14 Foundations of Information Theory (†)

To make the arguments in the previous chapter rigorous and generalizable to other settings we need some classic tools from information theory and statistics. In particular, we will need the concept of **relative entropy**, also known as the **Kullback-Leibler divergence** named for Solomon Kullback and Richard Leibler (KL divergence, for short).

The relative entropy has several interpretations. The one we will focus on here comes from the situation encountered by Alice, who wants to communicate with Bob. She wants to tell Bob the outcome of a sequence of independent random variables sampled from known distribution Q. Alice and Bob agree to communicate using a code that is fixed in advance in such a way that the expected message length is minimized. Then the **entropy** of Q is the expected length of the optimal code. The relative entropy between distributions P and Q is the price in terms of expected message length that Alice and Bob have to pay if they believe the random variables are sampled from Q, when in fact they are sampled from P.

Let X be a random variable that takes finitely many values, which without loss of generality we will assume is $X \in [N]$. We abbreviate $p_i = \mathbb{P}(X = i)$. Let us first discuss how to define the amount of information that observing X conveys. One way to start is to define information as the amount of communication needed if we want to tell a friend about the value we observed. We'll assume that Alice observes the value of X and wants to tell Bob what value she observed using a **binary code** that they agree upon in advance. For example, if $N = 4$, then they might agree on the following code: $1 \to 00, 2 \to 01, 3 \to 10, 4 \to 11$. Then if Alice observes a 3, she sends Bob a message containing 10. For our purposes, a code is a function $c : [N] \to \{0,1\}^*$ where $\{0,1\}^*$ is the set of finite sequences of zeros and ones.

Of course we demand that c is injective so that no two numbers (or **symbols**) have the same code. We also require that c is **prefix free**, which means that no code should be the prefix of any other. This is justified by supposing that Alice would like to tell Bob about multiple samples. Then Bob needs to know where the message for one symbol starts and ends, and he would like to do this with no back-tracking.

The easiest choice is to use $\lceil \log_2(N) \rceil$ bits no matter the value of X. This simple code is sometimes effective, but is not entirely satisfactory if X is far from uniform.

To understand why, suppose that N is extremely large and $\mathbb{P}(X = 1) = 0.99$ and the remaining probability mass is uniform over $[N] - \{1\}$. Then it seems preferable to have a short code for 1 and slightly longer codes for the alternatives. With this in mind, a natural objective is to find a code that minimizes the expected code length. That is

$$c = \operatorname{argmin}_c \sum_{i=1}^{N} p_i \ell(c(i)) \,, \tag{14.1}$$

where the argmin is taken over valid codes and $\ell(\cdot)$ is a function that returns the length of a code. The optimization problem in (14.1) can be solved by using Huffman coding and the optimal value lies within the following range

$$H_2(P) \leq \min_c \sum_{i=1}^{N} p_i \ell(c(i)) \leq H_2(P) + 1 \,,$$

where $H_2(P)$ is defined by

$$H_2(P) = \sum_{i \in [N] : p_i > 0} p_i \log_2 \left(\frac{1}{p_i} \right) .$$

Notice that if P is uniform, then $p_i = 1/N$ and the naive idea of using a code of uniform length is recovered, but for non-uniform distributions the code adapts to assign shorter codes to symbols with larger probability. What is not apparent from the expression above is that the code length for symbol i when using Huffman coding is never longer than $\log(1/p_i) + 1$. It is also worth pointing out that the sum is only over outcomes that occur with non-zero probability, which is motivated by observing that $\lim_{x \to 0} x \log(1/x) = 0$ or by thinking of the entropy as an expectation of the log-probability with respect to P and expectations should not change when the value of the random variable is perturbed on a measure zero set.

It turns out that $H_2(P)$ is not just an approximation on the expected length of the Huffman code, but is itself a fundamental quantity. We will not go into this in detail, but imagine that Alice wants to send a long string of symbols to Bob. She could use a Huffman code to send Bob each symbol one at a time, but this introduces 'rounding errors' that accumulate as the message length grows. Instead they can agree on a procedure, which Bob can still interpret sequentially without backtracking, and for which the expected average code-length averaged over the whole message tends towards $H_2(P)$ as the length of the message grows. Furthermore, the celebrated source coding theorem says that you cannot do better than this. A procedure for achieving this is called **arithmetic coding**. We will not need to actually code messages, however, so the precise details are not needed. Before moving on, we will replace the base 2 logarithm with its natural

counterpart and define the entropy of a random variable X by

$$H(P) = \sum_{i \in [N] : p_i > 0} p_i \log\left(\frac{1}{p_i}\right) . \tag{14.2}$$

This is nothing more than a scaling of the H_2 that is ultimately mathematically more convenient. Measuring information using base 2 logarithms has a unit of **bits** and for the natural logarithm it is called **nats**.

We hope you agree that $H(P)$ measures the (expected) information content of observing random variables sampled from P, at least in the long run. We now move towards defining the relative entropy.

14.1 The relative entropy

Suppose that Alice and Bob agree to use a code that is optimal when X is sampled from distribution Q. Unbeknownst to them, however, X is actually sampled from distribution P. The relative entropy between P and Q measures how much longer the messages are expected to be using the optimal code for Q than what would be obtained using the optimal code for P. Letting $p_i = P(X = i)$ and $q_i = Q(X = i)$ and working out the math leads to the definition of the relative entropy as

$$\mathrm{D}(P, Q) = \sum_{i \in [N] : p_i > 0} p_i \log\left(\frac{1}{q_i}\right) - \sum_{i \in [N] : p_i > 0} p_i \log\left(\frac{1}{p_i}\right) = \sum_{i \in [N] : p_i > 0} p_i \log\left(\frac{p_i}{q_i}\right) \tag{14.3}$$

If this quantity is large, then we expect to be able to tell that $P \neq Q$ with fewer independent observations sharing P than if this quantity was smaller. For example, if there exists an i with $p_i > 0$ and $q_i = 0$, then the first time we see symbol i, we can tell with certainty that the symbol was not sampled from Q. Looking at the definition of the relative entropy shows that in this case, $\mathrm{D}(P, Q) = \infty$.

Still poking around the definition, what happens when $q_i = 0$ and $p_i = 0$? This means that the symbol i is superfluous and the value of $\mathrm{D}(P, Q)$ should not be impacted by introducing superfluous symbols. And again, it does not by the definition of the expectations. We also see that the sufficient and necessary condition for $\mathrm{D}(P, Q) < \infty$ is that for each i such that $q_i = 0$, we also have that $p_i = 0$. The condition we discovered is also expressed as saying that P is absolutely continuous with respect to Q, which is also written as $P \ll Q$. Note that absolute continuity only implies a finite relative entropy when X takes on finitely many values. If instead $X \in \{2, 3, 4, \ldots\}$ and $\mathbb{P}(X = i) \propto 1/(i \log^2(i))$, then $H(X) = \infty$.

More generally, for two measures P, Q on a common measurable space $(\Omega, \mathcal{F})$, we say that P is **absolutely continuous** with respect to Q (and write $P \ll Q$) if for any $A \in \mathcal{F}$, $Q(A) = 0$ implies that $P(A) = 0$ (intuitively, $\ll$ is like $\leq$ except that it only constrains the values when the right-hand side is also zero). This

brings us back to defining relative entropy between two arbitrary probability distributions P, Q defined over a common probability space. The difficulty we face is that if $X \sim P$ takes on uncountably infinitely many values then we cannot really use the ideas that use communication because no matter what coding we use, we would need infinitely many symbols to describe some values of X. How can the entropy of X be defined at all? This seems to be a truly fundamental difficulty. Luckily, the impasse gets resolved automatically if we only consider relative entropy. While we cannot communicate X, for any finite 'discretization' of the possible values that X can take on, the discretized values can be communicated finitely and all our definitions will work. Formally, if X takes values in the measurable space $(\mathcal{X}, \mathcal{G})$, with $\mathcal{X}$ possibly having uncountably many elements, a discretization to $[N]$ levels would be specified using some function $f : \mathcal{X} \to [N]$ that is $\mathcal{G}/2^{[N]}$-measurable map. Then, the entropy of P relative Q, $\mathrm{D}(P, Q)$ can be defined as

$$\mathrm{D}(P, Q) = \sup_f \mathrm{D}(P_f, Q_f)\,, \tag{14.4}$$

where P_f is the distribution of $Y = f(X)$ when $X \sim P$ and Q_f is the distribution of $Y = f(X)$ when $X \sim Q$ and the supremum is for all $N \in \mathbb{N}^+$ and all maps f as defined above. In words, we take all possible discretizations f (with no limit on the 'fineness' of the discretization) and define $\mathrm{D}(P, Q)$ as the excess information when expecting to see $f(X)$ with $X \sim Q$ while reality is $X \sim P$. If this is finite, then we expect this to be a reasonable definition. As we shall see it soon, it is indeed a reasonable definition.

THEOREM 14.1 *Let $(\Omega, \mathcal{F})$ be a measurable space and let P and Q be measures on this space. Then,*

$$\mathrm{D}(P, Q) = \begin{cases} \int \log\left(\frac{dP}{dQ}(\omega)\right) dP(\omega)\,, & \text{if } P \ll Q\,; \\ \infty\,, & \text{otherwise}\,. \end{cases}$$

Note that by our earlier remark, this reduces to (14.3) for discrete measures. If λ is a common dominating σ-finite measure for P and Q (that is, $P \ll \lambda$ and $Q \ll \lambda$ both hold) then letting $p = \frac{dP}{d\lambda}$ and $q = \frac{dQ}{d\lambda}$, if also $P \ll Q$, the chain rule gives $\frac{dP}{dQ}\frac{dQ}{d\lambda} = \frac{dP}{d\lambda}$, which lets us write

$$\mathrm{D}(P, Q) = \int p \log\left(\frac{p}{q}\right) d\lambda\,,$$

which is perhaps the best known expression for relative entropy and is also often used as a definition. Note that for probability measures, a common dominating σ-finite measure can always be bound. For example, $\lambda = P + Q$ always dominates both P and Q.

Relative entropy is a kind of 'distance' measure between distributions P and Q. In particular, if $P = Q$, then $\mathrm{D}(P, Q) = 0$ and otherwise $\mathrm{D}(P, Q) > 0$. Strictly

speaking, the relative entropy is not a distance because it satisfies neither the triangle inequality nor is it symmetric. Nevertheless, it serves the same purpose.

The relative entropy between many standard distributions is often quite easy to compute. For example, the relative entropy between two Gaussians with means $\mu_1, \mu_2 \in \mathbb{R}$ and common variance σ^2 is

$$\mathrm{D}(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2} .$$

The dependence on the difference in means and the variance is consistent with our intuition. If μ_1 is close to μ_2, then the 'difference' between the distributions should be small, but if the variance is very small, then there is little overlap and the difference is large. The relative entropy between two Bernoulli distributions with means $p, q \in [0, 1]$

$$\mathrm{D}(\mathcal{B}(p), \mathcal{B}(q)) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) ,$$

where $0 \log(\cdot) = 0$.

We are nearing the end of our whirlwind tour of relative entropy. It remains to state the key lemma, sometimes called the **high probability Pinsker** inequality, that connects the relative entropy to the hardness of hypothesis testing.

THEOREM 14.2 *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$ and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-\mathrm{D}(P, Q)\right) , \tag{14.5}$$

where $A^c = \Omega \setminus A$ is the complement of A.

The proof may be found at the end of the chapter, but first some interpretation and a simple application. Suppose that $\mathrm{D}(P, Q)$ is small, then P is 'close' to Q in some sense. Since P is a probability measure we have $P(A) + P(A^c) = 1$. If Q is close to P, then we might expect $P(A) + Q(A^c)$ should be large. The purpose of the theorem is to quantify just how large. Note that if P is not absolutely continuous with respect to Q then $\mathrm{D}(P, Q) = \infty$ and the result is vacuous. Also note that the result is symmetric. We could replace $\mathrm{D}(P, Q)$ with $\mathrm{D}(Q, P)$, which sometimes leads to a stronger result because the relative entropy is not symmetric.

Returning to the hypothesis testing problem described in the previous chapter. Let X be normally distributed with unknown mean $\mu \in \{0, \Delta\}$ and variance $\sigma^2 > 0$. We want to bound the quality of a rule for deciding what is the real mean from a single observation. The decision rule is characterized by a measurable set $A \subseteq \mathbb{R}$ on which the predictor guesses $\mu = \Delta$ (it predicts $\mu = 0$ on the complement of A). Let $P = \mathcal{N}(0, \sigma^2)$ and $Q = \mathcal{N}(\Delta, \sigma^2)$. Then the probability of an error under P is $P(A)$ and the probability of error under Q is $Q(A^c)$. The reader surely knows what to do next. By Theorem 14.2 we have

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-\mathrm{D}(P, Q)\right) = \frac{1}{2} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) .$$

If we assume that the signal to noise ratio is small, $\Delta^2/\sigma^2 \leq 1$, then

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp\left(-\frac{1}{2}\right) \geq \frac{3}{10}\,,$$

which implies $\max\{P(A), Q(A^c)\} \geq 3/20$. This means that no matter how we chose our decision rule, we simply do not have enough data to make a decision for which the probability of error on either P or Q is smaller than $3/20$.

Proof of Theorem 14.2 For reals a, b we abbreviate $\max\{a,b\} = a \vee b$ and $\min\{a,b\} = a \wedge b$. The result is trivial if $\mathrm{D}(P,Q) = \infty$. On the other hand, by Theorem 14.1, $\mathrm{D}(P,Q) < \infty$ implies that $P \ll Q$. Let $\nu = P + Q$. Then $P, Q \ll \nu$, which by Theorem 2.5 ensures the existence of the Radon-Nikodym derivatives $p = \frac{dP}{d\nu}$ and $q = \frac{dQ}{d\nu}$. The chain rule gives

$$\frac{dP}{dQ}\frac{dQ}{d\nu} = \frac{dP}{d\nu} \qquad \text{and} \qquad \frac{dP}{dQ} = \frac{\frac{dP}{d\nu}}{\frac{dQ}{d\nu}}\,.$$

Therefore

$$\mathrm{D}(P,Q) = \int p \log\left(\frac{p}{q}\right) d\nu\,.$$

For brevity, when writing integrals with respect to ν, in this proof, we will drop $d\nu$. Thus, we will write, for example $\int p \log(p/q)$ for the above integral. Instead of (14.5), we prove the stronger result that

$$\int p \wedge q \geq \frac{1}{2}\exp(-\mathrm{D}(P,Q))\,. \tag{14.6}$$

This indeed is sufficient since $\int p \wedge q = \int_A p \wedge q + \int_{A^c} p \wedge q \leq \int_A p + \int_{A^c} q = P(A) + Q(A^c)$. We start with an inequality attributed to French mathematician Lucien Le Cam, which lower bounds the left-hand side of Eq. (14.6). The inequality states that

$$\int p \wedge q \geq \frac{1}{2}\left(\int \sqrt{pq}\right)^2\,. \tag{14.7}$$

Starting from the right-hand side above using $pq = (p \wedge q)(p \vee q)$ and Cauchy-Schwartz we get

$$\left(\int \sqrt{pq}\right)^2 = \left(\int \sqrt{(p \wedge q)(p \vee q)}\right)^2 \leq \left(\int p \wedge q\right)\left(\int p \vee q\right)\,.$$

Now, using $p \wedge q + p \vee q = p + q$, the proof is finished by substituting $\int p \vee q = 2 - \int p \wedge q \leq 2$ and dividing both sides by two.

Thus, it remains to lower bound the right-hand side of (14.7). For this, we use Jensen's inequality. First, we write $(\cdot)^2$ as $\exp(2\log(\cdot))$ and then move the log

inside the integral:

$$\begin{aligned}\left(\int \sqrt{pq}\right)^2 &= \exp\left(2\log\int\sqrt{pq}\right) = \exp\left(2\log\int p\sqrt{\frac{q}{p}}\right)\\ &\geq \exp\left(2\int p\,\frac{1}{2}\,\log\left(\frac{q}{p}\right)\right) = \exp\left(-\int_{pq>0} p\log\left(\frac{p}{q}\right)\right)\\ &= \exp\left(-\int p\log\left(\frac{p}{q}\right)\right) = \exp\left(-\mathrm{D}(P,Q)\right)\,.\end{aligned}$$

In the fourth and the last step we used that since $P \ll Q$, $q = 0$ implies $p = 0$ and so $p > 0$, which implies $q > 0$, and eventually $pq > 0$. The result is completed by chaining the inequalities. □

14.2 Notes

1 Theorem 14.1 connects our definition of relative entropies to densities (the 'classic definition'). It can be found in Section 5.2 of the book by Gray [2011].

2 The supremum in the definition given in Eq. (14.4) may often be taken over a smaller set. Precisely, let $(\mathcal{X}, \mathcal{G})$ be a measurable space and suppose that $\mathcal{G} = \sigma(\mathcal{F})$ where $\mathcal{F}$ is a field. Note that a field is defined by the same axioms as a σ-algebra except that being closed under countable unions is replaced by the condition that it be closed under finite unions. Then for measures P and Q on $(\mathcal{X}, \mathcal{G})$ it holds that

$$\mathrm{D}(P,Q) = \sup_f \mathrm{D}(P_f, Q_f)\,,$$

where the supremum is over $\mathcal{F}/2^{[n]}$-measurable functions. This result is known as **Dobrushin's theorem**, which is due to [Dobrushin, 1959]. An alternative source is Lemma 5.2.2 in the book by Gray [2011].

3 How tight is Theorem 14.2? We remarked already that $\mathrm{D}(P,Q) = 0$ if and only if $P = Q$. But in this case Theorem 14.2 only gives

$$1 = P(A) + Q(A^c) \geq \frac{1}{2}\exp\left(-\mathrm{D}(P,Q)\right) = \frac{1}{2}\,,$$

which does not seem so strong. From where does the weakness arise? The answer is in Eq. (14.7), which can be refined by

$$\left(\int\sqrt{pq}\right)^2 \leq \left(\int p\wedge q\right)\left(\int p\vee q\right) = \left(\int p\wedge q\right)\left(2 - \int p\wedge q\right)$$

By solving the quadratic inequality we have

$$\begin{aligned}P(A) + Q(A^c) &\geq \int p\wedge q \geq 1 - \sqrt{1 - \left(\int\sqrt{pq}\right)^2}\\ &\geq 1 - \sqrt{1 - \exp\left(-\mathrm{D}(P,Q)\right)}\,,\end{aligned} \tag{14.8}$$

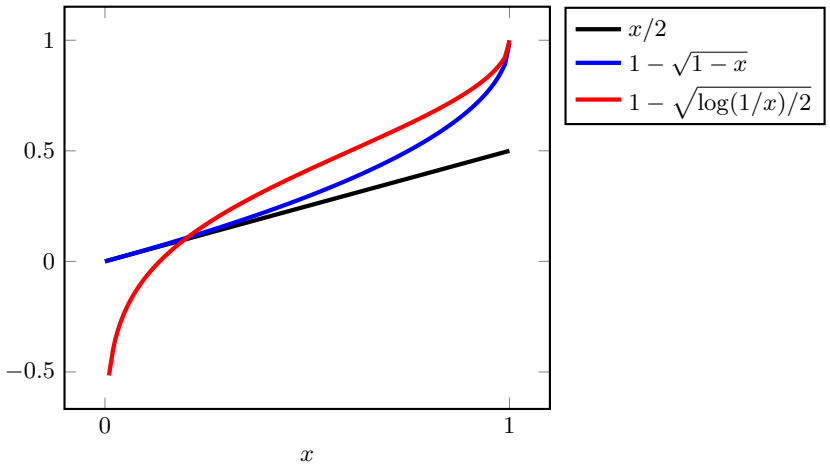

Figure 14.1 Tightening the inequality of Le Cam

which gives a modest improvement on Theorem 14.2 that becomes more pronounced when $\mathrm{D}(P, Q)$ is close to zero as demonstrated by Fig. 14.1. This stronger bound might be useful for fractionally improving constant factors in lower bounds, but we do not know of any application for which it is really crucial and the more complicated form makes it cumbersome to use. Part of the reason for this is that the situation where $\mathrm{D}(P, Q)$ is small is better dealt with using a different inequality as explained in the next note.

4 Another inequality from information theory is **Pinsker's inequality**, which states for measures P and Q on the same probability space $(\Omega, \mathcal{F})$ that

$$\delta(P, Q) = \sup_{A \in \mathcal{F}} P(A) - Q(A) \leq \sqrt{\frac{1}{2} \mathrm{D}(P, Q)}\,. \tag{14.9}$$

As an aside, the quantity on the left-hand side is call the **total variation distance** between P and Q, which actually is a distance on the space of probability measures (on the same probability space, of course). From this we can derive for any measurable $A \in \mathcal{F}$ that

$$P(A) + Q(A^c) \geq 1 - \sqrt{\frac{1}{2} \mathrm{D}(P, Q)} = 1 - \sqrt{\frac{1}{2} \log\left(\frac{1}{\exp(-\mathrm{D}(P, Q))}\right)}\,.$$

Examining Fig. 14.1 shows that this is an improvement on Eq. (14.8) when $\mathrm{D}(P, Q)$ is small.

5 We saw the total variation distance in Eq. (14.9). There are two other 'distances' that are occasionally useful. These are the **Hellinger distance** and the χ^2**-distance**, which using the notation in the proof of Theorem 14.2 are defined

by defined by

$$h(P,Q) = \sqrt{\int \left(\sqrt{p} - \sqrt{q}\right)^2} = \sqrt{2\left(1 - \int \sqrt{pq}\right)} \tag{14.10}$$

$$\chi^2(P,Q) = \int \frac{(p-q)^2}{q} = \int \frac{p^2}{q} - 1\,. \tag{14.11}$$

Notice that $h(P,Q)$ is bounded and exists for all probability measures P and Q, while a necessary condition for the χ^2-distance to exist is that $P \ll Q$. Like the total variation distance, the Hellinger distance is actually a distance (it is symmetric and satisfies triangle inequality), but the χ^2-'distance' is not. It is possible to show (see Chapter 2 of the book by Tsybakov [2008]) that

$$\delta(P,Q)^2 \le h(P,Q)^2 \le \mathrm{D}(P,Q) \le \chi^2(P,Q)\,. \tag{14.12}$$

Each of the inequalities are tight for some choices of P and Q, but the examples do not chain together as evidenced by Pinsker's inequality, which shows that $\delta(P,Q)^2 \le \mathrm{D}(P,Q)/2$ (which is also tight for some P and Q).

6 Let $P = (p_i)_i$ be a distribution on $[N]$. Another interpretation of the entropy $H(P)$ is as a measure of the amount of uncertainty in P. But what do we mean by uncertainty? One approach to define uncertainty is to think of how much one should be surprised to see a particular value of X (sampled from P). If x is deterministic, then there is no surprise at all and so the uncertainty measure should be zero. And indeed, $H(P) = 0$ when P is a Dirac measure. On the other hand, if X is uniformly distributed, then we should be equally surprised by any value, which provides some support defining the amount of 'surprise' when observing $X = i$ by $\log(1/p_i)$. Then entropy is the 'expected surprise'. Long story short, it turns out that reasonable definitions of uncertainty actually give rise to the definition of H in Eq. (14.2).

7 The entropy for distribution P was defined as $H(P)$ in Eq. (14.2). If X is a random variable, then $H(X)$ is defined to be the entropy of the law of X. This is a convenient notation because it allows one to write $H(f(X))$ and $H(XY)$ and similar expressions.

14.3 Bibliographic remarks

There are many references for information theory. Most well known (and comprehensive) is the book by Cover and Thomas [2012]. Another famous book is the elementary and enjoyable introduction by MacKay [2003]. The approach we have taken for defining and understanding the relative entropy is inspired by an excellent shorter book by Gray [2011].

14.4 Exercises

14.1 Let $(\Omega, \mathcal{F})$ be a measurable space and let $P, Q : \mathcal{F} \to [0,1]$ be probability measures. Let $a < b$ and $X : \Omega \to [a,b]$ be a $\mathcal{F}$-measurable random variable. Prove that

$$\left| \int_\Omega X(\omega) dP(\omega) - \int_\Omega X(\omega) dQ(\omega) \right| \leq (b-a)\delta(P,Q)\,.$$

14.2 Prove that each of the inequalities in Eq. (14.12) is tight.

14.3 Let Ω be a countable set and $p : \Omega \to [0,1]$ be a distribution on Ω so that $\sum_{\omega\in\Omega} p(\omega) = 1$. Let P be the measure associated with p, which means that $P(A) = \sum_{\omega\in A} p(\omega)$. The **counting measure** μ is the measure on $(\Omega, 2^\Omega)$ given by $\mu(A) = |A|$ if A is finite and $\mu(A) = \infty$ otherwise.

(a) Show that P is absolutely continuous with respect to μ.
(b) Show that the Radon-Nykodim $dP/d\mu$ exists and that $dP/d\mu(\omega) = p(\omega)$.

14.4 For each $i \in \{1,2\}$ let $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$ and $P_i = \mathcal{N}(\mu_i, \sigma_i^2)$. Show that

$$\mathrm{D}(P_1, P_2) = \frac{1}{2}\left(\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}\,.$$

14.5 Let λ be the Lebesgue measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. Find:

(a) a probability measure $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ that is not absolutely continuous with respect to λ.
(b) a probability measure P on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ that is absolutely continuous to λ with $\mathrm{D}(P,Q) = \infty$ where $Q = \mathcal{N}(0,1)$ is the standard Gaussian measure.

14.6 Let P and Q be measures on $(\Omega, \mathcal{F})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$ and $P_\mathcal{G}$ and $Q_\mathcal{G}$ be the restrictions of P and Q to $(\Omega, \mathcal{G})$. Show that $\mathrm{D}(P_\mathcal{G}, Q_\mathcal{G}) \leq \mathrm{D}(P,Q)$.

14.7 Let $(\Omega, \mathcal{F})$ be a measurable space and $P, Q : \mathfrak{B}(\mathbb{R}) \times \Omega \to [0,1]$ be a pair of probability kernels from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. Prove that

$$V = \{\omega \in \Omega : \mathrm{D}(P(\cdot \mid \omega), Q(\cdot \mid \omega)) = \infty\} \in \mathcal{F}\,.$$

Apply Dobrushin's theorem to the field of finite unions of rational-valued intervals in $\mathbb{R}$.

14.8 Let P and Q be measures on $(\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n))$ and for $t \in [n]$ let $X_t(x) = x_t$ be the coordinate project from $\mathbb{R}^n \to \mathbb{R}$. Then let P_t and Q_t be regular versions of X_t given $X_1, \ldots, X_{t-1}$ under P and Q respectively. Show that

$$\mathrm{D}(P,Q) = \sum_{t=1}^n \mathbb{E}_P\left[\mathrm{D}(P_t(\cdot \mid X_1, \ldots, X_{t-1}), Q_t(\cdot \mid X_1, \ldots, X_{t-1}))\right]\,. \tag{14.13}$$

This is a rather technical exercise. You will likely need to apply a monotone class argument [Kallenberg, 2002, Theorem 1.1]. For the definition of a regular version see [Kallenberg, 2002, Theorem 5.3] or Theorem 34.2 in Chapter 34. Briefly, P_t is a probability kernel from $(\mathbb{R}^{t-1}, \mathfrak{B}(\mathbb{R}^{t-1}))$ to $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ such that $P_t(A \mid x_1, \dots, x_{t-1}) = P(X_t \in A \mid X_1, \dots, X_{t-1})$ with P-probability one for all $A \in \mathfrak{B}(\mathbb{R})$.

14.9 Let P and Q be measures on $(\mathbb{R}^n, \mathfrak{B}(\mathbb{R}^n))$ and for $t \in [n]$ let $X_t(x) = x_t$ be the coordinate project from $\mathbb{R}^n \to \mathbb{R}$. Then let P_t and Q_t be regular versions of X_t given $X_1, \dots, X_{t-1}$ under P and Q respectively. Let τ be a stopping time adapted to the filtration generated by $X_1, \dots, X_n$ with $\tau \in [n]$ almost surely. Show that

$$\mathrm{D}(P_{|\mathcal{F}_\tau}, Q_{|\mathcal{F}_\tau}) = \mathbb{E}_P \left[\sum_{t=1}^{\tau} \mathrm{D}(P_t(\cdot \mid X_1, \dots, X_{t-1}), Q_t(\cdot \mid X_1, \dots, X_{t-1})) \right] .$$