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

# Lecture 17: Oct 31

*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.*

So far in the course, we have talked about lossless compression. In today's class, we will begin talking about lossy compression. We will also make connections to sufficient statistics and information bottleneck principle in machine learning.

## 17.1 Sufficient statistic

We start by reviewing what we discussed about sufficient statistics in last lecture. Sufficient statistics provide a (usually short) summary of the data, and are often useful for learning tasks.

A **statistic** is any function $T(X)$ of the data $X$. if $\theta$ parametrizes the class of underlying data-generating distributions, then for any statistic, we have the Markov chain

$$\theta \to X \to T(X)$$

i.e. $\theta \perp T(X)|X$ and data processing inequality tells us that $I(\theta, T(X)) \leq I(\theta, X)$.

A statistic is **sufficient** for a parameter $\theta$ if $\theta \perp X|T(X)$, i.e. we also have the Markov chain

$$\theta \to T(X) \to X.$$

In words, once we know $T(X)$, the remaining randomness in $X$ does not depend on $\theta$. This implies $p(X|T(X))$ does not depend on $\theta$ (this is a useful characterization when you may not want to think of $\theta$ as a random variable). Also, $I(\theta, T(X)) = I(\theta, X)$.

How do we find a sufficient statistic?

**Theorem 17.1 (Fisher-Neyman Factorization Theorem)** *$T(X)$ is a sufficient statistic for $\theta$ iff $p(X;\theta) = g(T(X), \theta)h(X)$.*

Here $p(X;\theta)$ is the joint distribution if $\theta$ is random, or is the likelihood of data under $p_\theta$ otherwise.

A sufficient statistic $T(X)$ is **minimal** if $T(X) = g(S(X))$ for all sufficient statistics $S(X)$. For example, if $X = X_1, \ldots, X_n \sim \text{Ber}(\theta)$, then $T(X) = \frac{1}{n}\sum_{i=1}^n X_i$ is a minimal sufficient for $\theta$ but $S(X) = X$ is not.

*Note:* A minimal sufficient statistic is not unique. For the above example, both $\frac{1}{n}\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i$ are minimal sufficient statistic.

Lets mention an alternate characterization of a sufficient statistic and minimal sufficient statistic.

- If $T(X)$ is sufficient, then

  $$I(\theta, T(X)) = \max_{T'} I(\theta, T'(X)) \qquad (= I(\theta, X))$$

  i.e. it is a statistic that has largest mutual information with $\theta$ among all other statistics $T'(X)$.

- If $T(X)$ is minimal sufficient statistic, then

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

  i.e. it is a statistic that has smallest mutual information with $X$ while having largest mutual information with $\theta$.

The following theorem limits interest in sufficient statistics for complex learning tasks.

**Theorem 17.2 (Pitman-Koopman-Darmois Theorem)** *For non-compactly supported distributions, exact sufficient statistics with bounded (not growing with $n$) dimensionality exist only for distributions in the exponential family.*

## 17.2 Information Bottleneck Principle

The **Information Bottleneck (IB) Principle** is an information-theoretic approach to machine learning that generalizes the notion of sufficient statistic. The IB principle suggests using a summary $T(X)$ of the data $X$ that preserves some amount of information about an auxiliary variable $Y$:

$$\min_{T} I(X,T) \quad s.t. \quad I(Y,T) \geq \gamma \qquad \equiv \qquad \min_{T} I(X,T) - \beta I(Y,T)$$

where $\beta \geq 0$. Thus, $T$ serves as an information bottleneck between $X$ and $Y$, extracting information from $X$ that is relevant to $Y$. It can be thought of as dimensionality reduction. As $\beta \to \infty$, $T$ preserves all information in $X$ relevant to $Y$ and becomes a minimal sufficient statistic for $Y$ based on data $X$.

Some examples of application of IB method in machine learning:

1. Speech compression - ideally, the speech waveform can be compressed down to its entropy. But in many applications, it is possible to compress further without losing important information about words, meaning, speaker identity, etc. The auxiliary variable $Y$ can be transcript of signal for speech recognition, part of speech for grammar checking, speaker identity, etc.

2. Protein folding - Given the protein sequence data, we might compress it only preserving information about its structure.

3. Neural coding - Given neural stimuli, we might seek a summary relevant to characterizing the neural responses.

4. Unsupervised problems - In clustering documents, $X$ can be document identity, $Y$ can be words in the document and $T$ will be clusters of documents with similar word statistics.

5. Supervised problems - In document classification, $X$ can be words in the document, $Y$ the topic labels and $T$ will be clusters of words that are sufficient for document classification.

Another viewpoint explored in [1] is to consider $I(Y,T)$ as the empirical risk and $I(X,T)$ as a regularization term. As $\beta \to \infty$, we are maximizing $I(Y,T)$ and overfitting to the data $X$, while for lower $\beta$ this is prevented as we minimize $I(X,T)$.

The IB method can also be considered as lossy compression of the data $X$ where the degree of information lost about $Y$ is controlled by $\beta$. In [2], it is argued that IB optimization provides a more general framework that rate distortion function used in lossy compression. We discuss this next.

[1] http://www.cs.huji.ac.il/labs/learning/Papers/ibgen.pdf

[2] http://www.cs.huji.ac.il/labs/learning/Papers/allerton.pdf

## 17.3 Lossy Compression and Rate distortion function

We have seen that for lossless compression, we can only compress a random variable $X$ down to entropy number of bits. What if we allowed some loss or distortion in how accurately we can recover $X$? This question is particularly inevitable for continuous random variables which cannot be compressed to infinite precision with finite number of bits. Thus, we talk about lossy compression, also known as rate distortion theory.

The rate-distortion function characterizes the smallest number of bits to which we can compress $X$ while allowing distortion $D$ in some distortion metric $d(X,\hat{X})$:

$$R(D) = \min_{\hat{X}} I(X,\hat{X}) \quad s.t. \quad \mathbb{E}[d(X,\hat{X})] \le D$$

Observe that if we set $D = 0$ i.e. lossless compression is desired, then the rate distortion function $R(0) = I(X,X) = H(X)$, the entropy.

Examples of distortion functions are:

- Squared error distortion: $d(X,\hat{X}) = (X-\hat{X})^2$
- Hamming distortion: $d(X,\hat{X}) = 1_{X \neq \hat{X}}$.

The IB principle can be considered as a more general framework than rate-distortion since in many applications (such as speaker identification) there may not be a natural distortion function and it is easier to state the goal as preserving information about an auxiliary variable $Y$ (speaker identity). However, as we will see in the next section, the IB method implicitly assumes the KL divergence as a distortion function.

Example of rate-distortion function:

- Gaussian source $\mathcal{N}(0,\sigma^2)$ and squared-error distortion

  We first find a lower bound for the rate distortion function

$$\begin{aligned} I(X,\hat{X}) &= H(X) - H(X|\hat{X}) \\ &\ge \frac{1}{2}\log(2\pi e)\sigma^2 - H(X-\hat{X}) \\ &\ge \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)D \\ &= \frac{1}{2}\log\frac{\sigma^2}{D} \end{aligned}$$

where $\sigma^2 \geq D$. Note that we used the fact that conditioning reduces entropy (1st inequality), and the maximum entropy under the second moment constraint $E[(X-\hat{X})^2] \leq D$ is achieved by the Gaussian distribution (2nd inequality). The lower bound can be achieved by the simple construction in Fig. 17.1 with $\hat{X} \sim \mathcal{N}(0, \sigma^2 - D)$ and $Z \perp \hat{X}$. It is easy to verify that $H(X|\hat{X}) = \frac{1}{2}\log(2\pi e)D$ and $E[(X-\hat{X})^2] = D$.

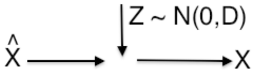

Figure 17.1: Joint distribution for Gaussian source.

If $\sigma^2 < D$, i.e. we are allowed mean square distortion larger than $\sigma^2$, then if we don't encode $X$ at all, i.e. $R = 0$, we can achieve mean square distortion $D$ trivially.

The rate distortion function is

$$R(D) = \min_{p(\hat{x}|x):E[d(X,\hat{X})]\leq D} I(X;\hat{X}) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & \text{if } D \leq \sigma^2 \\ 0, & \text{if } D > \sigma^2 \end{cases}$$

Notice that this only helps characterize the rate-distortion function (limit of lossy compression), but does not tell us how to do the encoding. Moreover, this limit can only be achieved asymptotically, i.e. by encoding large blocks of symbols, not individual symbols as the following result shows.

**Example (Compressing Gaussians):** 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

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

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 $X$,

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

Thus, the distortion is significantly (44% per symbol) higher than optimal. The intuition is that the rate-distoritio 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, i.e. do vector quantization.

## 17.4 Blahut-Arimoto algorithm for IB and R(D)

In some cases, such as Gaussian example above, the rate-distortion function can be computed explicitly. More generally, a closed form solution may not exist and we resort to iterative methods. The Blahut-Arimoto algorithm provides an iterative solution to both rate-distortion function and information bottleneck (and also channel capacity as we will see in next lecture). The solution will also reveal that IB is essentially using KL divergence as the distortion measure.

Lets first consider the rate-distortion function. The objective that we need to minimize can be written as (in analogy with the IB method):

$$I(X,T) + \beta \mathbb{E}_{X,T}[d(X,T)]$$

where $\beta \geq 0$, which we will optimize over $p(t|x)$ which is a probability distribution (integrates to 1 and is positive). The Lagrangian can be written as:

$$\mathcal{L} = I(X,T) + \beta \mathbb{E}_{X,T}[d(X,T)] + \sum_x \lambda_x [\sum_t p(t|x) - 1] - \nu_{x,t} p(t|x)$$

where $\nu_{x,t} \geq 0$.

The iterative approach starts with an initial guess $p(t)$ and iterates over updating $p(t|x)$ and $p(t)$.

The update equation for $p(t)$ is

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

First recall

$$\begin{aligned} I(X,T) &= \sum_{x,t} p(x,t) \log \frac{p(x,t)}{p(x)p(t)} \\ &= \sum_{x,t} p(t|x)p(x) \log \frac{p(t|x)}{p(t)} \\ &= \sum_{x,t} p(t|x)p(x) \log p(t|x) - \sum_{x,t} p(t|x)p(x) \log p(t) \\ &= \sum_{x,t} p(t|x)p(x) \log p(t|x) - \sum_t p(t) \log p(t) \end{aligned}$$

Taking derivative of $I(X,T)$ for fixed $x, t$

$$\begin{aligned} \frac{\partial I(X,T)}{\partial p(t|x)} &= p(x) \log p(t|x) + p(x) - \frac{\partial p(t)}{\partial p(t|x)} [\log p(t) + 1] \\ &= p(x)[\log p(t|x) + 1 - \log p(t) - 1] \\ &= p(x) \log \frac{p(t|x)}{p(t)} \end{aligned}$$

Also,

$$\frac{\partial \mathbb{E}_{X,T}[d(X,T)]}{\partial p(t|x)} = d(x,t)p(x)$$

Thus, we have the derivative of the Lagrangian is:

$$\frac{\partial \mathcal{L}}{\partial p(t|x)} = p(x) \left[ \log \frac{p(t|x)}{p(t)} + \beta d(x,t) + \frac{\lambda_x + \nu_{x,t}}{p(x)} \right]$$

Setting it equal to 0, we get the update equation for $p(t|x)$:

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

and coupled with the update equation for $p(t)$:

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

yields the Blahut-Arimoto algorithm for calculating the rate-distortion function.

*Note:* The Blahut-Arimoto algorithm only helps calculate the rate-distortion function and does not provide a way to do lossy compression.

*Note:* The algorithm assumes $p(x)$ is given. One can also jointly optimize over $p(x)$ and $p(t|x)$ in an EM-like fashion.

Similarly, we can derive an iterative algorithm for the IB method where the objective is

$$I(X,T) - \beta I(Y,T)$$

where $\beta \geq 0$, which we will optimize over $p(t|x)$ which is a probability distribution (integrates to 1 and is positive). The Blahut-Arimoto algorithm for IB yields the following update equations (assuming $p(x)$ and $p(y|x)$ are known):

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

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

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

last step follows since $Y$ is independent of $T$ given $X$.

Thus, the effective distortion measure in IB is $d(x,t) = KL(p(y|x)||p(y|t))$, the KL divergence between conditional distribution of $Y$ given $X$ and given $T$.