10-704: Information Processing and Learning **Spring 2012**

Lecture 2: Gibb's, Data Processing and Fano's Inequalities

Lecturer: Aarti Singh *Scribes: Min Xu*

2.1 Brief Review

Recall that, in last class, we introduced the following definitions:

- The **Shannon information content** of an outcome of a random variable is $\log_2 \frac{1}{p(x)}$ bits
- **Entropy** is the average uncertainty in random variable X: $H(X) = \mathbb{E}_p[\log_2 \frac{1}{p(x)}]$
- **Relative entropy** between two random variables is $D(p \,||\, q) = \mathbb{E}_p[\log_2 \frac{p(x)}{q(x)}]$, usually not equal to $D(q \,||\, p)$
- **Mutual information** between two random variables X, Y: $I(X,Y) = D(p(x,y) \,||\, p(x)p(y))$
- **Joint entropy** $H(X,Y) = \mathbb{E}_{p(x,y)}[\log_2 \frac{1}{p(x,y)}]$
- **Conditional entropy** $H(Y \mid X) = \mathbb{E}_{p(x,y)}[\log_2 \frac{1}{p(y \mid x)}]$

2.1.1 Fundamental Limits in Information Theory

The **data compression** model is the following:

$$\text{Source Model} \xrightarrow{X_1,\ldots,X_n} \boxed{\text{Compressor}} \xrightarrow{b_1,\ldots,b_m \text{ bits}} \boxed{\text{Decompressor}} \longrightarrow \text{Receiver}$$

Suppose the source data is generated from some distribution $p(X)$. If the rate of the compression, which is 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.

The **data transmission** model is the following:

$$\text{Src} \longrightarrow \boxed{\text{Encoder}} \xrightarrow{X} \boxed{\text{Channel}}_{p(y \mid x)} \xrightarrow{Y} \boxed{\text{Decoder}} \longrightarrow \text{Receiver}$$

If the rate of the code, which is $\frac{\#\text{message bits}}{\#\text{code bits}}$, is greater than the channel capactiy $C = \max_{p(X)} I(X,Y)$, then perfect reconstruction is not possible.

The **inference** problem is similar to the data transmission problem except we do not have an encoder:

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

$$f(X_1), f(X_2), \ldots \xrightarrow{p(y \mid f(x))} \boxed{\text{Channel}} \xrightarrow{Y_1, Y_2, \ldots} \boxed{\text{Decoder}} \xrightarrow{\hat{f}}$$

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

$$\theta \longrightarrow \boxed{\text{Channel}} \xrightarrow{X_1, X_2, \ldots} \boxed{\text{Decoder}} \xrightarrow{\hat{p}}$$

Under the log loss, we have that

$$\begin{aligned} \text{Excess Risk}(q) &= \text{Risk}(q) - \text{Risk}(p) \\ &= D(p \,||\, q) \end{aligned}$$

Fundamental limits of inference problems are often characterized by minimax 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 minimax excess risk is $\inf_q \sup_{p \in \mathcal{P}} D(p \,||\, q)$ and we will show that this is equal to the capacity C of the corresponding channel. This would imply that for all estimators q, $\sup_{p \in \mathcal{P}} D(p \,||\, q) \geq C$.

We will state and prove these results formally later in the course. Information theory will help us identify these fundamental limits of data compression, tranmission 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.

2.2 Useful Properties of Information Quantities

1. $H(X) \geq 0, = 0$ if and only if X is constant, that is, $p(x)$ is 1 for some outcome x.

   For example, consider a binary random variable $X \sim \text{Bernoulli}(\theta)$. 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 $\theta = \frac{1}{2}$ implies the distribution is uniform).

2. (**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 $\mathcal{X} = \{x : p(x) > 0\}$

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

   The first step is justified for $x \in \mathcal{X}$. The first inequality follows from Jensen's inequality since log is concave.

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

3. 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, $p(x,y) = p(x)p(y)$.

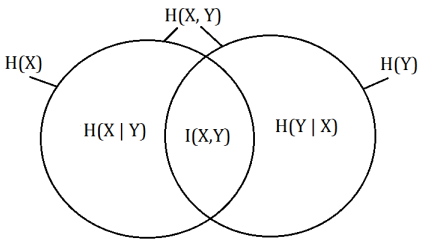
4. $H(X) \leq \log |\mathcal{X}|$ where $\mathcal{X}$ is the set of all outcomes with non-zero probability. Equality is achieved iff X is uniform.

   Proof: Let u be the uniform distribution over X, i.e. $u(x) = 1/|\mathcal{X}|$.

   $$\begin{aligned} D(p \,||\, u) &= \mathbb{E}_p[\log \frac{1}{u}] - \mathbb{E}_p[\log \frac{1}{p}] \\ &= \log |\mathcal{X}| - H(X) \end{aligned}$$

   The claim follows from the previously described property of relative entropy. ■

5. 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) $I(X,Y) = H(X) + H(Y) - H(X,Y)$

6. **Conditioning does not reduce entropy/Information always helps**

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

   i.e. observing additional information cannot increase the random-ness of X.

   Proof: Follows from properties 3 and 5(b) above. ■

   It is important to note that for a particular outcome y, it could be that $H(X \,|\, Y = y) > H(X)$, but it cannot happen when we average across all outcomes of Y.

2.3 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 (see homework 1), we know that $I(X,(Y,Z))$ can be decomposed in two ways:

$$\begin{aligned} I(X,(Y,Z)) &= I(X,Z) + I(X,Y \,|\, Z) \\ &= I(X,Y) + I(X,Z \,|\, Y) \end{aligned}$$

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 $I(X,Z) \leq I(X,Y)$. ■

2.4 Sufficient Statistics

Suppose we have a family of distributions parametrized by parameter θ: $\{f_\theta(x)\}$. Suppose data X is drawn from a distribution $f_\theta(x)$.

Definition 2.2 *A function of the data $T(X)$ is called a statistic. A function $T(X)$ is a **sufficient statistic** relative to parameter θ if $X \perp \theta \,|\, T(X)$.*

Here is visual way to remember sufficient statistic: for any statistic, it is true that

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

For sufficient statistic, it is also true that

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

That is, once we know $T(X)$, the remaining random-ness in X does not depend on θ.

Example:
Suppose $X_1, ..., X_n \sim Ber(\theta)$, then $T(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^n X_i$ is a sufficient statistic.

Suppose $X_1, ..., X_n \sim Unif(\theta, \theta+1)$, then $T(X_1, ..., X_n) = (\max_i X_i, \min_i X_i)$ is a sufficient statistic. You will prove this in the Homework.

2.5 Fano's Inequality

Suppose 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 error p_e should be 0.

Theorem 2.3 *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 $E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{else} \end{cases}$

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

$$\begin{aligned} H(E, X \,|\, Y) &= H(X \,|\, Y) + H(E \,|\, X, Y) \\ H(E, X \,|\, Y) &= \underbrace{H(E \,|\, Y)}_{\leq H(E) = H(Ber(p_e))} + H(X \,|\, E, Y) \end{aligned}$$

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. ■

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 For Random Functions

The following section, not in lecture, slightly generalizes Fano's Inequality to possibly random functions g.

Theorem 2.4 *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 possibly non-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 $E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{else} \end{cases}$

By the Chain rule, we have two ways of decomposing $H(E, X \mid \hat{X})$:

$$\begin{aligned} H(E, X \mid \hat{X}) &= H(X \mid \hat{X}) + H(E \mid X, \hat{X}) \\ H(E, X \mid \hat{X}) &= \underbrace{H(E \mid \hat{X})}_{\leq H(E) = H(Ber(p_e))} + H(X \mid E, \hat{X}) \end{aligned}$$

Also, $H(E \mid X, \hat{X}) = 0$ since E is deterministic once we know the values of X and $\hat{X}$. Thus, we have that

$$H(X \mid \hat{X}) \;\leq\; H(Ber(p_e)) + H(X \mid E, \hat{X})$$

To bound $H(X \mid E, \hat{X})$, we use the definition of conditional entropy:

$$H(X \mid E, \hat{X}) = H(X \mid E = 0, \hat{X})p(E = 0) + H(X \mid E = 1, \hat{X})p(E = 1)$$

We will first note that $H(X \mid E = 0, \hat{X}) = 0$ since $E = 0$ implies that $X = \hat{X}$ and hence, if we observe both $E = 0$ and $\hat{X}$, X is no longer random. Also, $P(E = 1) = p_e$.

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

To complete the proof, we just need to show that $H(X \mid \hat{X}) \geq H(X \mid Y)$. This holds since $X \to Y \to \hat{X}$ forms a Markov chain and thus

$$\begin{aligned} I(X, Y) &\geq I(X, \hat{X}) \text{(by data processing inequality)} \\ H(X) - H(X \mid Y) &\geq H(X) - H(X \mid \hat{X}) \text{ (by venn-diagram relation)} \\ H(X \mid Y) &\leq H(X \mid \hat{X}). \end{aligned}$$

■