THERE are a lot of documents that should be brought together into this one!! ALSO LOOK at my PROSPECTUS! AND INFO theory assignments. AND the two overleaf documents.

Chapter (CITE) introduced a number of divergences that can be used to quantify how different two probability distributions are from each other. One of them was *information divergence* (I-divergence), a concept with roots in communication theory. (SIDENOTE: The I-divergence is more commonly known as *Kullback-Leibler divergence* ("KL-divergence") or *relative entropy.* I prefer the term I-divergence because it comports with the I-projection terminology introduced later in this section.) However, the I-divergence, along with other concepts from communication theory, turns out to be central to a number of important questions in probability and statistics. The study of these concepts at the intersection of communication and probability comprises a field in its own right, known as *information theory.*

In this chapter, we'll discuss entropy, I-divergence, and mutual information, the core quantities of information theory, and point out some of the ways in which they arise in answers to communication and probability questions; you'll also find them continuing to pop up in future sections and chapters.

1 Definitions and basics

1.1 Entropy

(MAKE SURE I DEFINE "discrete set" in Chapter 0 and "discrete random variable" in Chapter 1)

Let X be a discrete random variable with density p with respect to the counting measure. The *entropy* of X is defined to be (SIDENOTE: For convenience, we will often simply use a probability measure as the argument (e.g. $H(P)$).)

$$H(X) := \mathbb{E} \log \frac{1}{p(X)} \tag{1}$$

(SIDENOTE: There are various notions of "entropy," especially in physics. A more specific name for (1) is *Shannon entropy.*) Notice that the actual values taken by X don't matter; only the probabilities are relevant. In fact, X doesn't have to be real-valued; it can be any discrete random element. And entropy is invariant under any one-to-one mapping f (from the support of X to *any* space): $H(X) = H(f(X))$.

1. Find the entropy of the Bernoulli distributions as a function of the Bernoulli parameter θ. What is the entropy when $\theta = 1/2$?

INCLUDE A PLOT (like in Cover and Thomas) of the Bernoulli entropy in bits as a function of θ.

If you work out Exercise (NUMBER), you will find that the entropy of a fair coin flip is $\log 2$. For now, let's assume that we are using the base-2 logarithm; the unit of measurement in that case is called "bits." So the entropy of a fair coin flip is 1 bit. If you repeat this exercise for a uniformly distributed random variable with a support of 4 elements, you will find that the entropy is 2 bits. With a support of 8 elements, the entropy is 3 bits. In each of these cases, the entropy is telling you how many symbols you would need to use to uniquely encode all of the possible elements with an equal codelength; each set of symbols representing an element is called the *codeword* for that element. (PICTURE OF TREE FOR 4-symbol CASE.)

We can extend this interpretation of entropy to arbitrary discrete distributions, if we take note of a few details. First, if the distribution isn't uniform, then we don't necessarily want equal codelengths. Instead. let's say we want the most efficient prefix-free code that we can achieve. A code is prefix-free if you can always tell when you're at the end of a codeword without having to look ahead; you can be sure that a code is prefix-free if every codeword is a leaf on the tree (REFER to PICTURES), meaning that it has no children. But which prefix-free code is most efficient?

To answer this question, let us observe that there is an equivalence between distributions and prefix-free codes. Given any discrete distribution P, one can construct a prefix-free code that assigns codewords of $l(x) = log(1/p(x))$ bits to each $x \in \mathcal{X}$. The entropy is the expected codelength when using this code when the data is generated by P. (SIDENOTE: We might want to refer to the $log(1/p(x))$ values as *idealized* codelengths, because we can't actually create those codes if the codelengths aren't integers. In that case, you can get arbitrarily close to the *idealized* expected codelength $H(P)$ by grouping consecutive symbols together before encoding them. This is an important practical detail, but it isn't important in most theoretical regards.) Likewise, given any prefix-free code for x that has codelengths $l(x)$, one can invert the above relationship to get $p(x) = e^{-l(x)}$; if the code isn't wasteful (i.e. if the tree terminates), these $p(x)$ values sum to one. (SIDENOTE: Otherwise if the code is wasteful, then the $p(x)$ sum to less than 1 and P is called a subprobability distribution.) The tree picture is the key to seeing this equivalence between prefix-free codelengths and probability distributions. Given any prefix-free coding, the corresponding set of exponential negative codelengths ($e^{-l(x)}$) is called the *coding distribution.*

In fact, the entropy is the best possible [idealized] expected prefix-free codelength. How do we know that no better code is possible? Let P be the distribution generating your data and Q be the coding distribution that you are using. Then the expected codelength is (SIDENOTE: This quantity is also known as the *cross entropy.*)

$$\begin{aligned}\mathbb{E}_P \log \frac{1}{q(X)} &= \mathbb{E}_P \log \frac{p(X)}{q(X)} + \mathbb{E}_P \log \frac{1}{p(X)} \\ &= D(P\|Q) + H(P) \end{aligned} \tag{2}$$

Because D is non-negative (recall CITE EXERCISE), this codelength is uniquely minimized when the coding distribution Q is equal to the generating distribution P. (SIDENOTE: Here we're thinking of the logs as base-2 but the I-divergence is still non-negative; it's simply being measured in a different unit that is a positive scalar multiple ($\log_e 2$, to be precise) of the usual units.)

Entropy seems like a natural way of quantifying "information" because it represents the expected number of bits needed to specify the value of a random variable. This idea motivates the term *information theory* to refer to the field of study summarized in this section.

CODE-LENGTHS - make a note of the bits/nats complications! AND the fact that the log reciprocal probabilities aren't typically integers. both of these details are conceptually unimportant for most purposes.

We will also find (1) a useful quantity when X has a continuous distribution (and its density is with respect to Lebesgue measure), although in that case it behaves a little differently. For instance, it can be negative and it isn't invariant under one-to-one mappings. (SIDENOTE: An alternative definition of entropy for continuous random variables that retains more of the familiar entropy interpretation is based on a concept called the "limiting density of discrete points.") When the distribution is continuous, the quantity (1) is often called *differential entropy* and denoted by a lower case h. We also use the lower case h when the random variable could be either discrete or continuous.

ANDREW's interpretation relating h to coding/knowledge.
MAKE SURE the observation $h(X + a) = h(X)$ comes up in an exercise. What about $h(BX)$? USE the LOCATION-SCALE families stuff from the previous section to address this. POINT out that this is different from the discrete case where one-to-one mappings don't affect entropy. Discrete distributions:

2. Express entropy in terms of a geometric expectation (CITE EXERCISE from section 1-2).

AN OVERALL codelength is the same as a sum of conditional codelengths. SEE DISCUSSION in my prospectus.
The *conditional entropy* of Y given another random variable X is defined to be

$$h(Y|X) := \mathbb{E} \log \frac{1}{p(Y|X)} \tag{3}$$

Note that unlike conditional expectations (e.g. $\mathbb{E}[Y|X]$), the conditional entropy is not random; the expectation in (3) is being taken over both X and Y.

In the discrete case, conditional entropy can be interpreted as the expected code-length for Y assuming that both yourself and the message receiver will learn the value of X and that you will then use an optimal code for Y given the observed value of X.
WHAT ABOUT "joint entropy" - cover it but point out that it's really nothing new! Just the entropy of a random vector.

3. Let X be a mixture of $X_1, X_2, \ldots$ with mixing probabilities $p_1, p_2, \ldots$; let θ be the random variable representing the selection of $j \in \{1, 2, \ldots\}$. Assume the X_j take values on disjoint supports from each other. Show that

   $$H(X) = H(\theta) + \sum_j p_j H(X_j)$$

   Then devise $X_1, X_2, \ldots$ and $p_1, p_2, \ldots$ such that $H(X) = \infty$.

4. Show that we can bound the entropy of P by

   $$\mathbb{E}_{X\sim P} f(X) + \mu e^{-f}$$

   for any function f such that $\mathbb{E}_{X\sim P} f(X) > -\infty$. (SIDENOTE: Here μ represents counting measure for discrete entropy and Lebesgue measure for differential entropy.) [Hint: Use the fact that cross entropies are always larger than entropy, as seen in (2).]

1.2 I-divergence

The I-divergence from P to Q is defined as

$$D(P\|Q) := \mathbb{E}_P \log \frac{p(X)}{q(X)} \tag{4}$$

(MAKE SURE I've pointed out and justified the convention $0 \log 0 = 0$.) We can see a communication theory interpretation from expressing D as cross entropy minus entropy as in (2): $D(P\|Q)$ measures the expected "price" in that you would pay by using a coding distribution for Q when the true distribution is P. (SIDENOTE: Notice that if any set has positive P measure but zero Q measure, then $D(P\|Q) = \infty$. In our communication theory interpretation, such a Q just can't code for P because it doesn't have a codeword for one (or more) of P's possible selections.)

Recall that $D(\cdot\|\cdot)$ is an f-divergence (with either order of the arguments!), so it inherits the f-divergence properties we showed in section (CITE SECTION). The fact that $D(P\|Q) \geq 0$ is called the *information inequality* or the *Gibbs inequality.*

ALSO TELL the reader that I-divergence is a Bregman divergence (CITE the section where it is covered) - YOU will show this in Exercise (NUMBER). SO POINT out that I-divergence inherits the properties we've already established for Bregman divergences. LIST the properties.

REMIND reader of some inequalities established in CITE Sections/Exercises IF there are any relevant ones.

FINALLY, Gather more facts and properties:
CAN be written as an expectation of a non-negative quantity. This justifies interchanges of order of integration!

Pinsker's inequality says that total variation distance can be bounded by a function of I-divergence. We've seen in Exercise (NUMBER in section 2-2) that I-divergence bounds squared Hellinger divergences and in Exercise (NUMBER in section 2-7) the squared Hellinger distance bounds squared total variation distance, so we can already conclude that $D(P\|Q) \geq d^2_{\mathrm{TV}}(P, Q)$. This can be strengthened slightly to

$$d_{\mathrm{TV}}(P, Q) \leq \sqrt{\frac{D(P\|Q)}{2}}$$

(CITE Pollard section 3.3) (SIDENOTE: And of course, because d_{TV} is symmetric, it is also bounded by $\sqrt{D(Q\|P)/2}$.) This tells us that convergence in I-divergence implies convergence in d_{TV}. (SIDENOTE: But there is no possible inequality in the other direction as long as the sample space has more than one atom. For instance, let Q concentrate all its mass on $x \in \Omega$, while P puts $1 - \epsilon$ mass on x and $\epsilon > 0$ mass on y. Then $d_{\mathrm{TV}}(P, Q) = \epsilon$ while $D(P\|Q)$ is infinite no matter how small ϵ gets.)

I-divergence is the only f-divergence that has iterative projection on linear subspaces, or something like that... see link:

1.3 Mutual information

The mutual information between X and Y is the entropy of Y minus the conditional entropy of Y given X.

$$I(X; Y) := h(Y) - h(Y|X)$$

You can think of it as the expected reduction in entropy of Y that you would achieve by learning X.

5. Show that

   $$I(X; Y) = D(P_{X,Y} \| P_X P_Y)$$

   where $P_{X,Y}$ is the joint distribution and P_X and P_Y are the marginals.

The identity from Exercise (CITE) shows us that mutual information is symmetric in its arguments: $I(X; Y) = I(Y; X)$. It also tells us that it is non-negative, which implies that conditional entropies can't be larger than the unconditional entropy. Finally, recall that relative entropy is zero iff the arguments are the same distribution; in that case, that means $P_{X,Y} = P_X P_Y$. This means that $I(X; Y)$ is zero iff X and Y are independent; mutual information can be thought of as quantifying the amount of dependence between its arguments. And the conditional entropy $h(Y|X)$ equals $h(Y)$ iff X and Y are independent.

It can be shown that $I(X; Y)$ is convex in $P_{Y|X}$ for any fixed P_X and that it is concave in P_X for any fixed $P_{Y|X}$. See (CITE Cover and Thomas Theorem 2.7.4) for the proofs.
Entropies, conditional entropies, and mutual informations can be visualized using Venn diagrams. EXPLAIN AND ILLUSTRATE.

Make sure I include the important stuff from Cover and Thomas Chapter 2, through section 2.5, AND Conditional mutual information (and its nonnegativity, Cover and Thomas Corollary 2.92) AND Theorem 2.6.6.

6. Show that entropy is strictly concave. That is given any P_1 and P_2 that are distinct probability measures on some measurable space, show that

   $$h(\lambda P_1 + (1 - \lambda)P_2) > \lambda h(P_1) + (1 - \lambda)h(P_2)$$

   for any $\lambda \in (0, 1)$

7. Show that

   $$I(X; Y) = \mathbb{E}_X D(P_{Y|X} \| P_Y)$$

   where $P_{Y|X}$ is the conditional distribution of Y given X (which is random because it is a function of X).

8. The *generalized I-divergence* extends I-divergence to the set of all finite signed measures on a subset of $(\mathbb{R}, \mathcal{B})$ or on a discrete space. It is defined by

   $$D(P\|Q) := \mu p \log \frac{p}{q} - \mu p + \mu q$$

   where p and q are densities of signed measures P and Q with respect to μ, which is either Lebesgue or counting measure. (SIDENOTE: Notice that when P and Q are probability measures, this reduces to the ordinary I-divergence.) We can similarly define *generalized entropy* as $h(P) := \mu p \log(1/p)$. Show that D is a Bregman divergence with $\phi(P) = -h(P) - \mu p$. (SIDENOTE: For the definition of Bregman divergence, in this case use the L_2 inner product: $\langle P, Q\rangle := \mu pq$.)

9. Show that

   $$\mathbb{E}D(Q\|P_\theta) = \mathbb{E}D(\tilde{P}_\theta\|P_\theta) + D(Q\|\tilde{P}_\theta)$$

10. Explain why

    $$\mathbb{E}D(P_\theta\|Q_\alpha) = \mathbb{E}D(\tilde{Q}_\alpha\|Q_\alpha) + \mathbb{E}D(P_\theta\|\bar{P}_\theta) + D(\bar{P}_\theta\|\tilde{Q}_\alpha)$$

Venn diagram picture - but warn the reader not to put too much stock into this picture when there are more than two variables! ISN'T IT possible to get negative $I(X, Y, Z)$? IF SO, give an example maybe. I GUESS $I(\cdot, \cdot, \cdot)$ isn't a "mutual information" anyway - the definition is limited to two RVs.

11. Find the [differential] entropy of a multivariate normal random variable. Use this to derive an expression for the mutual information $I(X, Y)$ of jointly normal random variables X and Y in terms of their correlation.

DATA processing inequality - Cover and Thomas sections 2.8-2.9 THIS IS WRITTEN up somewhere! Cover the mutual information formulation and the relative entropy formulation!!! And define sufficient statistic.

FANO's inequality - Cover and Thomas 2.10 TAKE this from my minimax document. (WHAT is its relation to mutual information, from that document??? THAT is a good way to make it fit in this subsection.) I DON'T think I have a proof typed up, do I? Probably not worth including, but point to Cover and Thomas.

2 I-divergence geometry

I HAVE A LOT TYPED UP FOR THIS in my documents.
GOOD DESCRIPTION of the relationship between I-projection and rI-projection for intersection of linear and exponential families - MAKES it seem like this result is very much like the Hilbert space projection idea.

12. Show that the set of distributions that have Q as their I-projection onto $\mathcal{S}$ is log-convex.

Sometimes there is no I-projection or rI-projection *in* the set, but there is a unique distribution (IN THE INFORMATION CLOSURE?) that behaves like the projection. (ARE THESE the I-projection and rI-projection onto the INFO and rINFO closures of the set?)
Let Q be a PM and $\mathcal{S}$ be a set of PMs on the same measurable space. If there exists a unique Q^* such that every sequence $(P_n) \subseteq \mathcal{S}$ for which $D(P_n\|Q) \to D(\mathcal{S}\|Q)$ I-converges to Q^* (that is, $D(P_n\|Q^*) \to 0$), then Q^* is called the generalized I-projection of Q onto $\mathcal{S}$. Alternatively, if there exists a unique Q^* such that every sequence $(P_n) \subseteq \mathcal{S}$ for which $D(Q\|P_n) \to D(Q\|\mathcal{S})$ rI-converges to Q^* (that is, $D(Q^*\|P_n) \to 0$), then Q^* is called the generalized rI-projection of Q onto $\mathcal{S}$. (SIDENOTE: When the I-projection exists, it is also the generalized I-projection, of course. Likewise for rI-projections.)
FROM I-PROJ revisited THEOREM 1: - CHECK IF THIS theorem is specific to prob measures or if it extends to general signed measures!

Theorem 2.1. *Let Q be a probability measure on a measurable space $(\mathcal{X}, \mathcal{A})$. If $\mathcal{S}$ is a convex set of probability measures on $(\mathcal{X}, \mathcal{A})$, then there exists a unique PM $Q_\mathcal{S}$ such that*

$$D(P\|Q) \geq D(P\|Q_\mathcal{S}) + D(\mathcal{S}\|Q)$$

for any $P \in \mathcal{S}$. If $\mathcal{S}$ is a log-convex set of probability measures on $(\mathcal{X}, \mathcal{A})$, then there exists a unique $Q^\mathcal{S}$ such that

$$D(Q\|P) \geq D(Q\|\mathcal{S}) + D(Q^\mathcal{S}\|P)$$

for any $P \in \mathcal{S}$.

CAN IT BE PROVEN by taking derivatives? (one of these can but i don't know about the other) also look at cziszar's proof - neat stuff about log-mixtures in there.
EXERCISE - prove theorem (or at least part of it)

13. Explain why $Q_\mathcal{S}$ is a generalized I-projection from Q to $\mathcal{S}$ and why $Q^\mathcal{S}$ is a generalized rI-projection from Q to $\mathcal{S}$.

OUT of curiosity, IS $D(\mathcal{S}\|Q) = D(Q^*\|Q)$?? Likewise for generalized rI-projection?

2.1 Maximum entropy

FINDING max ent distributions. WHY do people care about these? IT is a method for inference - discuss this in a future chapter.

14. Let U and X both have the same support $\mathcal{X}$, and let U be uniformly distributed. Show that

    $$h(U) = h(X) + D(P_X\|P_U) \tag{5}$$

By the information inequality and the fact that $I-divergence$ separates points, we see that the uniform distribution's entropy is strictly greater than than that of any other distribution on $\mathcal{X}$. Furthermore, we can explicitly state the entropy of U because we know that the density for U must be $(\mu\mathcal{X})^{-1}$

$$\begin{aligned} h(U) &= \mathbb{E} \log \frac{1}{p_U(X)} \\ &= \mathbb{E} \log \mu\mathcal{X} \\ &= \log \mu\mathcal{X} \end{aligned}$$

The entropy of a uniform distribution is the log of the "size" of its support, quantified by $\mu\mathcal{X}$. If X is discrete, then μ is the counting measure: $\mu\mathcal{X} = |\mathcal{X}|$. If X is continuous, then μ is Lebesgue measure; if $\mathcal{X}$ is, for example, an interval (a, b), then $\mu\mathcal{X} = (b - a)$.

With this expression for $h(U)$, let's refine (5). When a uniform distribution exists with the same support as X, the entropy of X can be expressed as

$$h(X) = \log \mu\mathcal{X} - D(P_X\|P_U)$$

15. Which Bernoulli distribution has the largest entropy?

16. Is the family of Bernoulli distributions concave in the Bernoulli parameter θ?

However, sometimes you are interested in a class of distributions for which a uniform distribution isn't possible (e.g. the real line).
For an fixed variance, the Gaussian is the distribution with the largest entropy. And conversely, for any fixed entropy, the Gaussian is the distribution with the smallest variance! (A similar statement holds for all of these exponential max-ent distributions!) THIS FOLLOWS from the statement that

$$H(X) \leq \frac{1}{2} \log 2\pi e \mathbb{V}(X)$$

with equality iff X is Gaussian.

17. Let X and Y be marginally Normal (but not necessarily jointly Normal) random variables with correlation $\rho_{X,Y}$. Use the result of Exercise (CITE) to find a tight lower bound on the mutual information $I(X, Y)$ in terms of $\rho_{X,Y}$.

2.2 I-divergence projections

I-projections and rI-projections!!!
More topics for this section:
I think Stirling's formula belongs in here somewhere as an exercise.

Solution

1. Let P_θ be the Bernoulli(θ) distribution. It only takes two possible values, so the expectation is easy to evaluate.

   $$\begin{aligned} H(P_\theta) &= \mathbb{E} \log \frac{1}{p_\theta(X)} \\ &= \mathbb{P}(X = 1) \log \frac{1}{p_\theta(1)} + \mathbb{P}(X = 0) \log \frac{1}{p_\theta(0)} \\ &= \theta \log \frac{1}{\theta} + (1 - \theta) \log \frac{1}{1 - \theta} \end{aligned}$$

2. The entropy is the log of the geometric expectation of one over the density.

   $$\begin{aligned} h(X) &= \mathbb{E} \log \frac{1}{p(X)} \\ &= \log e^{\mathbb{E} \log(1/p(X))} \\ &= \log \tilde{\mathbb{E}} \frac{1}{p(X)} \end{aligned}$$

3. Because the X_j take values on disjoint supports, knowledge of X tells you exactly what value θ must have.

   $$\begin{aligned} H(X) &= H(X, \theta) - \underbrace{H(\theta|X)}_{\text{zero}} \\ &= H(X, \theta) \\ &= H(\theta) + H(X|\theta) \\ &= H(\theta) + \sum_j p_j H(X_j) \end{aligned}$$

   To make the second term infinite, we need the $H(X_j)$ to grow quickly enough and the p_j to diminish slowly enough. There are plenty of choices. One simple example is to let each X_j be uniformly distributed over a set of size 2^j, so that its entropy is j, while letting the p_j be proportional to $1/j^2$. Then the sum comprises terms proportional to $1/j$, which is a divergent series.

4. Entropy is bounded by all cross entropies. So for the probability density $q := e^{-f}/\mu e^{-f}$,

   $$\begin{aligned} h(P) \leq \mathbb{E} \log \frac{1}{q(X)} & \\ &= \mathbb{E} \log \frac{1}{e^{-f(X)}/\mu e^{-f}} \\ &= \mathbb{E}f(X) + \log \mu e^{-f} \end{aligned} \tag{6}$$

   This might enable you to bound entropy in terms of more familiar or easier to compute quantities. Conversely, you could also use this inequality to lower bound the expectations of various functions in terms of the entropy.

   As an example, consider $f(x) := \frac{(x-\mu)^2}{2\sigma^2}$ where μ and σ^2 are the mean and variance for P. The first term is

   $$\begin{aligned} \mathbb{E}f(X) &= \mathbb{E} \frac{(X - \mu)^2}{2\sigma^2} \\ &= \frac{1}{2\sigma^2} \underbrace{\mathbb{E}(X - \mu)^2}_{\sigma^2} \\ &= \frac{1}{2} \end{aligned}$$

   If we're dealing with differential entropy, then the other term is a Lebesgue integral:

   $$\begin{aligned} \log \mu e^{-f} &= \log \int e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \log \sqrt{2\pi}\sigma \end{aligned}$$

   In fact, this particular sum $\frac{1}{2} + \frac{1}{2} \log 2\pi\sigma^2$ is exactly the Gaussian entropy. You'll see in section (REFERENCE) another explanation for why any continuous random variable has to have its entropy bounded by this function of its variance.

   The inequality (6) can also be used to check that entropy is finite: look for an f such that both terms are finite. When exactly is μe^{-f} finite? For the countable case, we can have $f(n) = -\log s_n$ for any convergent series: $\sum s_n < \infty$. Likewise, for the continuous case, $f(x) = -\log s(x)$ works if s is Lebesgue integrable.

5. In the derivation below, the expectation refers to the joint distribution $P_{X,Y}$.

   $$\begin{aligned} I(X; Y) &:= h(Y) - h(Y|X) \\ &= \mathbb{E} \log \frac{1}{p(Y)} - \mathbb{E} \log \frac{1}{p(Y|X)} \\ &= \mathbb{E} \log \frac{p(Y|X)}{p(Y)} \\ &= \mathbb{E} \log \frac{p(X)p(Y|X)}{p(X)p(Y)} \\ &= \mathbb{E} \log \frac{p(X, Y)}{p(X)p(Y)} \\ &= D(P_{X,Y}\|P_X P_Y) \end{aligned}$$

6. The following clever and elegant proof is from (CITE Cover and Thomas Theorem 2.7.3).

   Let θ be a random variable taking values 1 and 2 with probabilities λ and $1 - \lambda$. Consider the random distribution P_θ. Unconditionally, it is the mixture $\lambda P_1 + (1 - \lambda)P_2$. We simply need to invoke the fact that conditional entropies are smaller than entropies.

   $$\begin{aligned} h(\lambda P_1 + (1 - \lambda)P_2) &= h(P_\theta) \\ &\geq h(P_\theta|\theta) \\ &= \lambda h(P_1) + (1 - \lambda)h(P_2) \end{aligned}$$

   Because P_θ depends on θ, the inequality is actually strict.

7. This derivation is straight-forward if you start with (CITE Exercise number that gives the mutual info as a relative entropy).

   $$\begin{aligned} I(X; Y) &= D(P_{X,Y}\|P_X P_Y) \\ &= \mathbb{E} \log \frac{p(X, Y)}{p(X)p(Y)} \\ &= \mathbb{E} \log \frac{p(X)p(Y|X)}{p(X)p(Y)} \\ &= \mathbb{E} \log \frac{p(Y|X)}{p(Y)} \\ &= \mathbb{E}_X \mathbb{E}_{Y|X} \log \frac{p(Y|X)}{p(Y)} \\ &= \mathbb{E}_X D(P_{Y|X}\|P_Y) \end{aligned}$$

8. WHAT about the condition that the set of definition should be closed? IS that satisfied here?

   We need to verify that ϕ is strictly convex on the set of finite signed measures. The second term of ϕ is linear:

   $$\mu\,(\alpha p + \beta q) = \alpha\mu\, p + \beta\mu\, q$$

   So we need the first term ($-h$) to be strictly convex (i.e. h to be strictly concave). Indeed, Exercise (NUMBER) showed that h is strictly concave for probability measures. CAN WE EXTEND THAT REASONING TO ALL FINITE SIGNED MEASURES?

   I STILL NEED TO FIGURE OUT and explain WHY $\nabla\phi(Q) = \log q$. MAYBE I SHOULD JUST TELL THE READER TO ACCEPT that gradient, and point the interested reader to a paper that [I think] covers the derivation: (I should read that paper anyway - Functional Bregman Divergence and Bayesian Estimation of Distributions)

   $$\begin{aligned} d_\phi(P, Q) &= \phi(P) - \phi(Q) - \langle\nabla\phi(Q), p - q\rangle \\ &= (\mu\, p \log p - \mu\, p) - (\mu\, q \log q - \mu\, q) - \langle \log q, p - q\rangle \\ &= \mu\, p \log p - \mu\, q \log q - \mu\, p \log q + \mu\, q \log q - \mu\, p + \mu\, q \\ &= \mu\, p \log \frac{p}{q} - \mu\, p + \mu\, q \end{aligned}$$

9. COPY and paste this proof from Overleaf document. Maybe change the notation or simplify it.

   Let's abstract this reasoning into a pattern for identifying bv-decompositions. (SIDENOTE: I don't actually know of any other divergences (besides Bregman and reversed I-divergence) that have a bv-decomposition. But if I was going to look, this is the pattern I would use.) It is sufficient to express an expected divergence between a random and a constant quantity as

   $$\mathbb{E}d(X, a) = d_0(c, a) + \mathbb{E}r(X) \tag{7}$$

   where d_0 is any divergence. If d_0 is not the same divergence as a, then this is not as "pure" of a bv-decomposition, but the term can still be interpreted as bias². By plugging c in for a, we find that the remaining term $\mathbb{E}r(X)$ must be exactly $\mathbb{E}d(X, c)$ which we can interpret as a variance. Why are we happy to consider this c a "center" for X? Because its the minimizer of $\mathbb{E}d(X, a)$ (i.e. a centroid). Why? Because the second term of (7) doesn't depend on a, all we can do is minimize the first term; it attains its minimum value of zero when c is the input.

10. Because I-divergence is a Bregman divergence in its first argument, it has a bv-decomposition (REFER to section or Exercise). Exercise (NUMBER) showed that it also has a bv-decomposition in its second argument. Do the integrations in either order (justified by Tonelli's theorem), and do the two bv-decompositions in turn.

    $$\begin{aligned} \mathbb{E}D(P_\theta\|Q_\alpha) &= \mathbb{E}_\theta \mathbb{E}_\alpha D(P_\theta\|Q_\alpha) \\ &= \mathbb{E}_\theta[\mathbb{E}_\alpha D(\tilde{Q}_\alpha\|Q_\alpha) + D(P_\theta\|\tilde{Q}_\alpha)] \\ &= \mathbb{E}D(\tilde{Q}_\alpha\|Q_\alpha) + \mathbb{E}D(P_\theta\|\bar{P}_\theta) + D(\bar{P}_\theta\|\tilde{Q}_\alpha) \end{aligned}$$

11. Assume R_1 and R_2 both have Q as their I-projection. Consider the relative entropy to a log mixture from any $P \in \mathcal{S}$.

    $$\begin{aligned} D(P\|R_1 e^{\lambda \log r_2/r_1}) &= \mathbb{E}_P \log \frac{p(X)}{r_1^{1-\lambda}(X) r_2^\lambda(X)/A_\lambda(R_2\|R_1)} \\ &= \mathbb{E}_P \log \frac{p^{1-\lambda}(X)p^\lambda(X)}{r_1^{1-\lambda}(X) r_2^\lambda(X)} + \log \frac{1}{A_\lambda(R_2\|R_1)} \\ &= (1 - \lambda)D(P\|R_1) + \lambda D(P\|R_2) + \log \frac{1}{A_\lambda(R_2\|R_1)} \end{aligned}$$

    Only the first two terms depend on P, and in both cases, Q is the minimizing choice. Thus Q is the I-projection of all of the log mixtures between R_1 and R_2.

    One interesting implication of this fact is that if R has Q as its I-projection, then all of the distributions along the log-mixture path from R to Q also have Q as their I-projection. This follows from our result because trivially Q is its own I-projection.

    REVERSE I-projection seems to travel along mixture lines while I-projection seems to travel ALONG LOG-MIXTURE CURVES. THINK ABOUT THIS and WRITE about it.

12. A rearrangement of the first Pythagorean inequality from Theorem 2.1 is

    $$D(P\|Q) - D(\mathcal{S}\|Q) \geq D(P\|Q_\mathcal{S})$$

    Assume $D(P_n\|Q) \to D(\mathcal{S}\|Q)$. Then with P_n as P, the left-hand side of the inequality goes to zero, so the right-hand side must as well, showing that (P_n) I-converges to $Q_\mathcal{S}$.
    A rearrangement of the second Pythagorean inequality from Theorem 2.1 is

    $$D(Q\|P) - D(Q\|\mathcal{S}) \geq D(Q_\mathcal{S}\|P)$$

    Assume $D(Q\|P_n) \to D(Q\|\mathcal{S})$. Then with P_n as P, the left-hand side of the inequality goes to zero, so the right-hand side must as well, showing that (P_n) rI-converges to $Q_\mathcal{S}$.

13. We'll expand the relative entropy

    $$\begin{aligned} D(P_X\|P_U) &= \mathbb{E} \log \frac{p_X(X)}{p_U(X)} \\ &= \mathbb{E} \log \frac{1}{p_U(X)} - \mathbb{E} \log \frac{1}{p_X(X)} \\ &= \mathbb{E} \log \frac{1}{p_U(U)} - \mathbb{E} \log \frac{1}{p_X(X)} \\ &= h(U) - h(X) \end{aligned}$$

    The tricky step requires realizing that $p_U(X) = p_U(U)$ because p_U is constant on their shared support $\mathcal{X}$.

14. A Bernoulli random variable can only take two values. We've seen from Exercise (CITE) that uniform distributions maximize entropy, so the Bernoulli(1/2) must be the entropy-maximizer. More specifically,

    $$\begin{aligned} H(P_\theta) &= H(P_{1/2}) - D(P_\theta\|P_{1/2}) \\ &= 1 \text{ bit } - D(P_\theta\|P_{1/2}) \end{aligned}$$

15. We've already seen in Exercise (Number) that entropy is concave on the set of all probability measures on a given measurable space. Taking convex combinations of θ is equivalent to taking the corresponding convex mixture of Bernoulli distributions:

    $$\text{Bernoulli}(\lambda\theta_1 + [1 - \lambda]\theta_2) \equiv \lambda\text{Bernoulli}(\theta_1) + [1 - \lambda]\text{Bernoulli}(\theta_2)$$

    So the entropy is concave in θ:

    $$\begin{aligned} H(\text{Bernoulli}(\lambda\theta_1 + [1 - \lambda]\theta_2)) &= H(\lambda\text{Bernoulli}(\theta_1) + [1 - \lambda]\text{Bernoulli}(\theta_2)) \\ &> \lambda H(\text{Bernoulli}(\theta_1)) + [1 - \lambda]H(\text{Bernoulli}(\theta_2)) \end{aligned}$$

16. Letting P be the d-dimensional Normal distribution with mean vector μ and [non-degenerate] covariance matrix Σ,

    $$\begin{aligned} h(P) &= \mathbb{E} \log \frac{1}{p(X)} \\ &= \mathbb{E} \log \left[(2\pi)^{d/2}|\Sigma|^{1/2} e^{\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)}\right] \\ &= \frac{d}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma| + \frac{1}{2}\mathbb{E}(X - \mu)'\Sigma^{-1}(X - \mu) \end{aligned}$$

    Notice that each X is transformed into a standard Normal random vector Z:

    $$\begin{aligned} \mathbb{E}(X - \mu)'\Sigma^{-1}(X - \mu) &= \mathbb{E}[\Sigma^{-1/2}(X - \mu)]'[\Sigma^{-1/2}(X - \mu)] \\ &= \mathbb{E}Z'Z \\ &= \mathbb{E}\sum_{i=1}^d Z_i^2 \\ &= \sum_{i=1}^d \mathbb{E}Z_i^2 \\ &= d \end{aligned}$$

    Thus the overall entropy is $\frac{d}{2}[1 + \log(2\pi)] + \frac{1}{2} \log |\Sigma|$.

    Remark: It is interesting to note that $\mathbb{E}(X - \mu)'\Sigma^{-1}(X - \mu)$ is equal to d even when X isn't Normal, as long as it has mean μ and covariance Σ. To see this, use the fact that the trace is invariant under cyclic permutation and that expectation commutes with trace.

    $$\begin{aligned} \mathbb{E}(X - \mu)'\Sigma^{-1}(X - \mu) &= \mathbb{E}\operatorname{tr}(X - \mu)'\Sigma^{-1}(X - \mu) \\ &= \mathbb{E}\operatorname{tr}\Sigma^{-1}(X - \mu)(X - \mu)' \\ &= \operatorname{tr}\Sigma^{-1}\mathbb{E}(X - \mu)(X - \mu)' \\ &= \operatorname{tr}\Sigma^{-1}\Sigma \\ &= \operatorname{tr} I_d \\ &= d \end{aligned}$$

    Now, let X and Y be bivariate Normal random variables.

    $$\begin{aligned} I(X, Y) &= h(X) + h(Y) - h(X, Y) \\ &= \frac{1}{2}[1 + \log(2\pi)] + \frac{1}{2} \log \sigma_X^2 + \frac{1}{2}[1 + \log(2\pi)] + \frac{1}{2} \log \sigma_Y^2 \\ &\quad - \left(\frac{2}{2}[1 + \log(2\pi)] + \frac{1}{2} \log(\sigma_X^2\sigma_Y^2 - \sigma_{X,Y}^2)\right) \\ &= \frac{1}{2} \log \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2\sigma_Y^2 - \sigma_{X,Y}^2} \\ &= -\frac{1}{2} \log \frac{\sigma_X^2\sigma_Y^2 - \sigma_{X,Y}^2}{\sigma_X^2\sigma_Y^2} \\ &= -\frac{1}{2} \log[1 - \rho_{X,Y}^2] \end{aligned}$$

    where $\rho_{X,Y}$ is the correlation $\sigma_{X,Y}/(\sigma_X\sigma_Y)$. As we might expect, mutual information is a monotonically increasing function of the squared correlation.

    CITE the blog linked below.

17. First, observe that for any specified means and covariance matrix, the entropy-maximizing distribution is the multivariate Normal distribution. (THIS SHOULD BE AN EASY[?] CONSEQUENCE of the general result in the text above. BUT POINT it out explicitly. ACTUALLY it may not be so easy. It might involve rescaling uncorrelated RVs. OR DOES that log determinant formula only work for multivariate normal?)

    Let X' and Y' have the same marginal distributions of X and Y, but assume they are jointly Normal and have correlation $\rho_{X',Y'} = \rho_{X,Y}$.

    Because X' has the same marginal distribution as X, they must have the same entropy. Likewise for Y' and Y. And because the multivariate Normal distribution (X', Y') maximizes entropy over the class of distributions whose means and covariances are the same as those of (X, Y), we know that $h(X', Y') \geq h(X, Y)$. Thus

    $$\begin{aligned} I(X, Y) &= h(X) + h(Y) - h(X, Y) \\ &\geq h(X') + h(Y') - h(X', Y') \\ &= I(X', Y') \\ &= -\frac{1}{2} \log[1 - \rho_{X',Y'}^2] \\ &= -\frac{1}{2} \log[1 - \rho_{X,Y}^2] \end{aligned}$$

    A lower bound is "tight" if no greater lower bound is possible. Because jointly Normal (X, Y) achieves the bound, we know it can't be improved.

    It's perfectly intuitive that for fixed marginal entropies, maximizing the joint entropy is equivalent to minimizing the mutual information.

    CITE the blog AGAIN!