Lecture 2: Information Theory and Lower Bounds

Daniel Kane
Scribe: Thomas Tucker

April 5, 2017

1 Review

Our goal is to learn a distribution p on $[n]$ to Variational Distance ϵ. Last time, we showed that if we take $O(n/\epsilon^2)$ samples and return the empirical distribution, we will achieve this goal. We now seek to show that n/ϵ^2 samples also provides the lower bound on the necessary number of samples.

2 Entropy

The most basic thing you want in Information Theory is given a variable p, determine how much information p encdes. In other words, if it is given that $p = x$, determine how "surprising" this is. We measure "surprisingness" as $\log(\frac{1}{\Pr(p=x)})$

Shannon Entropy (or, more commonly, just *Entropy*) is defined as such:

$$H(p) = \sum_{x \in X} \Pr(p = x) \log\left(\frac{1}{\Pr(p = x)}\right)$$

If p is supported on $[n]$, then it is not hard to see that $\log(n) \geqslant H(p) \geqslant 0$.

Note that the base of the logarithm is not typically important. When it is specified, the base defines the "units" of the entropy; when the logarithm is base 2, the entropy is in "bits," and when the log is base e, the entropy is in "nats.",

Also note that this only applies to discrete variables. While there is an analogous quantity for continuous variables, we will not likely be covering it in this course.

Example Take a variable p which is uniform over some $[n]$. We can then calculate the entropy as follows:

$$H(p) = \sum_{i=1}^{n} \frac{1}{n} \log(n) = \log(n)$$

Example Take a weighted coin p with probabilities q and q', such that $q+q' = 1$. The entropy is calculated as follows:

$$H(p) = q \log\left(\frac{1}{q}\right) + q' \log\left(\frac{1}{q'}\right)$$

2.1 Joint Entropy

We can also take entropies over joint distributions. If we have two variables p and q, we can not only find $H(p)$ and $H(q)$, we can also find $H(p, q)$. If p can have any value in X and q can have any value in Y and p and q are independent, we can calculate this joint entropy as follows:

$$H(p,q) = \sum_{x\in X, y\in Y} p_x q_x \log\left(\frac{1}{p_x q_y}\right) = \sum_{x\in X} p_x \log\left(\frac{1}{p_x}\right) + \sum_{y\in Y} q_y \log\left(\frac{1}{q_y}\right) = H(p) + H(q)$$

2.2 Relative Entropy

Relative Entropy covers the subject of how much entropy exists in one variable, $p \in X$, when the value of another variable $q \in Y$ is known. It is written and calculated as follows:

$$\begin{aligned} H(p|q) = \mathbb{E}_{q=y}\left[H(p|q=y)\right] &= \sum_{y\in Y} q_y \sum_{x\in X} \frac{1}{q_y} \Pr(p-x, q=y) \log\left(\frac{q_y}{\Pr(p=x, q=y)}\right) \\ &= \sum_{y\in Y, x\in X} \Pr(p-x, q=y)\left(\log\left(\frac{1}{\Pr(p=x,q=y)}\right) - \log\left(\frac{1}{q_y}\right)\right) \\ &= \sum_{y\in Y, x\in X} \Pr(p-x, q=y)\log\left(\frac{1}{\Pr(p=x,q=y)}\right) - \sum_{y\in Y} \Pr(q=y)\log\left(\frac{1}{q_y}\right) \\ &= H(p,q) - H(q) \end{aligned}$$

A quick explanation: Going from the first line to the second line, we cancel the q_y and $\frac{1}{q_y}$ terms, and expand the logarithm. Going from the second line ot the third line, we distribute the probability over the two log terms, and sum the right hand side of the subtraction over X to combine the p terms into 1, where they become irrelevant. We then use the definition of entropy to finalize the equation.

This also gives us an interesting chain rule that we may use as follows:

$$H(p_1, p_2, p_3, ..., p_n) = H(p_1) + H(p_2|p_1) + H(p_3|p_1, p_2) + ... + H(p_n|p_1, ..., p_{n-1})$$

3 Shared Information

Given variables p and q, we may wonder how much q tells us about p. Shared Information captures this, and is calculated as follows:

$$I(p:q) = H(p) - H(p|q) = H(p) + H(q) - H(p,q)$$

This is the difference between the uncertainty of p and the uncertainty of p when q is known. This also shows how much we learn about p when we learn q.

Example Take a random element of $\mathbb{F}_2^n$.
p gives coordinates of positions in S.
q gives coordinates of positions in T.

$$H(p) = |S|,\ H(q) = |T|, \text{ and } H(p,q) = |S \cup T|$$
$$I(p:q) = |S| + |T| - |S \cup T| = |S \cap T|$$

Example p and q are correlated coins. Individually, p and q are fair coins; however, with probability x, it will be the case that $p = q$.

$$H(q) = \log(2),\ H(q|p) = x\log\left(\frac{1}{x}\right) + (1-x)\log\left(\frac{1}{1-x}\right)$$

$$I(p:q) = H(q) - H(q|p) = \log(2) - \left(x\log\left(\frac{1}{x}\right) + (1-x)\log\left(\frac{1}{1-x}\right)\right)$$

In the case of these two coins, $I(p:q) \geqslant 0$ with equality if and only if $x = \frac{1}{2}$. If $x \in \{0,1\}$, then $I(p:q) = 1$. If $x = \frac{1}{2} + \epsilon$ then $I(p:q) = \Theta(\epsilon^2)$.

Lemma $I(p:q) \geqslant 0$; that is, knowing one variable cannot cause you to know less about another variable.

Proof $I(p:q) = H(p) - H(p|q) = H(\mathbb{E}[p|q]) - \mathbb{E}[H(p|q)]$. It is the case that $H(p)$ is concave in p. Due to Jensen's Inequality, the entropy of the expectation is at least the expectation of the entropy.

3.1 Relative Shared Information

Written as $I(p:q|R)$, this captures the expected shared information between p and q when the value of R is known. It is calculated as follows:

$$I(p:q|R) = H(p|R) + H(q|R) - H(p,q|R)$$

There is also a chain rule for shared information:

$$I(p:q,R) = I(p:R) + I(p:q|R)$$

The Information Processing Inequality is as follows: $I(p:q) \geqslant I(f(p):q)$. The intuition behind this is that $f(p)$ cannot have more information than p.

Proof $I(p:q) = I(p, f(p):q)$ due to p and $f(p)$ being effectively the same variable. We then split this into $I(f(p):q) + I(p:q|f(p)) \geqslant I(f(p):q)$, as $I(p:q|f(p))$ can be no less than zero.

4 Lower Bounds and the Adversarial Method

The adversarial method is a competition between nature and the algorithm. Nature picks a distribution p, and the algoirthm seeks to learn it. Nature chooses p from a set of distributions which are difficult to identify from one another.

Looking at the proof for the upper bound algorithm, we notice that it is only tight when most of the probabilities with the different bins are approximately equal. As such, we are looking at a similar set of bins: $p_i \approx \frac{1}{n}$. In addition, we are selecting bins which are approximately ϵ far from each other; otherwise, the algorithm would be able to output any of the distributions, and know that it is within ϵ of the correct answer. Similarly, the distributions cannot be too much farther than ϵ apart from one another, as this eases the difficulty of distinguishing the distributions.

First, we assume that we have a distribution over $2n$ bins. The probability of each bin should be $\frac{1}{2n} \pm \frac{5\epsilon}{2n}$. We must also ensure that the sum of the probabilities is still one; we do this by looking at pairs of adjacent probabilities (p_{2i-1}, p_{2i}). These probabilities should either be $\left(\frac{1+5\epsilon}{2n}, \frac{1-5\epsilon}{2n}\right)$ or $\left(\frac{1-5\epsilon}{2n}, \frac{1+5\epsilon}{2n}\right)$. This gives us a distribution $\mathcal{D}$ over possible values of p.

Suppose there is an algorithm $p^N \to f(p^N) = q$. We want it such that with high probability, $d_{TV}(p,q) \leqslant \epsilon$. We're going to process this a little bit more to get q' so that for each pair (q_{2i-1}, q_{2i}), we get $\left(\frac{1\pm5\epsilon}{2n}, \frac{1\mp5\epsilon}{2n}\right)$, and it must be the case that $d_{TV}(p,q') \leqslant 2\epsilon$. q' is defined as the closest distribution in the support of $\mathcal{D}$ to q that satisfies the pairing requirement above, and that $d_{TV}(q,q') \leqslant \epsilon$.

The point now is that for any given pair of bins, if the guess was correct, the distance will be zero; and if the guess was incorrect, the distance will be $\frac{10\epsilon}{n} \cdot (\#$ of bins for which it is wrong). This tells us that if we know q', then the number of possibilities for p (assuming we got q' somewhat right) cannot be very large, as instead of us being wrong on all n pairs of bins, we are only wrong on $\frac{n}{5}$ pairs of bins, which leaves us in a much better spot.

As such, it follows that $I(\mathcal{D}, q') \geqslant \Omega(n)$ and $I(\mathcal{D}, q') \leqslant I(\mathcal{D}, p)$.

From here, we get that if we do not have enough samples, we will not have enough information to create an approximation.