PHYS 7895: Quantum Information Theory Fall 2015

Lecture 17 — October 21, 2015

Prof. Mark M. Wilde *Scribe: Mark M. Wilde*

1 Overview

In the previous lecture, we finished our development of the trace distance and the fidelity.

In this lecture, we now move on to entropies, beginning with classical entropies.

2 Entropy of a Random Variable

Consider a random variable X. Suppose that each realization x of random variable X belongs to an alphabet $\mathcal{X}$. Let $p_X(x)$ denote the probability density function of X so that $p_X(x)$ is the probability that realization x occurs. The information content $i(x)$ of a particular realization x is a measure of the surprise that one has upon learning the outcome of a random experiment:

$$i(x) \equiv -\log\left(p_X(x)\right). \tag{1}$$

The logarithm is base two and this choice implies that we measure surprise or information in bits.

The information content is a useful measure of surprise for particular realizations of random variable X, but it does not capture a general notion of the amount of surprise that a given random variable X possesses. The entropy $H(X)$ captures this general notion of the surprise of a random variable X—it is the expected information content of random variable X:

$$H(X) \equiv \mathbb{E}_X\left\{i(X)\right\}. \tag{2}$$

At a first glance, the above definition may seem strangely self-referential because the argument of the probability density function $p_X(x)$ is itself the random variable X, but this is well-defined mathematically. Evaluating the above formula gives an expression which we take as the definition for the entropy $H(X)$:

Definition 1 (Entropy)**.** *The entropy of a discrete random variable X with probability distribution $p_X(x)$ is*

$$H(X) \equiv -\sum_x p_X(x)\log\left(p_X(x)\right). \tag{3}$$

We adopt the convention that $0 \cdot \log(0) = 0$ for realizations with zero probability. The fact that $\lim_{\varepsilon\to 0} \varepsilon \cdot \log(1/\varepsilon) = 0$ intuitively justifies this latter convention. (We can interpret this convention as saying that the fact that the event has probability zero is more important than or outweighs the fact that you would be infinitely surprised if such an event would occur.)

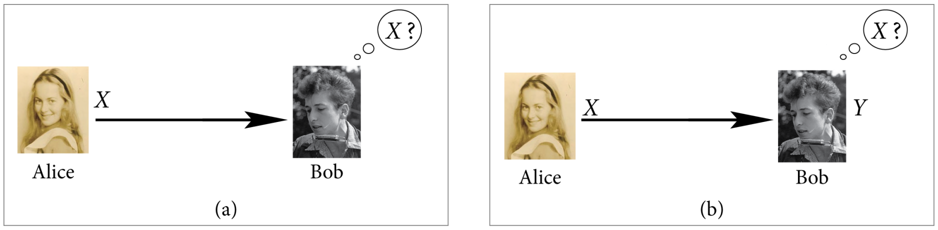


Figure 1: (a) The entropy $H(X)$ is the uncertainty that Bob has about random variable X before learning it. (b) The conditional entropy $H(X|Y)$ is the uncertainty that Bob has about X when he already possesses Y.

The entropy admits an intuitive interpretation. Suppose that Alice performs a random experiment in her lab that selects a realization x according to the density $p_X(x)$ of random variable X. Suppose further that Bob has not yet learned the outcome of the experiment. The interpretation of the entropy $H(X)$ is that it quantifies Bob's uncertainty about X before learning it—his expected information gain is $H(X)$ bits upon learning the outcome of the random experiment. Shannon's noiseless coding theorem makes this interpretation precise by proving that Alice needs to send Bob bits at a rate $H(X)$ in order for him to be able to decode a compressed message. Figure 1(a) depicts the interpretation of the entropy $H(X)$, along with a similar interpretation for the conditional entropy that we introduce in Section 3.

2.1 Mathematical Properties of Entropy

We now discuss five important mathematical properties of the entropy $H(X)$.

Property 2 (Non-Negativity)**.** *The entropy $H(X)$ is non-negative for any discrete random variable X with probability density $p_X(x)$:*

$$H(X) \geq 0. \tag{4}$$

Proof. Non-negativity follows because entropy is the expected information content of $i(X)$, and the information content itself is non-negative. It is perhaps intuitive that the entropy should be non-negative because non-negativity implies that we always learn some number of bits upon learning random variable X (if we already know beforehand what the outcome of a random experiment will be, then we learn zero bits of information once we perform it). In a classical sense, we can never learn a negative amount of information! □

Property 3 (Concavity)**.** *The entropy $H(X)$ is concave in the probability density $p_X(x)$.*

Proof. We justify this result with a heuristic "mixing" argument for now, and provide a formal proof in Section 8.1. Consider two random variables X_1 and X_2 with two respective probability density functions $p_{X_1}(x)$ and $p_{X_2}(x)$ whose realizations belong to the same alphabet. Consider a Bernoulli random variable B with probabilities $q \in [0, 1]$ and $1 - q$ corresponding to its two respective realizations $b = 1$ and $b = 2$. Suppose that we first generate a realization b of random variable B and then generate a realization x of random variable X_b. Random variable X_B then denotes a mixed version of the two random variables X_1 and X_2. The probability density of X_B is $p_{X_B}(x) = q p_{X_1}(x) + (1 - q)\, p_{X_2}(x)$. Concavity of entropy is the following inequality:

$$H(X_B) \geq q H(X_1) + (1 - q)\, H(X_2). \tag{5}$$

Our heuristic argument is that this mixing process leads to more uncertainty for the mixed random variable X_B than the expected uncertainty over the two individual random variables. We can think of this result as a physical situation involving two gases. Two gases each have their own entropy, but the entropy increases when we mix the two gases together. We later give a more formal argument to justify concavity. □

Property 4 (Permutation Invariance)**.** *The entropy is invariant with respect to permutations of the realizations of random variable X.*

Proof. That is, suppose that we apply some permutation π to realizations x_1, x_2, …, $x_{|\mathcal{X}|}$ so that they respectively become $\pi(x_1)$, $\pi(x_2)$, …, $\pi(x_{|\mathcal{X}|})$. Then the entropy is invariant under this shuffling because it depends only on the probabilities of the realizations, not the values of the realizations. □

Property 5 (Minimum Value)**.** *The entropy vanishes for a deterministic variable.*

Proof. We would expect that the entropy of a *deterministic* variable should vanish, given the interpretation of entropy as the uncertainty of a random experiment. This intuition holds true and it is the degenerate probability density $p_X(x) = \delta_{x,x_0}$, where the realization x_0 has all the probability and other realizations have vanishing probability, that gives the minimum value of the entropy: $H(X) = 0$ when X has a degenerate density. □

Sometimes, we may not have any prior information about the possible values of a variable in a system, and we may decide that it is most appropriate to describe them with a probability density function. How should we assign this probability density if we do not have any prior information about the values? Theorists and experimentalists often resort to a "principle of maximum entropy" or a "principle of maximal ignorance"—we should assign the probability density to be the one that maximizes the entropy.

Property 6 (Maximum Value)**.** *The maximum value of the entropy $H(X)$ for a random variable X taking values in an alphabet $\mathcal{X}$ is $\log|\mathcal{X}|$:*

$$H(X) \leq \log|\mathcal{X}|\,. \tag{6}$$

The inequality is saturated if and only if X is a uniform random variable on $\mathcal{X}$.

We give a proof of this statement after developing relative entropy.

3 Conditional Entropy

Let us now suppose that Alice possesses random variable X and Bob possesses some other random variable Y. Random variables X and Y share correlations if they are not statistically independent, and Bob then possesses "side information" about X in the form of Y. Let $i(x|y)$ denote the conditional information content:

$$i(x|y) \equiv -\log\left(p_{X|Y}(x|y)\right). \tag{7}$$

The entropy $H(X|Y=y)$ of random variable X conditioned on a particular realization y of random variable Y is the expected conditional information content, where the expectation is with respect to $X|Y=y$:

$$\begin{aligned} H(X|Y=y) &\equiv \mathbb{E}_{X|Y=y}\left\{i(X|y)\right\} && (8)\\ &= -\sum_x p_{X|Y}(x|y)\log\left(p_{X|Y}(x|y)\right). && (9) \end{aligned}$$

The relevant entropy that applies to the scenario where Bob possesses side information is the conditional entropy $H(X|Y)$, defined as follows:

Definition 7 (Conditional Entropy)**.** *Let X and Y be discrete random variables with joint probability distribution $p_{X,Y}(x,y)$. The conditional entropy $H(X|Y)$ is the expected conditional information content, where the expectation is with respect to both X and Y:*

$$\begin{aligned} H(X|Y) &\equiv \mathbb{E}_{X,Y}\left\{i(X|Y)\right\} && (10)\\ &= \sum_y p_Y(y)H(X|Y=y) && (11)\\ &= -\sum_y p_Y(y)\sum_x p_{X|Y}(x|y)\log\left(p_{X|Y}(x|y)\right) && (12)\\ &= -\sum_{x,y} p_{X,Y}(x,y)\log\left(p_{X|Y}(x|y)\right) && (13) \end{aligned}$$

The conditional entropy $H(X|Y)$ as well deserves an interpretation. Suppose that Alice possesses random variable X and Bob possesses random variable Y. The conditional entropy $H(X|Y)$ is the amount of uncertainty that Bob has about X given that he already possesses Y. Figure 1(b) depicts this interpretation.

The above interpretation of the conditional entropy $H(X|Y)$ immediately suggests that it should be less than or equal to the entropy $H(X)$. That is, having access to a side variable Y should only decrease our uncertainty about another variable. We state this idea as the following theorem and give a formal proof in Section 8.1.

Theorem 8 (Conditioning Does Not Increase Entropy)**.** *The entropy $H(X)$ is greater than or equal to the conditional entropy $H(X|Y)$:*

$$H(X) \geq H(X|Y), \tag{14}$$

and equality occurs if and only if X and Y are independent random variables. As a consequence of the fact that $H(X|Y) = \sum_y p_Y(y)H(X|Y=y)$, we see that the entropy is concave.

Non-negativity of conditional entropy follows from non-negativity of entropy because conditional entropy is the expectation of the entropy $H(X|Y=y)$ with respect to the density $p_Y(y)$. It is again intuitive that conditional entropy should be non-negative. Even if we have access to some side information Y, we always learn some number of bits of information upon learning the outcome of a random experiment involving X. Perhaps strangely, we will see that *quantum* conditional entropy can become negative, defying our intuition of information in the classical sense given here.

4 Joint Entropy

What if Bob knows neither X nor Y? The natural entropic quantity that describes his uncertainty is the joint entropy $H(X,Y)$. The joint entropy is merely the entropy of the joint random variable (X,Y):

Definition 9 (Joint Entropy)**.** *Let X and Y be discrete random variables with joint probability distribution $p_{X,Y}(x,y)$. The joint entropy $H(X,Y)$ is defined as*

$$H(X,Y) \equiv \mathbb{E}_{X,Y}\left\{i\left(X,Y\right)\right\} \tag{15}$$

$$= -\sum_{x,y} p_{X,Y}(x,y)\log\left(p_{X,Y}(x,y)\right). \tag{16}$$

5 Mutual Information

We now introduce an entropic measure of the common or mutual information that two parties possess. Suppose that Alice possesses random variable X and Bob possesses random variable Y.

Definition 10 (Mutual Information)**.** *Let X and Y be discrete random variables with joint probability distribution $p_{X,Y}(x,y)$. The mutual information $I(X;Y)$ is the marginal entropy $H(X)$ less the conditional entropy $H(X|Y)$:*

$$I(X;Y) \equiv H(X) - H(X|Y). \tag{17}$$

It quantifies the dependence or correlations of the two random variables X and Y. The mutual information measures how much knowing one random variable reduces the uncertainty about the other random variable. In this sense, it is the common information between the two random variables. Bob possesses Y and thus has an uncertainty $H(X|Y)$ about Alice's variable X. Knowledge of Y gives an information gain of $H(X|Y)$ bits about X and then reduces the overall uncertainty $H(X)$ about X, the uncertainty were he not to have any side information at all about X.

Exercise 11. *Show that the mutual information is symmetric in its inputs:*

$$I(X;Y) = I(Y;X), \tag{18}$$

implying additionally that

$$I(X;Y) = H(Y) - H(Y|X). \tag{19}$$

We can also express the mutual information $I(X;Y)$ in terms of the respective joint and marginal probability density functions $p_{X,Y}(x,y)$ and $p_X(x)$ and $p_Y(y)$:

$$I(X;Y) = \sum_{x,y} p_{X,Y}(x,y)\log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right). \tag{20}$$

The above expression leads to two insights regarding the mutual information $I(X;Y)$. Two random variables X and Y possess zero bits of mutual information if and only if they are statistically independent (recall that the joint density factors as $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ when X and Y are independent). That is, knowledge of Y does not give any information about X when the random variables are statistically independent. Also, two random variables possess $H(X)$ bits of mutual information if they are perfectly correlated in the sense that $Y = X$.

Theorem 12 below states that the mutual information $I(X;Y)$ is non-negative for any random variables X and Y—we provide a formal proof in Section 8.1. However, this follows naturally from the definition of mutual information in (17) and "conditioning does not increase entropy" (Theorem 8).

Theorem 12. *The mutual information $I(X;Y)$ is non-negative for any random variables X and Y:*

$$I(X;Y) \geq 0, \tag{21}$$

and $I(X;Y) = 0$ if and only if X and Y are independent random variables (i.e., if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$).

6 Relative Entropy

The relative entropy is another important entropic quantity that quantifies how "far" one probability density function $p(x)$ is from another probability density function $q(x)$. It can be helpful to have a more general definition in which we allow $q(x)$ to be a function taking non-negative values. Before defining the relative entropy, we need the notion of the support of a function.

Definition 13 (Support)**.** *Let $\mathcal{X}$ denote a finite set. The support of a function $f : \mathcal{X} \to \mathbb{R}$ is equal to the subset of $\mathcal{X}$ that takes non-zero values under f:*

$$\operatorname{supp}(f) \equiv \{x : f(x) \neq 0\}. \tag{22}$$

Definition 14 (Relative Entropy)**.** *Let p be a probability distribution defined on the alphabet $\mathcal{X}$, and let $q : \mathcal{X} \to [0, \infty)$. The relative entropy $D(p\|q)$ is defined as follows:*

$$D(p\|q) \equiv \left\{ \begin{array}{cc} \sum_x p(x) \log\left(p(x)/q(x)\right) & \textit{if } \operatorname{supp}(p) \subseteq \operatorname{supp}(q) \\ +\infty & \textit{else} \end{array} \right. . \tag{23}$$

According to the above definition, the relative entropy is equal to the following expected log-likelihood ratio:

$$D(p\|q) = \mathbb{E}_X \left\{ \log\left(\frac{p(X)}{q(X)}\right) \right\}, \tag{24}$$

where X is a random variable distributed according to p.

The above definition implies that the relative entropy is not symmetric under interchange of $p(x)$ and $q(x)$. Thus, the relative entropy is not a distance measure in the strict mathematical sense because it is not symmetric (nor does it satisfy a triangle inequality).

The relative entropy has an interpretation in source coding, if we let $q(x)$ be a probability distribution. Suppose that an information source generates a random variable X according to the density $p(x)$. Suppose further that Alice (the compressor) mistakenly assumes that the probability density of the information source is instead $q(x)$ and codes according to this density. Then the relative entropy quantifies the inefficiency that Alice incurs when she codes according to the mistaken probability density—Alice requires $H(X) + D\left(p\|q\right)$ bits on average to code (whereas she would only require $H(X)$ bits on average to code if she used the true density $p(x)$).

We might also see now that the mutual information $I(X;Y)$ is equal to the relative entropy $D(p_{X,Y}(x,y)\|p_X(x)p_Y(y))$ by comparing the definition of relative entropy in (23) and the expression for the mutual information in (20). In this sense, the mutual information quantifies how far the two random variables X and Y are from being independent because it calculates the "distance" of the joint density $p_{X,Y}(x,y)$ to the product $p_X(x)p_Y(y)$ of the marginals.

Let p_{X_1} and p_{X_2} be two probability distributions defined over the same alphabet. The relative entropy $D(p_{X_1}\|p_{X_2})$ admits a pathological property. It can become infinite if the distribution $p_{X_1}(x_1)$ does not have all of its support contained in the support of $p_{X_2}(x_2)$ (i.e., if there is some realization x for which $p_{X_1}(x) \neq 0$ but $p_{X_2}(x) = 0$). This can be somewhat bothersome if we like the interpretation of relative entropy as a notion of distance. In an extreme case, we would think that the distance between a deterministic binary random variable X_2 where $\Pr\left\{X_2 = 1\right\} = 1$ and one with probabilities $\Pr\left\{X_1 = 0\right\} = \varepsilon$ and $\Pr\left\{X_1 = 1\right\} = 1 - \varepsilon$ should be on the order of ε (this is true for the classical trace distance). However, the relative entropy $D(p_{X_1}\|p_{X_2})$ in this case is infinite, in spite of our intuition that these distributions are close. The interpretation in lossless source coding is that it would require an infinite number of bits to code a distribution p_{X_1} losslessly if Alice mistakes it as p_{X_2}. Alice thinks that the symbol $X_2 = 0$ never occurs, and in fact, she thinks that the typical set consists of just one sequence of all ones and every other sequence is atypical. But in reality, the typical set is quite a bit larger than this, and it is only in the limit of an infinite number of bits that we can say her compression is truly lossless.

Exercise 15. *Verify that the definition of relative entropy in Definition 14 is consistent with the following limit:*

$$D(p\|q) = \lim_{\varepsilon \searrow 0} D(p\|q + \varepsilon \mathbf{1}), \tag{25}$$

where $\mathbf{1}$ *denotes a vector of ones, so that the elements of* $q + \varepsilon\mathbf{1}$ *are* $q(x) + \varepsilon$*.*

7 Conditional Mutual Information

What is the common information between two random variables X and Y when we have some side information embodied in a random variable Z? The entropic quantity that quantifies this common information is the conditional mutual information.

Definition 16 (Conditional Mutual Information)**.** *Let X, Y, and Z be discrete random variables. The conditional mutual information is defined as follows:*

$$\begin{aligned} I(X;Y|Z) &\equiv H(Y|Z) - H(Y|X,Z) && (26)\\ &= H(X|Z) - H(X|Y,Z) && (27)\\ &= H(X|Z) + H(Y|Z) - H(X,Y|Z). && (28) \end{aligned}$$

Theorem 17 (Strong Subadditivity)**.** *The conditional mutual information $I(X;Y|Z)$ is non-negative:*

$$I(X;Y|Z) \geq 0, \tag{29}$$

and the inequality is saturated if and only if $X - Z - Y$ is a Markov chain (i.e., if $p_{X,Y|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$).

Proof. The proof of the above theorem is a straightforward consequence of the non-negativity of mutual information (Theorem 12). Consider the following equality:

$$I(X;Y|Z) = \sum_{z} p_Z(z) I(X;Y|Z=z), \tag{30}$$

where $I(X;Y|Z=z)$ is a mutual information with respect to the joint density $p_{X,Y|Z}(x,y|z)$ and the marginal densities $p_{X|Z}(x|z)$ and $p_{Y|Z}(y|z)$. Non-negativity of $I(X;Y|Z)$ then follows from non-negativity of $p_Z(z)$ and $I(X;Y|Z=z)$. The saturation conditions then follow immediately from those for mutual information given in Theorem 12 (considering that the conditional mutual information is a convex combination of mutual informations). □

The proof of the above classical version of strong subadditivity is perhaps trivial in hindsight (it requires only a few arguments). The proof of the quantum version of strong subaddivity is highly non-trivial on the other hand. We discuss strong subadditivity of quantum entropy in the next chapter.

Exercise 18. *The expression in* (29) *represents the most compact way to express the strong subadditivity of entropy. Show that the following inequalities are equivalent ways of representing strong subadditivity:*

$$H(XY|Z) \leq H(X|Z) + H(Y|Z), \tag{31}$$

$$H(XYZ) + H(Z) \leq H(XZ) + H(YZ), \tag{32}$$

$$H(X|YZ) \leq H(X|Z). \tag{33}$$

Exercise 19. *Prove the following chaining rule for mutual information:*

$$\begin{aligned} &I(X_1, \ldots, X_n; Y) \\ &\qquad = I(X_1;Y) + I(X_2;Y|X_1) + \cdots + I(X_n;Y|X_1,\ldots,X_{n-1}). \end{aligned} \tag{34}$$

8 Entropy Inequalities

The entropic quantities introduced in the previous sections each have bounds associated with them. These bounds are fundamental limits on our ability to process and store information. We introduce several bounds in this section: the non-negativity of relative entropy, two data-processing inequalities, Fano's inequality, and a uniform bound for continuity of entropy. Each of these inequalities plays an important role in information theory, and we describe these roles in more detail in the forthcoming subsections.

8.1 Non-Negativity of Relative Entropy

The relative entropy is always non-negative. This seemingly innocuous result has several important implications—namely, the maximal value of entropy, conditioning does not increase entropy (Theorem 8), non-negativity of mutual information (Theorem 12), and strong subadditivity (Theorem 17) are straightforward corollaries of it. The proof of this entropy inequality follows from the application of a simple inequality: $\ln x \leq x - 1$.

Theorem 20 (Non-Negativity of Relative Entropy)**.** *Let $p(x)$ be a probability distribution over the alphabet $\mathcal{X}$ and let $q : \mathcal{X} \to [0, 1]$ be a function such that $\sum_x q(x) \leq 1$. Then the relative entropy $D(p\|q)$ is non-negative:*

$$D(p\|q) \geq 0, \tag{35}$$

and $D(p\|q) = 0$ if and only if $p = q$.

Proof. First, suppose that $\text{supp}(p) \not\subseteq \text{supp}(q)$. Then the relative entropy $D(p\|q) = +\infty$ and the inequality is trivially satisfied.

Now, suppose that $\text{supp}(p) \subseteq \text{supp}(q)$. A proof relies on the inequality $\ln x \leq x - 1$ that holds for all $x \geq 0$ and saturates for $x = 1$. (Brief justification: Let $f(x) = x - 1 - \ln x$. Observe that $f(1) = 0$, $f'(1) = 0$, $f'(x) \geq 0$ for $x \geq 1$ and $f'(x) \leq 0$ for $x \leq 1$. So $f(x)$ has a minimum at $x = 1$ and is increasing when going away from $x = 1$.) Figure 2 plots the functions $\ln x$ and $x - 1$ to compare them.

We first prove the inequality in (35). Consider the following chain of inequalities:

$$\begin{aligned} D(p\|q) &= \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right) &(36)\\ &= -\frac{1}{\ln 2} \sum_x p(x) \ln \left(\frac{q(x)}{p(x)} \right) &(37)\\ &\geq \frac{1}{\ln 2} \sum_x p(x) \left(1 - \frac{q(x)}{p(x)} \right) &(38)\\ &= \frac{1}{\ln 2} \left(\sum_x p(x) - \sum_x q(x) \right) &(39)\\ &\geq 0. &(40) \end{aligned}$$

The sole inequality follows because $-\ln x \geq 1 - x$ (a simple rearrangement of $\ln x \leq x - 1$). The last inequality is a consequence of the assumption that $\sum_x q(x) \leq 1$.

Now suppose that $p = q$. It is then clear that $D(p\|q) = 0$. Finally, suppose that $D(p\|q) = 0$. Then we necessarily have $\text{supp}(p) \subseteq \text{supp}(q)$, and the condition $D(p\|q) = 0$ implies that both inequalities above are saturated. So, first we can deduce that q is a probability distribution since we assumed that $\sum_x q(x) \leq 1$ and the last inequality above is saturated. Next, the inequality in the third line above is saturated, which implies that $\ln (q(x)/p(x)) = 1 - q(x)/p(x)$ for all x for which $p(x) > 0$. But this happens only when $q(x)/p(x) = 1$ for all x for which $p(x) > 0$, which allows us to conclude that $p = q$. □

We can now quickly prove several corollaries of the above theorem.

Proofs of Property 6, Theorem 12, Theorem 8. Recall in Section 2.1 that we proved that the entropy $H(X)$ takes the maximal value $\log |\mathcal{X}|$, where $|\mathcal{X}|$ is the size of the alphabet of X. The proof method involved Lagrange multipliers. Here, we can prove this result simply by computing the relative entropy $D(p_X(x)\|\{1/|\mathcal{X}|\})$, where $p_X(x)$ is the probability density of X and $\{1/|\mathcal{X}|\}$ is

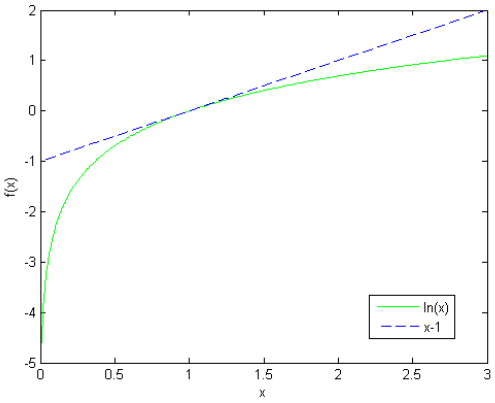


Figure 2: A plot that compares the functions $\ln x$ and $x - 1$, showing that $\ln x \leq x - 1$ for all positive x.

the uniform density, and applying the non-negativity of relative entropy:

$$
\begin{aligned}
0 &\leq D(p_X(x) \| \{1/|\mathcal{X}|\}) && (41)\\
&= \sum_x p_X(x) \log \left(\frac{p_X(x)}{\frac{1}{|\mathcal{X}|}} \right) && (42)\\
&= -H(X) + \sum_x p_X(x) \log |\mathcal{X}| && (43)\\
&= -H(X) + \log |\mathcal{X}| . && (44)
\end{aligned}
$$

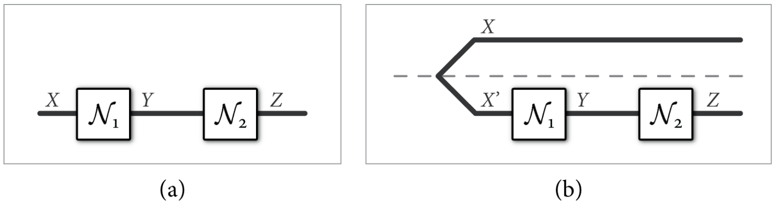
It then follows that $H(X) \leq \log |\mathcal{X}|$ by combining the first line with the last. Non-negativity of mutual information (Theorem 12) follows by recalling that $I(X;Y) = D(p_{X,Y}(x,y) \| p_X(x)p_Y(y))$ and applying the non-negativity of relative entropy. The equality conditions follow from those for equality of $D(p\|q) = 0$. Conditioning does not increase entropy (Theorem 8) follows by noting that $I(X;Y) = H(X) - H(X|Y)$ and applying Theorem 12. □

8.2 Data-Processing Inequality

Another important inequality in classical information theory is the *data-processing inequality*. There are at least two variations of it. The first one states that correlations between random variables can only decrease after we process one variable according to some stochastic function that depends only on that variable. The next one states that the relative entropy cannot increase if a channel is applied to both of its arguments. These data-processing inequalities find application in the converse proof of a coding theorem (the proof of the optimality of a communication rate).

8.2.1 Mutual Information Data-Processing Inequality

We detail the scenario that applies for the first data-processing inequality. Suppose that we initially have two random variables X and Y. We might say that random variable Y arises from random



(a) (b)

Figure 3: Two slightly different depictions of the scenario in the data-processing inequality. (a) The map $\mathcal{N}_1$ processes random variable X to produce some random variable Y, and the map $\mathcal{N}_2$ processes the random variable Y to produce the random variable Z. The inequality $I(X;Y) \geq I(X;Z)$ applies here because correlations can only decrease after data processing. (b) This depiction of data processing helps us to build intuition for data processing in the quantum world. The protocol begins with two perfectly correlated random variables X and X'—perfect correlation implies that $p_{X,X'}(x,x') = p_X(x)\delta_{x,x'}$ and further that $H(X) = I(X;X')$. We process random variable X' with a stochastic map $\mathcal{N}_1$ to produce a random variable Y, and then further process Y according to the stochastic map $\mathcal{N}_2$ to produce random variable Z. By the data-processing inequality, the following chain of inequalities holds: $I(X;X') \geq I(X;Y) \geq I(X;Z)$.

variable X by processing X according to a stochastic map $\mathcal{N}_1 \equiv p_{Y|X}(y|x)$. That is, the two random variables arise by first picking X according to the density $p_X(x)$ and then processing X according to the stochastic map $\mathcal{N}_1$. The mutual information $I(X;Y)$ quantifies the correlations between these two random variables. Suppose then that we process Y according to some other stochastic map $\mathcal{N}_2 \equiv p_{Z|Y}(z|y)$ to produce a random variable Z (note that the map can also be deterministic because the set of stochastic maps subsumes the set of deterministic maps). Then the first data-processing inequality states that the correlations between X and Z must be less than the correlations between X and Y:

$$I(X;Y) \geq I(X;Z), \tag{45}$$

because data processing according to any map $\mathcal{N}_2$ can only decrease correlations. Figure 3(a) depicts the scenario described above. Figure 3(b) depicts a slightly different scenario for data processing that helps build intuition for the forthcoming notion of quantum data-processing. Theorem 21 below states the classical data-processing inequality.

The scenario described in the above paragraph contains a major assumption and you may have picked up on it. We assumed that the stochastic map $p_{Z|Y}(z|y)$ that produces random variable Z depends on random variable Y only—it has no dependence on X, meaning that

$$p_{Z|Y,X}(z|y,x) = p_{Z|Y}(z|y). \tag{46}$$

This assumption is called the Markovian assumption and is the crucial assumption in the proof of the data-processing inequality. We say that the three random variables X, Y, and Z form a *Markov chain* and use the notation $X \to Y \to Z$ to indicate this stochastic relationship.

Theorem 21 (Data-Processing Inequality)**.** *Suppose three random variables X, Y, and Z form a Markov chain: $X \to Y \to Z$. Then the following data-processing inequality applies:*

$$I(X;Y) \geq I(X;Z). \tag{47}$$

Proof. The Markov condition $X \to Y \to Z$ implies that random variables X and Z are conditionally independent through Y because

$$p_{X,Z|Y}(x,z|y) = p_{Z|Y,X}(z|y,x)p_{X|Y}(x|y) \tag{48}$$

$$= p_{Z|Y}(z|y)p_{X|Y}(x|y). \tag{49}$$

We prove the data-processing inequality by manipulating the mutual information $I(X;YZ)$. Consider the following equalities:

$$I(X;YZ) = I(X;Y) + I(X;Z|Y) = I(X;Y). \tag{50}$$

The first equality follows from the chain rule for mutual information (Exercise 19). The second equality follows because the conditional mutual information $I(X;Z|Y)$ vanishes for a Markov chain $X \to Y \to Z$—i.e., X and Z are conditionally independent through Y (recall Theorem 17). We can also expand the mutual information $I(X;YZ)$ in another way to obtain

$$I(X;YZ) = I(X;Z) + I(X;Y|Z). \tag{51}$$

Then the following equality holds for a Markov chain $X \to Y \to Z$ by exploiting (50):

$$I(X;Y) = I(X;Z) + I(X;Y|Z). \tag{52}$$

The inequality in Theorem 21 follows because $I(X;Y|Z)$ is non-negative for any random variables X, Y, and Z (recall Theorem 17). □

By inspecting the above proof, we find the following:

Corollary 22. *The following inequality holds for a Markov chain $X \to Y \to Z$:*

$$I(X;Y) \geq I(X;Y|Z). \tag{53}$$

8.2.2 Relative Entropy Data-Processing Inequality

Another kind of data-processing inequality holds for the relative entropy, known as monotonicity of relative entropy. This also is a consequence of the non-negativity of relative entropy in Theorem 20.

Corollary 23 (Monotonicity of Relative Entropy)**.** *Let p be a probability distribution on an alphabet $\mathcal{X}$ and let $q : \mathcal{X} \to [0,\infty)$. Let $N(y|x)$ be a conditional probability distribution (i.e., a classical channel). Then the relative entropy does not increase after the channel $N(y|x)$ acts on p and q:*

$$D(p\|q) \geq D(Np\|Nq), \tag{54}$$

where Np is a probability distribution with elements $(Np)(y) \equiv \sum_x N(y|x)p(x)$ and Nq is a vector with elements $(Nq)(y) = \sum_x N(y|x)q(x)$. Let R be the channel defined by the following set of equations:

$$R(x|y)(Nq)(y) = N(y|x)q(x). \tag{55}$$

The inequality in (54) is saturated (i.e., $D(p\|q) = D(Np\|Nq)$) if and only if $RNp = p$, where RNp is a probability distribution with elements $(RNp)(x) = \sum_{y,x'} R(x|y)N(y|x')p(x')$.

Proof. First, if p and q are such that $\operatorname{supp}(p) \not\subseteq \operatorname{supp}(q)$, then the inequality is trivially true because $D(p\|q) = +\infty$ in this case. So let us suppose that $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$, which implies that $\operatorname{supp}(Np) \subseteq \operatorname{supp}(Nq)$. Our first step is to rewrite the terms in the inequality. To this end, consider the following algebraic manipulations:

$$\begin{aligned}
&D(Np\|Nq) \\
&= \sum_y (Np)(y) \log\left(\frac{(Np)(y)}{(Nq)(y)}\right) && (56)\\
&= \sum_{y,x} N(y|x)p(x) \log\left(\frac{(Np)(y)}{(Nq)(y)}\right) && (57)\\
&= \sum_x p(x) \left[\sum_y N(y|x) \log\left(\frac{(Np)(y)}{(Nq)(y)}\right)\right] && (58)\\
&= \sum_x p(x) \log\exp\left[\sum_y N(y|x) \log\left(\frac{(Np)(y)}{(Nq)(y)}\right)\right]. && (59)
\end{aligned}$$

This implies that

$$D(p\|q) - D(Np\|Nq) = D(p\|r), \tag{60}$$

where

$$r(x) \equiv q(x) \exp\left[\sum_y N(y|x) \log\left(\frac{(Np)(y)}{(Nq)(y)}\right)\right]. \tag{61}$$

Now consider that

$$\begin{aligned}
\sum_x r(x) &= \sum_x q(x) \exp\left[\sum_y N(y|x) \log\left(\frac{(Np)(y)}{(Nq)(y)}\right)\right] && (62)\\
&\le \sum_x q(x) \sum_y N(y|x) \exp\left[\log\left(\frac{(Np)(y)}{(Nq)(y)}\right)\right] && (63)\\
&= \sum_x q(x) \sum_y N(y|x) \left(\frac{(Np)(y)}{(Nq)(y)}\right) && (64)\\
&= \sum_y \left[\sum_x q(x)N(y|x)\right] \frac{(Np)(y)}{(Nq)(y)} && (65)\\
&= \sum_y (Np)(y) && (66)\\
&= 1. && (67)
\end{aligned}$$

The inequality in the second line follows from convexity of the exponential function. Since r is a vector such that $\sum_x r(x) \le 1$, we can conclude from Theorem 20 that $D(p\|r) \ge 0$, which by (60) is the same as (54).

We now comment on the saturation conditions. First, suppose that $RNp = p$. By monotonicity of relative entropy (what was just proved) under the application of the channel R, we find that

$$D(Np\|Nq) \ge D(RNp\|RNq) = D(p\|q), \tag{68}$$

where the equality follows from the assumption that $RNp = p$, and the fact that $RNq = q$. This last statement follows because

$$(RNq)(x) = \sum_y R(x|y)(Nq)(y) = \sum_y N(y|x)q(x) = q(x). \tag{69}$$

The other implication $D(p\|q) = D(Np\|Nq) \Rightarrow RNp = p$ is a consequence of a later development (see book). □

8.3 Continuity of Entropy

That the entropy is a continuous function follows from the fact that the function $-x \log x$ is continuous. However, it can be useful in applications to have explicit continuity bounds. Before we give such bounds, we should establish how we measure distance between probability distributions. A natural way for doing so is to use the classical trace distance, defined as follows:

Definition 24 (Classical Trace Distance)**.** *Let $p, q : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a finite alphabet. The classical trace distance between p and q is then*

$$\|p - q\|_1 \equiv \sum_x |p(x) - q(x)|\,. \tag{70}$$

The classical trace distance is a special case of the trace distance for quantum states, in which we place the entries of p and q along the diagonal of some square matrices. If p and q are probability distributions, then the operational meaning of the classical trace distance is the same as it was in the fully quantum case: it is the bias in the probability with which one could successfully distinguish p and q by means of any binary hypothesis test.

We can now state an important entropy continuity bound.

Theorem 25 (Zhang-Audenaert)**.** *Let X and Y be discrete random variables taking values in a finite alphabet $\mathcal{A}$. Let p_X and p_Y denote their distributions, respectively. Then the following bound holds*

$$|H(X) - H(Y)| \leq T \log(|\mathcal{A}| - 1) + h_2(T), \tag{71}$$

where $T \equiv \frac{1}{2}\|p_X - p_Y\|_1$. Furthermore, this bound is optimal, meaning that there exists a pair of random variables saturating the bound for every $T \in [0, 1]$ and alphabet size $|\mathcal{A}|$.