i + \sigma_i^2) - \frac{1}{2} \sum_{i=1}^{n} \log 2\pi e \sigma_i^2 \\
&= \max_{\{P_i\}_{i=1}^n} \frac{1}{2} \sum_{i=1}^{n} \log \left(1 + \frac{P_i}{\sigma_i^2}\right)
\end{aligned}$$

Since the $P_i \geq 0$ and $\sum_i P_i \leq P$, the Lagrangian multiplier of the above optimization problem is

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}\sum_{i=1}^{n}\log\left(1+\frac{P_i}{\sigma^2}\right)+\lambda(\sum_i P_i - P)+\lambda_i P_i \\ \frac{\partial \mathcal{L}}{\partial P_i} =& \frac{1}{2}\frac{1}{1+\frac{P_i}{\sigma_i^2}}\cdot\frac{1}{\sigma_i^2}+\lambda+\lambda_i = 0 \\ \Rightarrow P_i+\sigma_i^2 =& constant, \forall i \quad \text{s.t.} \lambda_i = 0\end{aligned}$$

Since complementary slackness implies that either $P_i = 0$ or $\lambda_i = 0$, the solution is either to put no power in a channel or to put enough power so that the sum of power and noise variance is a constant for all channels with non-zero power. Thus, we are putting more power to less noisy channels through a water filling way, where we first try to add power to least noisy channels until its "height" is same with the second least one, and continue until all power is allocated. ■

In next class, we will evaluate the capacity of correlated channels and the multi-antenna channel capacity.

18.3.3 General channels

For general distributions, it may not be possible to analytically calculate the capacity of a channel. Interestingly, we can still use a variation of the Blhaut-Arimoto algorithm for calculating capacity since mutual information is concave in $p(x)$ and we can write the capacity as a dual maximization:

$$\max_{p(x)} I(X,Y) = \max_{p(x)}\sum_{x,y} p(x)p(y|x)\log\frac{p(x|y)}{p(x)} = \max_{p(x)}\max_{q(x|y)}\sum_{x,y} p(x)p(y|x)\log\frac{q(x|y)}{p(x)}$$

Now, we are given $p(y|x)$ which characterizes the channel, and an iterative procedure that does maximization over $p(x)$ and then over $q(x|y) = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}$ can be derived as follows: Initialize $p(x)$. For fixed $p(x)$, update

$$q(x|y) = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}.$$

This update step is straight-forward. Now fix $q(x|y)$ and update $p(x)$ as

$$p(x) = \frac{\prod_y q(x|y)^{p(y|x)}}{\sum_{x'}\prod_y q(x'|y)^{p(y|x')}}.$$

To derive this step, differentiate objective wrt $p(x)$ for fixed $q(x|y)$.

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.