ECE 6980
Algorithmic and Information-Theoretic Methods in Data Science

Instructor: Jayadev Acharya
Scribe: Chengrun Yang

Lecture #4
6th September, 2017

1 Introduction

Lower bounds help in getting relevant results such as sample complexity and time complexity (For example, we prefer $O(n)$ to $O(2^n)$). We have shown the following lower bounds for the order of required samples in the past lectures:

1. Hypothesis testing between two distributions: $\Omega(\frac{1}{\epsilon^2})$.
2. Learning distribution p over $[k]$: $\Omega(\frac{k}{\epsilon^2})$.
3. Uniformity testing: $\Omega(\frac{\sqrt{k}}{\epsilon^2})$.

In this lecture, we are going to review some basics about information theory, which will be helpful in proving lower bounds.

2 Information Theory Basics

Note: Please refer to Chapter 2 of Cover & Thomas [1] for more information.

2.1 Entropy

Definition 1. Given a probability distribution p, the entropy of that distribution $H(p)$ is $\sum_x p(x) \log_2 \frac{1}{p(x)}$.

Entropy is used to describe the amount of randomness for a given probability distribution. Note that $H(p) = \mathbb{E}_p(\log \frac{1}{p(x)})$[1] according to the convexity of $f(x) = \log \frac{1}{x}$.

The following theorem shows that for all distributions over $[k]$, the uniform distribution has the largest entropy.

Theorem 1. *Given a distribution with k elements, we have $0 \le H(p) \le \log(k)$.*

Proof. According to the concavity of $f(x) = \log(x)$, $H(p) \le \log(\mathbb{E}_p[\frac{1}{p(x)}]) = \log(\sum p(x) \frac{1}{p(x)}) = \log(k)$. □

2.2 Kullback-Leibler (KL) Divergence

Definition 2. Given two probability distributions p and q, the KL divergence of these two distributions $KL(p,q)$ (or $D(p||q)$) is $\sum_x p(x) \log \frac{p(x)}{q(x)}$.

If q is a uniform distribution u over $[k]$, then $D(p||u) = \sum p(x) \log(p(x)k) = \log(k) - H(p)$

[1] We omit the base of log from now on.

Theorem 2. *$D(p||q)$ is convex in p and q, i.e. $\forall \lambda \in [0,1]$, $\lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \geq D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2)$.*

Theorem 3. *$\forall$ probability distributions p, q, $D(p||q) \geq 0$.*

Proof.

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x)(-\log \frac{q(x)}{p(x)}) \\ &\geq -\log \sum_x q(x) \\ &= 0 \end{aligned} \tag{1}$$

The inequality holds due to the convexity of $f(x) = -log(x)$. □

2.3 Conditional Entropy

Definition 3. Given random variables X, Y and distributions $\mathcal{X}$, $\mathcal{Y}$, the conditional entropy of X given Y is

$$\begin{aligned} H(X|Y) &\triangleq \sum_y P(Y=y)H(X|Y=y) = \sum_y P(Y=y) \sum_x P(X=x|Y=y) \log \frac{1}{P(X=x|Y=y)} \\ &= \sum_{x,y} P(X=x, Y=y) \log \frac{P(Y=y)}{P(X=x,Y=y)} \\ &= \sum_{x,y} P(X=x, Y=y) \log \frac{1}{P(X=x,Y=y)} - \sum_y P(Y=y) \log \frac{1}{P(Y=y)} \\ &= H(X,Y) - H(Y) \end{aligned} \tag{2}$$

From the above we know $H(X|Y) = H(X,Y) - H(Y)$. Intuitively, this means conditional entropy of X given Y captures the remaining randomness in X after knowing Y.

Theorem 4. *Chain Rule of Entropy: $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$*

Specifically, if X is independent of Y, then $P(X|Y) = P(X)$, $H(X,Y) = H(X) + H(Y)$. The latter equality is equivalent to $H(X|Y) = H(X)$.

For a series of random variables $X, Y, Z, ...$, we have $H(X,Y,Z) = H(X) + H(Y|X) + H(Z|X,Y) + ...$

2.4 Mutual Information

Mutual information of two random variables X and Y tells how much information the random variable Y (or X) gives about X (or Y).

Definition 4. Mutual information of X and Y is

$$\begin{aligned} I(X;Y) &\triangleq H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{3}$$

Theorem 5. $I(X;Y) \geq 0$.

Proof.

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x,y} P(X=x, Y=y) \log \frac{1}{P(X=x)} + \sum_{x,y} P(X=x, Y=y) \log \frac{1}{P(Y=y)} \\ &- \sum_{x,y} P(X=x, Y=y) \log \frac{1}{P(X=x, Y=y)} \\ &= \sum_{x,y} P(X=x, Y=y) \log \frac{P(X=x, Y=y)}{P(X=x)P(Y=y)} \geq 0 \end{aligned} \tag{4}$$

The last inequality holds since the left-hand-side of the last line is actually the KL Divergence of probability distributions $P(X,Y)$ and $P(X)P(Y)$. □

2.5 Multiway Classification and Channel Capacity

First, let's have a look at the multiway classification problem. Its settings are:

1. Given m possible messages (distributions) p_1, p_2, ..., p_m.
2. The true distribution M is selected uniformly at random from $\{1, ..., m\}$.
3. Observe output Y from source distribution p_m.
4. Predict M from $\hat{M}(Y)$.

This can be regarded as a message passing problem $M - Y - \hat{M}$. Generally, given the message passing procedure $X - Y - Z$. If this is a Markov chain, then we have $P(Z|Y) = P(Z|Y,X)$, $H(Z|Y) = H(Z|Y,X)$.

We can formulate a simple message passing problem, which is from one Bernoulli distribution $\{0,1\}$ to the other. Denote the random variables in the two distributions as X and Y. If $P(Y=0|X=0) = P(Y=1|X=0) = \frac{1}{2}$, then no information can be sent from X to Y; If $P(Y=0|X=0) = 0.9$, $P(Y=1|X=0) = 0.1$, then some amount of information can be sent.

We often use the mutual information $I(X;Y)$ to quantify the channel capacity.

2.6 Data Processing Inequality

Theorem 6. *Given a data processing procedure $X - Y - Z$, we have $I(X;Z) \leq I(Y,Z)$ if this procedure is a Markov chain.*

Proof.

$$I(X;Z) = H(Z) - H(Z|X) \leq H(Z) - H(Z|X,Y) = H(Z) - H(Z|Y) = I(Y;Z) \tag{5}$$

□

2.7 Fano's Inequality

Definition 5. Given a message passing procedure $M - Y - \hat{M}$, we have

$$I(M;Y) \geq P(\text{correct}) \log(m-1) - \log 2$$

Proof. Denote $\mathbb{C} \triangleq \mathbb{I}\{M = \hat{M}\}$.

$$\begin{aligned} H(M, \mathbb{C}|\hat{M}) &= H(M|\hat{M}) + H(\mathbb{C}|M, \hat{M}) \\ &= H(M|\hat{M}) \end{aligned} \tag{6}$$

The last equality holds since we can get $\mathbb{C}$ unambiguously when knowing M and $\hat{M}$.

Note that we also have

$$\begin{aligned} H(M, \mathbb{C}|\hat{M}) &= H(\mathbb{C}|\hat{M}) + H(M|\hat{M}, \mathbb{C}) \\ &\leq H(\mathbb{C}) + P(\mathbb{C} = 1) H(M|\hat{M}, \mathbb{C} = 1) \\ &+ P(\mathbb{C} = 0) H(M|\hat{M}, \mathbb{C} = 0) \end{aligned} \tag{7}$$

Since $\mathbb{C}$ is a Bernoulli random variable, $H(\mathbb{C}) \leq \log 2$. Also, $H(M|\hat{M}, \mathbb{C} = 1) = 0$ since we can know M for sure given $\hat{M}$ and $\mathbb{C}$. $H(M|\hat{M}, \mathbb{C} = 0) \leq \log(m-1)$ from concavity of $f(x) = log(x)$. Thus we have

$$H(M|\hat{M}) \leq \log 2 + P(\text{error}) \log(m-1) \tag{8}$$

Thus the mutual information

$$\begin{aligned} I(M; \hat{M}) &= H(M) - H(M|\hat{M}) \\ &\geq log(m) - P(\text{error}) \log(m-1) - \log 2 \\ &\geq P(\text{correct}) \log(m-1) - \log 2 \end{aligned} \tag{9}$$

Finally, using the data processing inequality, we have

$$I(M;Y) \geq I(M;\hat{M}) \geq P(\text{correct}) \log(m-1) - \log 2$$

□

In the next lecture, we are going to show that "testing" a multiway classification problem not more difficult than "learning" it, in the sense that $n_{\text{learn}} \geq n_{\text{test}}$, in which n denotes the number of required samples.

References

[1] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.