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

# Lecture 2: January 15

*Lecturer: Akshay Krishnamurthy* *Scribe: Kyle Soska*

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



## 2.1 Brief Review

In the previous class, we defined the following:

- The **Shannon Information Content** of an outcome $X$ which occurs with probability $p(X)$ is

$$\log_2 \frac{1}{p(x)}.$$

- The **entropy** in bits is the average uncertainty of a random variable $X$:

$$H(X) = \sum_{x \epsilon \mathcal{X}} p(x) \log_2 \frac{1}{p(x)} = -\mathbb{E}_{X \sim p}[\log_2 p(X)]$$

- The **joint entropy** in bits of two random variables $X, Y$ with join distribution $p(x, y)$ is

$$H(X, Y) = \sum_{x \epsilon \mathcal{X}, y \epsilon \mathcal{Y}} p(x, y) \log_2 \left( \frac{1}{p(x, y)} \right)$$

- The **conditional entropy** in bits of $Y$ conditioned on $X$ is the average uncertainty about $Y$ after observing $X$.

$$H(Y|X) = \sum_{x \epsilon \mathcal{X}} p(x) H(Y|X = x) = \sum_{x \epsilon \mathcal{X}} p(x) \sum_{y \epsilon \mathcal{Y}} p(y|x) \log_2 \left( \frac{1}{p(y|x)} \right)$$

- Given two distributions $p, q$ for a random variable $X$, the **relative entropy** between $p$ and $q$ is

$$D(p||q) = \mathbb{E}_{X \sim p}[\log \left( \frac{1}{q(X)} \right)] - \mathbb{E}_{X \sim p}[\log \left( \frac{1}{p(X)} \right)] = \mathbb{E}_p[\log \left( \frac{p}{q} \right)] = \sum_x p(x) \log \left( \frac{p(x)}{q(x)} \right)$$

- The **Mutual Information** between $X$ and $Y$ is

$$I(X; Y) = D\left(p(x, y)||p(x)p(y)\right)$$

## 2.2 Fundamental Limits in Information Theory

The **source coding** model is as follows:

$$\text{Source Model} \xrightarrow{X_1..X_n} \boxed{\text{Compressor}} \xrightarrow{b_1,\ldots,b_m bits} \boxed{\text{Decompresser}} \rightarrow \text{Receiver}$$

Let the data source be generated according to some distribution $p(X)$. If the average number of bits used to encode one source symbol is less than the source entropy $H(X)$, that is $\mathbb{E}_p[\frac{\text{codelength}}{\#\text{src symbols}}] < H(X)$ then perfect reconstruction is not possible. A distribution cannot be compressed below its entropy without loss.

The **channel coding** model is as follows:

$$\text{Source} \rightarrow \boxed{\text{Encoder}} \xrightarrow{X} \boxed{\text{Channel } p(y|x)} \xrightarrow{Y} \boxed{\text{Decoder}} \rightarrow \text{Receiver}$$

If the data source is generated according to $p$, the rate of the code is defined as $\mathbf{E}_p \frac{\#\text{ source symbols}}{\#\text{ message bits}}$. If the rate of the code is greater than the channel capacity $C = max_{p(X)} I(X,Y)$, then perfect reconstruction is not possible.

The **inference** problem is similar to the channel coding problem except we do not design the encoder:

In the regression setting with $f \epsilon \mathcal{F}$:

$$\xrightarrow{f(X_1),\ldots,f(X_n)} \boxed{\text{Channel: } p(y|f(x))} \xrightarrow{Y_1,\ldots,Y_n} \boxed{\text{Decoder}} \xrightarrow{f}$$

In the density estimation setting with $p_\theta, \theta \epsilon \Theta$:

$$\theta \rightarrow \boxed{\text{Channel: } p_\theta(X)} \xrightarrow{X_1,\ldots,X_n} \boxed{\text{Decoder}} \xrightarrow{\hat{p}}$$

Under log loss:

$$\text{Excess Risk}(q) = \text{Risk}(q) - \text{Risk}(p) = D(p||q)$$

Fundamental limits of inference problems are often characterized by minmax lower bounds, i.e. the smallest possible excess risk that any estimator can achieve for a class of models. For the density estimation problem, the minmax excess risk is $\inf_q \sup_{p \epsilon \mathcal{P}} D(p||q)$ and we will show that this is equal to the capacity $C$ of the corresponding channel. This would imply that for all estimators q, $\sup_{p \epsilon \mathcal{P}} D(p||q) \geq C$.

We will state and prove these results formally later in the course. Information theory will help us identify these fundamental limits of data compression, transmission and inference; and in some cases also demonstrate that the limits are achievable. The design of efficient encoders / decoders / estimators that achieve these limits is the common objective of Signal Processing and Machine Learning algorithms.

## 2.3 Useful Properties of Information Quantities

1. Chain Rule: $H(X,Y) = H(X) + H(Y|X)$

2. Entropy is always non-negative: $H(X) \geq 0, H(X) = 0 \Leftrightarrow X$ is constant
   **Proof:** $0 \leq p(x) \leq 1$ implies that $\log \frac{1}{p(x)} \geq 0$ ■

   For example, consider a binary random variable $X \sim Bernoulli(\theta)$ Then $\theta = 0$ or $\theta = 1$ implies that $H(X) = 0$. If $\theta = \frac{1}{2}$, then $H(X) = 1$ (which is the maximum entropy for a binary random variable since the distribution is uniform).

3. (**Gibbs Information Inequality**) $D(p||q) \geq 0, = 0$ if and only if $p(x) = q(x)$ for all $x$.
   **Proof:** Define the support of $p$ to be $\mathcal{X} = \{x : p(x) > 0\}$

$$\begin{aligned} -D(p||q) &= -\sum_{x \epsilon \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \epsilon \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \\ &\leq \log \sum_{x \epsilon \mathcal{X}} p(x) \frac{q(x)}{p(x)} \\ &= \log \sum_{x \epsilon \mathcal{X}} q(x) \leq \log 1 = 0 \end{aligned}$$

   The first inequality is Jensen's inequality since log is concave. Because log is strictly concave we have equality in the first inequality only if $p$ is a constant distribution or if $\frac{q(x)}{p(x)}$ is a constant c, for all x (i.e. if $q(x) = cp(x)$). The second inequality is tight only when that constant $c = 1$ since $\sum_{x \epsilon \mathcal{X}} p(x) = 1$. ■

4. As a corollary, we get that $I(X, Y) = D(p(x, y)||p(x)p(y)) \geq 0$ and $= 0$ iff $X, Y$ are independent, that is, $p(x, y) = p(x)p(y)$.

5. $H(X) \leq \log |\mathcal{X}|$ where $\mathcal{X}$ is the set of all outcomes with non-zero probability. Equality is achieved iff $X$ is uniform.

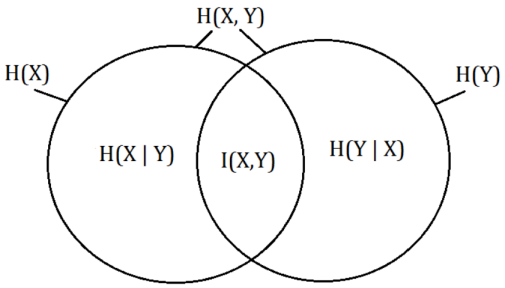
   **Proof:** Let u be the uniform distribution over $X$, i.e. $u(x) = \frac{1}{|\mathcal{X}|}$ and let $p(x)$ be the probability mass function for $X$.

$$D(p||u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X)$$

$$0 \leq D(p||u) = \log |\mathcal{X}| - H(X) \text{ by non negativity of relative entropy}$$

   ■

6. The following relations hold between entropy, conditional entropy, joint entropy, and mutual information:



(a) H(X,Y) = H(X) + H(Y|X) = H(Y) + H(Y|X)
(b) I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y,X)
(c) I(X,Y) = H(X,Y) - H(X|Y) - H(Y|X)
(d) I(X,Y) = H(X) + H(Y) - H(X,Y)

7. Conditioning cannot increase entropy, i.e. information always helps.

$$H(X|Y) \leq H(X)$$

with equality iff X and Y are independent.

**Proof:** $0 \leq I(X;Y) = H(X) - H(X|Y)$ ■

## 2.4 Data Processing Inequality

Intuitively, the data processing inequality says that no clever transformation of the received code (channel output) $Y$ can give more information about the sent code (channel input) $X$ than $Y$ itself.

**Theorem 2.1** *(Data processing inequality)*
*Suppose we have a probability model described by the following (Markov Chain):*

$$X \to Y \to Z$$

*where $X \perp Z|Y$, then it must be that $I(X,Y) \geq I(X,Z)$.*

**Proof:** By the Chain rule, we know that I(X, (Y,Z)) can be decomposed in two ways:

$$I(X,(Y,Z)) = I(X,Z) + I(X,Y|Z)$$

$$= I(X,Y) + I(X,Z|Y)$$

Because $I(X,Z|Y) = 0$ by assumption ($X \perp Z|Y$), we have that $I(X,Z) + I(X,Y|Z) = I(X,Y)$. Since mutual information is always non-negative, we get that $I(X,Z) \leq I(X,Y)$. ■

## 2.5 Fano's Inequality

Suppose that we want to predict the sent code or channel input $X$ from the received code or channel output $Y$. If $H(X|Y) = 0$, then intuitively, the probability of the error $p_e$ should be 0.

**Theorem 2.2** *Suppose $X$ is a random variable with finite outcomes in $\mathcal{X}$. Let $\hat{X} = g(Y)$ be the predicted value of $X$ for some deterministic function $g$ that also takes values in $\mathcal{X}$. Then we have:*

$$p_e \equiv p(\hat{X} \neq X) \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|}$$

*Or, stated more strongly:*

$$H(Ber(p_e)) + p_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

*where $Ber(p_e)$ refers to the bernoulli error random variable $E$ with $Pr(E = 1) = p_e$.*

**Proof:** Define random variable $E = \begin{cases} 1 \; if \; \hat{X} \neq X \\ 0 \; else \end{cases}$

By the Chain rule, we have two ways of decomposing $H(E, X|Y)$:

$$H(E, X|Y) = H(X|Y) + H(E|X, Y)$$
$$H(E, X|Y) = H(E|Y) + H(X|E, Y)$$

Also, $H(E|X, Y) = 0$ since $E$ is deterministic once we know the values of $X$ and $Y$ (and $g(Y)$). Thus we have that

$$H(X|Y) \leq H(Ber(p_e)) + H(X|E, Y)$$

To bound $H(X|E, Y)$, we use the definition of conditional entropy:

$$H(X|E, Y) = H(X|E = 0, Y)p(E = 0) + H(X|E = 1, Y)p(E = 1)$$

We will first note that $H(X|E = 0, Y) = 0$ since $E = 0$ implies that $X = g(Y)$ and hence, if we observe both $E = 0$ and $Y, X = g(Y)$ is no longer random. Also, $P(E = 1) = p_e$.

Next, we note that $H(X|E = 1, Y) \leq \log(|\mathcal{X}| - 1)$. This is because if we observe $E = 1$ and $g(Y)$, then $X$ cannot be equal to $g(Y)$ and thus can take on at most $|\mathcal{X} - 1|$ values.

Putting everything together, we have

$$H(Ber(p_e)) + p_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

as desired.

■

An alternate proof is based on the intuition that a good reconstruction algorithm can be used to compress the source. In what follows, we will design a compression algorithm using the estimator $g(Y)$. Imagine that $g$ is an estimator that has error probability $p_e$. Here is the compression scheme:

1. Compute $g(Y)$
2. Check if $g(Y) = X$
3. if yes, output $Y, 0$
4. if no, output $Y, 1, X$

Not counting $Y$, the entropy of the resulting string is $H(p_e) + p_e \log(|\mathcal{X}| - 1)$, since we know that $g(Y)$ has error probability $p_e$ so that additional entropy of the indicator bit is $H(p_e)$ and we include the source symbol $X$ only if we make an error. Since from the compression one can perfectly recover the source symbol $X$, the additional information must account for the remaining entropy in $X$ after seeing $Y$, that is $H(X|Y)$. From this it follows that $H(p_e) + p_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$, which is precisely Fano's Inequality.

Fano's inequality will be critical to establishing the fundamental limits of data compression and transmission. We will use it to characterize when reconstruction of sent code is not possible, i.e. the probability of error is bounded away from zero. Similarly, Fano's inequality will be used to establish the fundamental limits of inference in machine learning problems by demonstrating when the probability of error of recovering the true model from data is bounded away from zero.

### 2.5.1 Submodularity

Submodularity introduces and describes a notion of diminishing returns when adding a new element to a set. Submodularity intuitively states that adding an element to a smaller set helps more than adding it to a larger set.

**Definition 2.3** *Let $\Omega$ be a finite set. A set function $f : 2^\Omega \to \mathbb{R}$ is a submodularity function if any of the following hold (they are all equivalent):*

1. *Adding an element to a subset set has more value than adding the same element to a super set:*

$$\forall X, Y \subseteq \Omega, \quad X \subseteq Y, \quad \forall x \in \Omega \setminus Y, \quad f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$$

2. $$\forall S, T \subseteq \Omega \quad f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$$

3. $$\forall X \subseteq \Omega \quad x_1, x_2 \in \Omega \backslash X \quad f(X \cup \{x_1\}) + f(X \cup \{x_2\}) \geq f(X \cup \{x_1, x_2\}) + f(X)$$

Examples:

1. Linear functions are submodular. A linear set-valued function is parametrized by a vector $w \in \mathbb{R}^{|\Omega|}$ and defined by
$$f(S) = \sum_{x \in S} w_x = w^T \mathbf{1}_S$$
where $\mathbf{1}_S \in \{0, 1\}^{|\Omega|}$ is the characteristic (indicator) vector for the set $S$.

2. Entropy of a collection of random variables is submodular. Let $\Omega = \{X_1, X_2, ..., X_n\}$. Specifically, the function is the joint entropy of a subset of random variables.
$$f(S) = H(\{X_s | s \in S\}) = -\sum p(X_s) \log p(X_s)$$

Submodularity is a nice property in part because greedy maximization of a submodular function. If we want to maximize a submodular function $f$ over all sets of size at most $K$, then a natural algorithm is based on greedy maximization: Starting with the emptyset, we find the single element that leads to the largest increase in $f$ and add it to the candidate solution, and repeat. This algorithm comes with the following approximation guarantee:

**Theorem 2.4** *Consider the optimization problem $maximize_S f(S) subject to |S| \leq K$ for $f$ submodular and non-decreasing with $f(\emptyset) = 0$ then*

$$f(\hat{S}) \geq (1 - \frac{1}{e})(f(OPT)$$

*where $f(\hat{S})$ is a the greedy solution and OPT is the optimal solution.*

Intuitively what this says is that performance of the greedy solution, i.e. adding the next element which results in the largest incremental benefit performs reasonably well in terms of the optimal solution. Note that finding the optimal solution in general requires exhaustive search over all $\binom{|\Omega|}{K}$ subsets.