EE 231A
Information Theory
Instructor: Rick Wesel

Handout #5, Solution Set 1
April 7, 2020

108 pts
Reading: Chapter 2 of *Elements of Information Theory*

**Lecture 1B: Entropy**

1. (10 pts) *Coin flips.*

   (a) (4 pts) The number $X$ of tosses till the first head appears has the geometric distribution with parameter $p = 1/2$, where $P(X = n) = pq^{n-1}$, $n \in \{1, 2, \ldots\}$. Hence the entropy of $X$ is

$$\begin{aligned} H(X) &= -\sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\ &= -\left[ \sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q \right] \\ &= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\ &= \frac{-p \log p - q \log q}{p} \\ &= H(p)/p \text{ bits.} \end{aligned}$$

   (b) (2 pts) If $p = 1/2$, then $H(X) = 2$ bits.

   (c) (4 pts) Intuitively, it seems clear that the best questions are those that have equally likely chances of receiving a yes or a no answer. Consequently, one possible guess is that the most "efficient" series of questions is: Is $X = 1$? If not, is $X = 2$? If not, is $X = 3$? ... with a resulting expected number of questions equal to $\sum_{n=1}^{\infty} n(1/2^n) = 2$. This should reinforce the intuition that $H(X)$ is a measure of the uncertainty of $X$. Indeed in this case, the entropy is exactly the same as the average number of questions needed to define $X$, and in general $E(\#\text{ of questions}) \geq H(X)$. This problem has an interpretation as a source coding problem. Let 0 =no, 1 =yes, $X$ =Source, and $Y$ =Encoded Source. Then the set of questions in the above procedure can be written as a collection of $(X, Y)$ pairs: $(1, 1)$, $(2, 01)$, $(3, 001)$, etc. . In fact, this intuitively derived code is the optimal (Huffman) code minimizing the expected number of questions.

2. (4 pts) We wish to find *all* probability vectors $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ which minimize

$$H(\mathbf{p}) = -\sum_i p_i \log p_i.$$

Now $-p_i \log p_i \geq 0$, with equality iff $p_i = 0$ or 1. Hence the only possible probability vectors which minimize $H(\mathbf{p})$ are those with $p_i = 1$ for some $i$ and $p_j = 0, j \neq i$. There are $n$ such vectors, i.e., $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$, and the minimum value of $H(\mathbf{p})$ is 0. These points are the corners of the simplex.

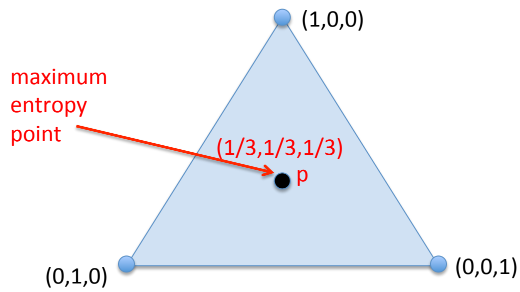



Figure 1: Illustration of minimum-entropy points (corners) for $n = 3$ simplex.

3. (12 pts) *Entropy of functions of a random variable.*

   (a) $H(X, g(X)) = H(X) + H(g(X)|X)$ by the chain rule for entropies.

   (b) $H(g(X)|X) = 0$ since for any particular value of X, g(X) is fixed, and hence $H(g(X)|X) = \sum_x p(x)H(g(X)|X = x) = \sum_x 0 = 0$.

   (c) $H(X, g(X)) = H(g(X)) + H(X|g(X))$ again by the chain rule.

   (d) $H(X|g(X)) \geq 0$, with equality iff $X$ is a function of $g(X)$, i.e., $g(.)$ is one-to-one. Hence $H(X, g(X)) \geq H(g(X))$.

   Combining parts (b) and (d), we obtain $H(X) \geq H(g(X))$.

## Lecture 1C: Relative Entropy

4. (4 pts) *Computing Relative Entropy for 2D $p$ and $q$.*

   Let $p(x, y)$ be given by

| $X/\mathcal{Y}$ | 0 | 1 |
|---|---|---|
| 0 | $\frac{1}{6}$ | $\frac{7}{12}$ |
| 1 | $\frac{1}{6}$ | $\frac{1}{12}$ |

   Let $q(x, y)$ be given by

| $X/\mathcal{Y}$ | 0 | 1 |
|---|---|---|
| 0 | $\frac{1}{4}$ | $\frac{1}{2}$ |
| 1 | $\frac{1}{12}$ | $\frac{1}{6}$ |

   Find $D(p||q)$.

   **Solution:**

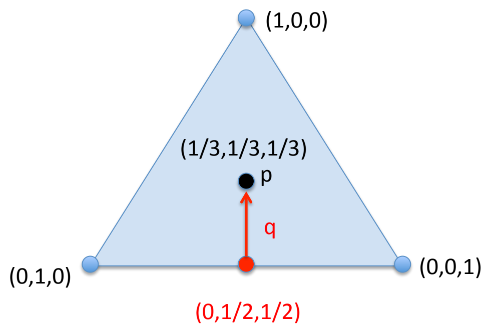
$$D(p||q) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{q(x,y)}\right) \tag{1}$$
$$= \frac{1}{6}\log\frac{4}{6} + \frac{1}{6}\log\frac{12}{6} + \frac{7}{12}\log\frac{14}{12} + \frac{1}{12}\log\frac{6}{12} \tag{2}$$
$$= \frac{1}{6}\log\frac{2}{3} + \frac{1}{6} + \frac{7}{12}\log\frac{7}{6} - \frac{1}{12} \tag{3}$$
$$= 0.1156 \tag{4}$$

5. (16 pts) *Computing Relative Entropy for $p$ and $q$ on a line in the 3D Simplex.*

Let $p(x)$ and $q(x)$ be three-outcome PMFs with the possible outcomes $\mathcal{X} = \{a, b, c\}$ so that $p$ and $q$ lie on the 3D simplex which is a 2D triangle in 3D space. Furthermore, let the PMF for $p$ be the point $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and the PMF for $q_\lambda$ be the point $(\frac{\lambda}{3}, \frac{3-\lambda}{6}, \frac{3-\lambda}{6})$ in the simplex.

(a) (4 pts) As in lecture, draw a triangle representing the simplex, show the point $p$ and the line segment that shows the trajectory of $q$ as $\lambda$ varies between 0 and 1.

**Solution:**




(b) (4 pts) Find $D(p||q_\lambda)$ as a function of $\lambda$ as $\lambda$ varies between 0 and 1 and use MATLAB to make a nice plot of $D(p||q_\lambda)$ vs. $\lambda$. You may not be able to plot $D(p||q_\lambda)$ in MATLAB for values of $\lambda$ near zero, but please evaluate (compute) what the value should be at $\lambda = 0$ (possibly infinity or a finite value).

**Solution:**

$$D(p||q_\lambda) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q_\lambda(x)} \tag{5}$$

$$= \frac{1}{3} \log \frac{1}{\lambda} + \frac{2}{3} \log \frac{2}{3-\lambda} \tag{6}$$

See the curve for $D(p||q_\lambda) = \infty$ in Fig. **??**. At $\lambda = 0$, $D(p||q_\lambda) = \infty$.

(c) (4 pts) Find $D(q_\lambda||p)$ as a function of $\lambda$ as $\lambda$ varies between 0 and 1 and use MATLAB to make a nice plot of $D(q_\lambda||p)$ vs. $\lambda$. Include your plot from the previous part for comparison. You may not be able to plot $D(q_\lambda||p)$ for values of $\lambda$ near zero, but please evaluate (compute) what the value should be at $\lambda = 0$ (possibly infinity or a finite value).

**Solution:**

$$D(q_\lambda||p) = \sum_{x \in \mathcal{X}} q_\lambda(x) \log \frac{q_\lambda(x)}{p(x)} \tag{7}$$

$$= \frac{\lambda}{3} \log \lambda + \frac{3-\lambda}{3} \log \frac{3-\lambda}{2} \tag{8}$$

See the curve for $D(q_\lambda||p)$ in Fig. **??**. Note that $0 \log 0 = 0$ so at $\lambda = 0$, $D(p||q_\lambda) = \log \frac{3}{2} = 0.585$.

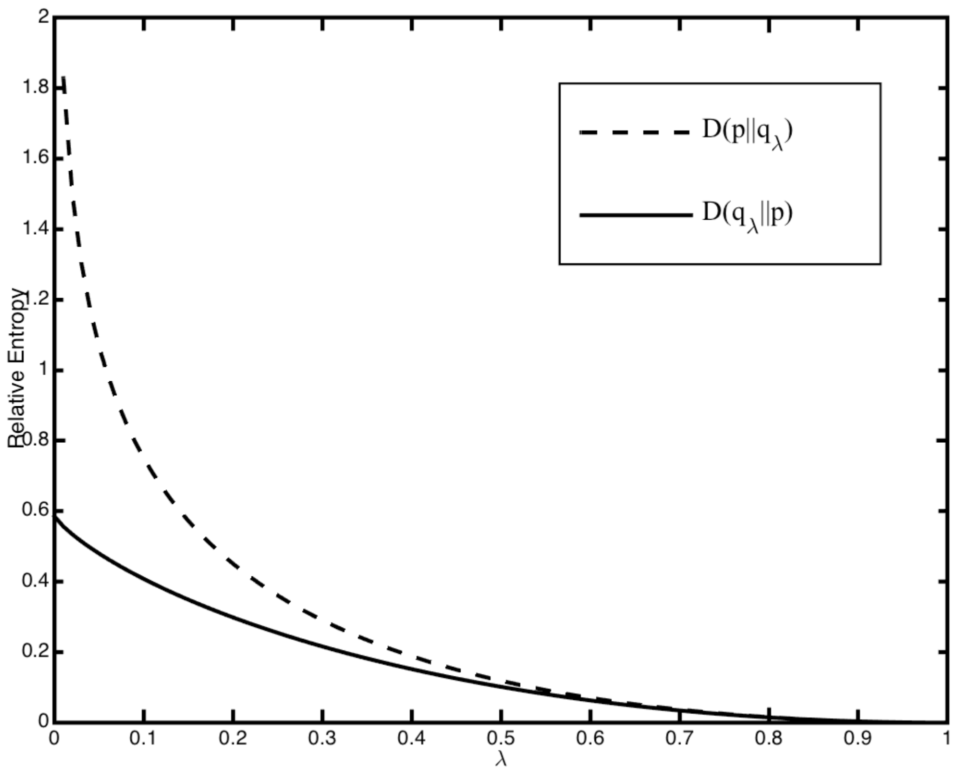



Figure 2: $D(p||q_\lambda)$ and $D(q_\lambda||p)$ vs. $\lambda$.

(d) (4 pts) Discuss the differences between $D(p||q_\lambda)$ and $D(q_\lambda||p)$. We learned that one interpretation of $D(p||q)$ is that it is a penalty for using the wrong distribution for determining description length. How come this penalty is infinitely larger at $\lambda = 0$ in one case as compared to the other?

**Solution:** Looking at the curves in Fig. **??** we can see that certainly $D(p||q_\lambda) \neq D(q_\lambda||p)$, but the differences are negligible for $\lambda$ near one and become infinite as $\lambda$ approaches zero. Lets consider the $\lambda = 0$ points from the perspective of $D(p||q_\lambda)$ and $D(q_\lambda||p)$ being measures of the penalty of using the distribution on the right of the $||$ to determine description length when the distribution on the left of the

$||$ is the actual distribution producing the symbols to be compressed. In the $q_\lambda$ distribution, the probability of $x = a$ goes to zero as $\lambda \to 0$. When $q_\lambda$ is the true distribution this means that the penalty in compression only applies to $x = b$ and $x = c$ since $x = a$ happens with probability zero. The penalty is that we use the description length $-\log \frac{1}{3}$ which is longer than the description length $-\log \frac{1}{2}$ we should have used for $x = b$ and $x = c$ . This is a finite penalty.

Now consider the case where $p$ is the true distribution. The three outcomes $x = a$, $x = b$, and $x = c$ all happen with equal probability of $\frac{1}{3}$ but in the $q_\lambda$ distribution, the probability of $x = a$ goes to zero as $\lambda \to 0$, which means that the appropriate description length for $x = a$ goes to infinity as $\lambda \to 0$. We end up using an infinitley long description length a third of the time which leads to an infinite description length and hence an infinite "penalty" $D(q_\lambda||p)$.

## Lecture 1D: Mutual Information

6. (4 pts) *Mutual Information?.*

   Can the relative entropy computed in the previous problem be expressed as a mutual information? Explain fully.

   **Solution:** Yes. In fact the relative entropy computed above is exactly $I(X;Y)$ since it turns out that $q(x,y) = p(x)p(y)$.

7. (12 pts) *Example of joint entropy*

   (a) $H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = 0.918$ bits $= H(Y)$.
   (b) $H(X|Y) = \frac{1}{3}H(X|Y=0) + \frac{2}{3}H(X|Y=1) = 0.667$ bits $= H(Y|X)$.
   (c) $H(X,Y) = 3 \times \frac{1}{3}\log 3 = 1.585$ bits.
   (d) $H(Y) - H(Y|X) = 0.251$ bits.
   (e) $I(X;Y) = H(Y) - H(Y|X) = 0.251$ bits.
   (f) See Figure 2.2 in *Elements of Information Theory.*

8. (8 pts) *Mutual Information and the Weather*

   (a)
   $$I(P_s;W) = H(P_s) - H(P_s|W) = 0 - 0 = 0 \tag{9}$$

   (b)
   $$I(P_w;W) = H(W) - H(W|P_w) = H(.9) - .75 \times H(0) - .25 \times H(.4) = 0.2262 \tag{10}$$

   (c) Wendy, provides the most information. In fact, Stormy provides no information at all.

   (d) Plant your tulip bulbs when Wendy forecasts rain.

## Lecture 2A: Convexity

9. (6 pts) *Concavity of entropy*

   (a) Show that $\log x$ is concave in $x$ for positive $x$. **Solution:**

$$\frac{d^2}{dx^2}\log_2 x = \frac{d^2}{dx^2}(\log_2 e)\ln x \tag{11}$$
$$= (\log_2 e)\frac{d}{dx}x^{-1} \tag{12}$$
$$= -(\log_2 e)x^{-2}\,, \tag{13}$$

   which is negative for positive $x$ so $\log x$ is concave in $x$ for positive $x$.

   (b) Show that $x\log x$ is convex in $x$ for positive $x$. **Solution:**

$$\frac{d^2}{dx^2}x\log_2 x = \frac{d^2}{dx^2}(\log_2 e)x\ln x \tag{14}$$
$$= (\log_2 e)\frac{d}{dx}(\ln x + 1) \tag{15}$$
$$= (\log_2 e)x^{-1}\,, \tag{16}$$

   which is positive for positive $x$ so $x\log x$ is convex in $x$ for positive $x$.

   (c) Use the second derivative to show that $H(p) = -p\log p - (1-p)\log(1-p)$ is concave in $p$ for $0 \le p \le 1$.

   **Solution:** While one can essentially refer to slide 44 of lecture 2 as follows: $H(p) = \log|\mathcal{X}| - D(p||u)$ so the convexity of relative entropy implies the concavity of entropy, this exercise required that you show the concavity by differentiation as in the other two parts.

$$\frac{d^2}{dp^2} - p\log p - (1-p)\log(1-p) = -(\log_2 e)\left(p^{-1} + (1-p)^{-1}\right)\,, \tag{17}$$

   which is negative for $0 \le p \le 1$ so $H(p)$ is concave in $p$ for $0 \le p \le 1$.

## Lecture 2B: Jensen's Inequality and its Applications

10. (4 pts) *Maximum entropy.* What is the maximum value of $H(p_1, ..., p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of $n$-dimensional probability vectors? Find all $\mathbf{p}$'s which achieve this maximum.
    **Solution:** Slides 16-17 of Lecture 2 showed that entropy is upper bounded by $H(X) \le \log|\mathcal{X}|$ because $H(p) = \log|\mathcal{X}| - D(p||u)$. We can achieve $H(X) = \log|\mathcal{X}|$ with a uniform distribution (i.e. all probabilities equal to $|\mathcal{X}|^{-1}$. This is the only distribution that achieves the maximum value since any other distribution will have a nonzero $D(p||u)$.

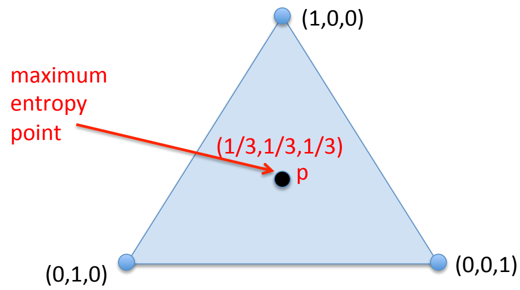



Figure 3: Illustration of maximum entropy point for $n = 3$ simplex.

11. (8 pts) *Drawing with and without replacement.* Intuitively, it is clear that if the balls are drawn with replacement, the number of possible choices for the $i$-th ball is larger, and therefore the conditional entropy is larger. But computing the conditional distributions is slightly involved. It is easier to compute the unconditional entropy.

    - With replacement. In this case the conditional distribution of each draw is the same for every draw. Thus

    $$X_i = \begin{cases} \text{red} & \text{with prob.} \frac{r}{r+w+b} \\ \text{white} & \text{with prob.} \frac{w}{r+w+b} \\ \text{black} & \text{with prob.} \frac{b}{r+w+b} \end{cases} \tag{18}$$

    and therefore

    $$H(X_i|X_{i-1}, \ldots, X_1) = H(X_i) \tag{19}$$

    $$= \log(r+w+b) - \frac{r}{r+w+b}\log r - \frac{w}{r+w+b}\log w - \frac{b}{r+w+b}\log b \tag{20}$$

    - Without replacement. The unconditional probability of the $i$-th ball being red is still $r/(r+w+b)$, etc. Thus the unconditional entropy $H(X_i)$ is still the same as with replacement. The conditional entropy $H(X_i|X_{i-1}, \ldots, X_1)$ is less than the unconditional entropy (We showed that as an application of Jensen's inequality.), and therefore the entropy of drawing without replacement is lower.

### Lecture 2C: Markov Chains and the Data Processing Inequality

12. (10 pts) *Conditional Mutual Information.*

    (a) (5 pts) Show that if $X \to Y \to Z$ forms a Markov chain, $I(X;Y|Z) \le I(X;Y)$.

    $$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \tag{21}$$

    $$= I(X;Y) + I(X;Z|Y) \tag{22}$$

For a Markov chain $X \to Y \to Z$, $I(X;Z|Y) = 0$. Thus

$$\begin{aligned} I(X;Y|Z) &= I(X;Y) - I(X;Z) && (23)\\ &\le I(X;Y)\,. && (24) \end{aligned}$$

(b) (5 pts) Is it always true that $I(X;Y|Z) \le I(X;Y)$ (i.e even for every case where $X \to Y \to Z$ does not form a Markov chain? No. Consider this example, which is also given in the text. Let $X, Y$ be independent fair binary random variables and let $Z = X + Y$. In this case we have that,

$$I(X;Y) = 0$$

and,

$$I(X;Y \mid Z) = H(X \mid Z) = 1/2.$$

So $I(X;Y) < I(X;Y \mid Z)$. Note that in this case $X, Y, Z$ are not Markov.

13. (10 pts) *Find the gap.*
You know that for $X \to Y \to Z$, $I(X;Z) \le I(Y;Z)$. Find the exact value of the gap between these mutual informations. i.e. Find $I(Y;Z) - I(X;Z)$ for the Markov chain $X \to Y \to Z$.

For full credit your answer must be a single information theoretic expression such as an entropy, a mutual information, or a conditional mutual information.

The answer is $I(Y;Z) - I(X;Z) = I(Y;Z|X)$ for a Markov chain.

Following the proof of Theorem 2.8.1 on pages 32-33, one approach is to write a mutual information two ways via the chain rule for mutual information:

$$\begin{aligned} I(X,Y;Z) &= I(X;Z) + I(Y;Z|X) && (25)\\ &= I(Y;Z) + I(X;Z|Y)\,. && (26) \end{aligned}$$

Realizing that $I(X;Z|Y) = 0$ for Markov chains completes the proof.

Another technique is the following:

$$\begin{aligned} I(X;Z) &= H(Z) - H(Z|X) && (27)\\ &= H(Z) - H(Z|X) + H(Z|Y) - H(Z|Y) && (28)\\ &= I(Y;Z) - \Big(H(Z|X - H(Z|Y)\Big) && (29)\\ &= I(Y;Z) - \Big(H(Z|X - H(Z|Y,X)\Big) \qquad \text{Since } X \to Y \to Z. && (30)\\ &= I(Y;Z) - I(Y;Z|X) && (31) \end{aligned}$$