2 Basic Information Theory

We begin with a short primer on basic notions of information theory. We review the various measures of information and their properties, without discussing their operational meanings.

At a high level, we can categorize information measures into three categories: measures of uncertainty, measures of mutual information, and measures of statistical distances. Quantities such as Shannon entropy fall in the first category, and provide a quantitative measure of uncertainty in a random variable. In addition to Shannon's classic notion of entropy, we also need a related notion of entropy due to Rényi. These quantities will be used throughout this book, but will be central to our treatment of randomness extraction (namely, the problem of generating random and independent coin tosses using a biased coin, correlated with another random variable).

The measures of mutual information capture the information revealed by a random variable about another random variable. As opposed to statistical notions such as mean squared error, which capture the notion of estimating or failing to estimate, mutual information allows us to quantify partial information and gives a mathematical equivalent of the heuristic phrase: "X gives *a bit of information* about Y." For us, Shannon's mutual information and a related quantity (using total variation distance) will be used to measure the "information leaked" to an adversary. These quantities are central to the theme of this book.

Finally, we need the notions of distances between probability distributions. Statistical inference entails determining the properties of the distribution generating a random variable X, by observing X. The closer two distributions P and Q are, the more difficult it is to distinguish whether X is generated by P or by Q. Information theory provides a battery of measures for "distances" between two probability distributions. In fact, our notions of information-theoretic security will be defined using these distances.

In addition, we present the properties of these quantities and inequalities relating them. To keep things interesting, we have tried to present "less well-known" proofs of these inequalities – even a reader familiar with these bounds may find something interesting in this chapter. For instance, we prove Fano's inequality using data processing inequality, we prove continuity of entropy using Fano's

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

inequality, we provide multiple variational formulae for information quantities, and we provide two different proofs of Pinsker's inequality.

We start by covering the essential notions of probability.

2.1 Very Basic Probability

Since our focus will be on discrete random variables, the reader only needs to be familiar with very basic probability theory. We review the main concepts and notations in this section.

Let Ω be a set of finite cardinality. A discrete *probability distribution* P on Ω can be described using its *probability mass function* (pmf) $p\colon \Omega \to [0,1]$ satisfying $\sum_{\omega\in\Omega} p(\omega) = 1$. We can think of Ω as the underlying "probability space" with "probability measure" P. Since we restrict our attention to discrete random variables throughout this book, we will specify the probability distribution P using the corresponding pmf. In particular, with an abuse of notation, we use $\mathrm{P}(\omega)$ to denote the probability of the event $\{\omega\}$ under P.

For a probability distribution P on Ω, we denote by $\mathtt{supp}(\mathrm{P})$ its *support set* given by

$$\mathtt{supp}(\mathrm{P}) = \{\omega \in \Omega : \mathrm{P}(\omega) > 0\}.$$

A *discrete random variable* X is a mapping $X\colon \Omega \to \mathcal{X}$ where $\mathcal{X}$ is a set of finite cardinality. Without loss of generality, we will assume that the mapping X is onto, namely $\mathcal{X} = \mathtt{range}(X)$. It is not necessary that the probability space (Ω, P) is discrete; but it suffices for our purpose. We can associate with a random variable X a distribution P_X, which is the distribution "induced" on the output $\mathcal{X}$ by X. Since X is a discrete random variable, P_X is a discrete probability distribution given by

$$\mathrm{P}_X(x) = \Pr(X = x) = \sum_{\omega\in\Omega: X(\omega)=x} \mathrm{P}(\omega);$$

often, we will simply say $\mathrm{P} = \mathrm{P}_X$ is a probability distribution on $\mathcal{X}$, without referring to the underlying probability space Ω. Throughout the book, unless otherwise stated, when we say random variable, we refer only to discrete random variables.

Notation for random variables: In probability, we often choose the set $\mathcal{X}$ as $\mathbb{R}$, the set of real numbers. But in this book we will often treat random binary vectors or messages as random variables. Thus, we will allow for an arbitrary finite set $\mathcal{X}$. We will say that "X is a random variable taking values in the set $\mathcal{X}$," for any finite set $\mathcal{X}$. We denote the probability distribution of X by P_X and, for every realization $x \in \mathcal{X}$, denote by $\mathrm{P}_X(x)$ the probability $\Pr(X = x)$. Throughout the book, we denote the random variables by capital letters such as X, Y, Z, etc., realizations by the corresponding small letters such as x, y, z, etc., and the corresponding range-sets by calligraphic letters such as $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, etc.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

Associated with a random variable X taking values in $\mathcal{X} \subset \mathbb{R}$, there are two fundamental quantities: its *expected value* $\mathbb{E}[X]$ given by

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \mathrm{P}_X(x)\,,$$

and its *variance* $\mathbb{V}[X]$ given by

$$\mathbb{V}[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]\,.$$

An important property of expectation is its *linearity*. Namely,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]\,.$$

Before proceeding, denoting by $\mathbf{1}[\mathcal{S}]$ the *indicator function* for the set $\mathcal{S}$, we note a useful fact about the binary random variable $\mathbf{1}[X \in \mathcal{A}]$:

$$\mathbb{E}[\mathbf{1}[X \in \mathcal{A}]] = \Pr(X \in \mathcal{A})\,,$$

for any subset $\mathcal{A}$ of $\mathcal{X}$. Heuristically, the expected value of X serves as an estimate for X and the variance of X serves as an estimate for the error in the estimate. We now provide two simple inequalities that formalize this heuristic.

THEOREM 2.1 (Markov's inequality) *For a random variable X taking values in $[0, \infty)$ and any $t > 0$, we have*

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof We note that any random variable X can be expressed as

$$X = X\mathbf{1}[X \geq t] + X\mathbf{1}[X < t]\,.$$

Then,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X\mathbf{1}[X \geq t]] + \mathbb{E}[X\mathbf{1}[X < t]] \\ &\geq \mathbb{E}[X\mathbf{1}[X \geq t]] \\ &\geq \mathbb{E}[t\mathbf{1}[X \geq t]] \\ &= t\Pr(X \geq t)\,,\end{aligned}$$

where the first inequality uses the fact that X is nonnegative and the second inequality uses the fact that $X\mathbf{1}[X \geq t]$ is either 0 or exceeds t with probability 1. □

Applying Markov's inequality to $|X - \mathbb{E}[X]\,|^2$, we obtain the following bound.

THEOREM 2.2 (Chebyshev's inequality) *For a random variable X taking values in $\mathbb{R}$ and any $t > 0$, we have*

$$\Pr(|X - \mathbb{E}[X]\,| \geq t) \leq \frac{\mathbb{V}[X]}{t^2}.$$

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

The previous bound says that the random variable X lies within the interval $[\mathbb{E}[X] - \sqrt{\mathbb{V}[X]/\varepsilon}, \mathbb{E}[X] + \sqrt{\mathbb{V}[X]/\varepsilon}]$ with probability exceeding $1-\varepsilon$. Such an interval is called the $(1-\varepsilon)$-confidence interval. In fact, the $\sqrt{1/\varepsilon}$ dependence of the accuracy on the probability of error can often be improved to $\sqrt{\ln 1/\varepsilon}$ using a "Chernoff bound." We will present such bounds later in the book.

When we want to consider multiple random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, we consider their joint distribution P_{XY} specified by the joint probability distribution $\mathrm{P}_{XY}(x,y)$. For discrete random variables, we can define the *conditional distribution* $\mathrm{P}_{X|Y}$ using the conditional probabilities given by[1]

$$\mathrm{P}_{X|Y}(x|y) := \frac{\mathrm{P}_{XY}(x,y)}{\mathrm{P}_Y(y)}$$

for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ such that $\mathrm{P}_Y(y) > 0$; if $\mathrm{P}_Y(y) = 0$, we can define $\mathrm{P}_{X|Y}(x|y)$ to be an arbitrary distribution on $\mathcal{X}$. A particular quantity of interest is the *conditional expectation* $\mathbb{E}[X|Y]$ of X given Y, which is a random variable that is a function of Y, defined as

$$\mathbb{E}[X|Y](y) := \sum_{x \in \mathcal{X}} x \mathrm{P}_{X|Y}(x|y), \quad \forall\, y \in \mathcal{Y}.$$

Often, we use the alternative notation $\mathbb{E}[X|Y=y]$ for $\mathbb{E}[X|Y](y)$. Note that $\mathbb{E}[X|Y]$ denotes the random variable obtained when y is replaced with the random Y with distribution P_Y.

In information theory, it is customary to use the terminology of a *channel* in place of conditional distributions. Simply speaking, a channel is a randomized mapping. For our use, we will define this mapping using the conditional distribution it induces between the output and the input. Formally, we have the definition below.

DEFINITION 2.3 (Channels) For finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, a channel W with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is given by probabilities $W(y|x)$, $y \in \mathcal{Y}, x \in \mathcal{X}$, where $W(y|x)$ denotes the probability of observing y when the input is x.

Often, we abuse the notation and use $W\colon \mathcal{X} \to \mathcal{Y}$ to represent the channel. Also, for a channel $W\colon \mathcal{X} \to \mathcal{Y}$ and input distribution P on $\mathcal{X}$, we denote by $\mathrm{P} \circ W$ the distribution induced on the output Y of the channel when the input X has distribution X. Namely,

$$(\mathrm{P} \circ W)(y) = \sum_{x \in \mathcal{X}} \mathrm{P}(x) W(y|x).$$

Furthermore, we denote by $\mathrm{P} \times W$ the joint distribution of (X,Y).

[1] In probability theory, conditional probability densities are technically difficult to define and require several conditions on the underlying probability space. However, since we restrict our attention to discrete random variables, the conditional pmf serves this purpose for us.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

2.2 The Law of Large Numbers

Chebyshev's inequality tells us that a good estimate for a random variable X is its expected value $\mathbb{E}[X]$, up to an accuracy of roughly $\pm\sqrt{\mathbb{V}[X]}$. In fact, this estimate is pretty sharp asymptotically when X is a sum of independent random variables.

Specifically, let $X_1, \ldots, X_n$ be *independent and identically distributed* (i.i.d.) random variables, that is, they are independent and have the same (marginal) distribution, say, P_X. An important fact to note about independent random variables is that their variance is *additive*. Indeed, for independent random variables X and Y,

$$\begin{aligned}\mathbb{V}[X+Y] &= \mathbb{E}\big[(X+Y)^2\big] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}\big[X^2\big] + \mathbb{E}\big[Y^2\big] + 2\mathbb{E}[XY] - \left(\mathbb{E}[X]^2 + \mathbb{E}[Y]^2 + 2\mathbb{E}[X]\,\mathbb{E}[Y]\right) \\ &= \mathbb{V}[X] + \mathbb{V}[Y] + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\right) \\ &= \mathbb{V}[X] + \mathbb{V}[Y]\,,\end{aligned}$$

where in the final identity we used the observation that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ for independent random variables.

In fact, X and Y such that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ are called *uncorrelated*, and this is the only property we need to get additivity of variance. Note that in general, random variables may be uncorrelated but not independent.[2]

Returning to i.i.d. random variables $X_1, X_2, \ldots, X_n$ with common distribution P_X, by the linearity of expectation we have

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\mathbb{E}[X]\,,$$

and since the variance is additive for independent random variables,

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i] = n\mathbb{V}[X]\,.$$

Therefore, by Chebyshev's inequality,

$$\Pr\left(\left|\sum_{i=1}^{n} X_i - n\mathbb{E}[X]\right| \geq t\right) \leq \frac{n\mathbb{V}[X]}{t^2},$$

or equivalently,

$$\Pr\left(\left|\sum_{i=1}^{n} X_i - n\mathbb{E}[X]\right| \geq \sqrt{\frac{n\mathbb{V}[X]}{\varepsilon}}\right) \leq \varepsilon$$

[2] For example, consider the random variable X taking values $\{-1, 0, 1\}$ with equal probabilities and $Y = 1 - |X|$. For these random variables, $\mathbb{E}[X] = 0 = \mathbb{E}[X]\,\mathbb{E}[Y]$ and $\mathbb{E}[XY] = 0$ since $Y = 0$ whenever $X \neq 0$. But clearly X and Y are not independent.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

for every $\varepsilon \in (0,1)$. Thus, with large probability, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is roughly within $\pm\sqrt{\mathbb{V}[X]/n}$ of its expected value $\mathbb{E}[X]$. We have proved the *weak law of large numbers*.

THEOREM 2.4 (Weak law of large numbers) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with common distribution P_X over a finite set $\mathcal{X} \subset \mathbb{R}$. For every $\delta > 0$ and $\varepsilon \in (0,1)$, we have*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| > \delta\right) \leq \varepsilon$$

for every n sufficiently large.

Alternatively, we can express the previous result as follows: for every $\delta > 0$,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| > \delta\right) = 0.$$

In fact, a stronger version of the result above holds – we can exchange the limit and the probability. This result is a bit technical, and it says that for all "sample paths" $\{X_i\}_{i=1}^{\infty}$ the sample average $\frac{1}{n}\sum_{i=1}^{n}$ converges to $\mathbb{E}[X]$ as n goes to infinity. We state this result without proof.

THEOREM 2.5 (Strong law of large numbers) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with common distribution P_X over a finite set $\mathcal{X} \subset \mathbb{R}$. Then,*

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X]\right) = 1.$$

In summary, in this section we have learnt that the average of a large number of independent random variables can be approximated, rather accurately, by its expected value.

2.3 Convex and Concave Functions

Convex and concave functions play an important role in information theory. Informally speaking, convex and concave functions, respectively, are those whose graphs look like a "cup" and "cap." In particular, a function f is concave if $-f$ is convex. We provide the formal definition and a key property below.

The domain of a convex function must be a *convex set*, which we define first. For simplicity, we restrict ourselves to convex sets that are subsets of $\mathbb{R}^d$.

DEFINITION 2.6 (Convex set) For a natural number $d \in \mathbb{N}$, a set $\mathcal{S} \subset \mathbb{R}^d$ is a convex set if for any two points s_1 and s_2 in $\mathcal{S}$ and any $\theta \in [0,1]$, we must have $\theta s_1 + (1-\theta)s_2 \in \mathcal{S}$. Namely, if two points belong to $\mathcal{S}$, then all the points in the straight line joining these two points also belong to $\mathcal{S}$.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

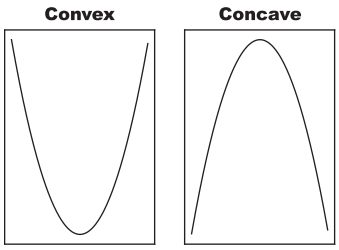

Figure 2.1 Depiction of convex and concave functions.

DEFINITION 2.7 (Convex and concave functions) For $d \in \mathbb{N}$, let $\mathcal{S} \subset \mathbb{R}^d$ be a convex set. Then, a function $f\colon \mathcal{S} \to \mathbb{R}$ is a convex function if for every $\theta \in [0,1]$ and every pair of points $s_1, s_2 \in \mathcal{S}$, we have

$$f(\theta s_1 + (1-\theta)s_2) \leq \theta f(s_1) + (1-\theta) f(s_2), \tag{2.1}$$

and it is concave if

$$f(\theta s_1 + (1-\theta)s_2) \geq \theta f(s_1) + (1-\theta) f(s_2). \tag{2.2}$$

In particular, when strict inequality holds in (2.1) (respectively (2.2)) for every $\theta \in (0,1)$ and $s_1 \neq s_2$, then the function is a strict convex function (respectively strict concave function).

Simply speaking, a function is convex (respectively concave) if the value of the function at the (weighted) average of two points is less than (respectively more than) the average of the values at the point. Note that linear functions are both convex and concave, but not strict convex nor strict concave. Examples of strict convex functions include e^x, e^{-x}, x^2, etc. and examples of strict concave functions include $\ln x$ (natural logarithm), $\sqrt{x}$, etc. We depict the shape of convex and concave functions in Figure 2.1.

It is clear from the definition of convex functions that, for a random variable X taking values in a finite set $\mathcal{X}$ and a convex function f, we must have $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$. This is the powerful Jensen inequality.

LEMMA 2.8 (Jensen's inequality) *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function and X a random variable taking values in $\mathbb{R}^n$. Then,*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]\,. \tag{2.3}$$

Similarly, if f is a concave function, then

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]\,. \tag{2.4}$$

In particular, when f is strict convex (respectively strict concave), then strict inequality holds in (2.3) (respectively (2.4)) unless $X = \mathbb{E}[X]$ with probability 1.

In fact, this inequality holds for more general random variables than those considered in this book – there is no need for the assumption of finiteness or

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

even discreteness. However, the proof is technical and beyond the scope of this book.

Note that for the convex function $f(x) = x^2$, this inequality implies $\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \geq 0$. Another very useful implication of this inequality is for the concave function $f(x) = \log x$. Instead of showing that $\log x$ is concave and using Jensen's inequality, we derive a self-contained inequality which is very handy.

LEMMA 2.9 (Log-sum inequality) *For nonnegative numbers* $\{(a_i, b_i)\}_{i=1}^n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i},$$

with equality[3] *if and only if* $a_i = b_i$ *for all* i.

Proof The inequality is trivial if all a_i are 0, or if there exists an i such that $b_i = 0$ but $a_i \neq 0$. Otherwise, by rearranging the terms, it suffices to show that

$$\sum_{i=1}^{n} a_i \log \frac{a_i / \sum_{j=1}^n a_j}{b_i / \sum_{j=1}^n b_j} \geq 0,$$

which holds if and only if

$$\sum_{i=1}^{n} a_i' \log \frac{a_i'}{b_i'} \geq 0,$$

where $a_i' = a_i / \sum_{j=1}^n a_j$ and $b_i' = b_i / \sum_{j=1}^n b_j$. Note that the previous inequality is simply our original inequality for the case when $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i = 1$. Thus, without loss of generality we can assume that a_i and b_i, $1 \leq i \leq n$, both constitute pmfs. Then, since $\ln x \leq x - 1$, $\log x \leq (x-1)\log e$, applying this inequality for $x = a_i/b_i$ we get

$$\sum_{i=1}^{n} a_i \log \frac{b_i}{a_i} \leq \log e \sum_{i=1}^{n} a_i \left(\frac{b_i}{a_i} - 1 \right) = 0,$$

which establishes the inequality.

Equality can hold only if equality holds for every instance of $\ln x \leq x - 1$ used in the proof, which happens only if $x = 1$. Thus, equality holds only if $a_i = b_i$ for every $i \in \{1, \ldots, n\}$. □

2.4 Total Variation Distance

Suppose we observe a sample X taking values in a discrete set $\mathcal{X}$. How difficult is it to determine if X is generated from a distribution P or Q? We need a precise quantitative handle on this difficulty for this book. Indeed, the formal notion of information-theoretic security that we define in this book will rely on difficulty in solving such problems. This problem is one of the fundamental problems in

[3] We follow the convention that $0 \log(0/0) = 0 \log 0 = 0$ throughout the book.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

statistics – the *binary hypothesis testing* problem. Later in the book, we will revisit this problem in greater detail. In this section, we only present a quick version that will help us motivate an important notion.

Formally, we can apply a channel $T\colon \mathcal{X} \to \{0,1\}$ to make this decision, where the outputs 0 and 1 indicate P and Q, respectively. Such a decision rule is called a (hypothesis) *test*. For this channel denoting the test, let $T(1|x) = 1 - T(0|x)$ denote the probability with which T declares 1 for input $x \in \mathcal{X}$. When X was generated using P, the test T makes an error if its output is 1, which happens with probability $\sum_{x\in\mathcal{X}} \mathrm{P}(x)T(1|x)$. Similarly, when X was generated using Q, the probability of making an error is $\sum_{x\in\mathcal{X}} \mathrm{Q}(x)T(0|x)$.

One simple notion of performance of the test T is the average of these two errors. It corresponds to the probability of error in T's output when the input distribution for X is chosen to be P or Q with equal probability. Specifically, the average probability of error $P_{\mathtt{err}}(T)$ is given by

$$P_{\mathtt{err}}(T|\mathrm{P},\mathrm{Q}) := \frac{1}{2}\left[\sum_{x\in\mathcal{X}} \mathrm{P}(x)T(1|x) + \sum_{x\in\mathcal{X}} \mathrm{Q}(x)T(0|x)\right],$$

and a measure of difficulty of the hypothesis testing problem described above is the minimum average probability of error

$$P_{\mathtt{err}}(\mathrm{P},\mathrm{Q}) = \min_{T} P_{\mathtt{err}}(T|\mathrm{P},\mathrm{Q}),$$

where the minimum is over all channels $T\colon \mathcal{X} \to \{0,1\}$.

In fact, the minimizing T is easy to find (it is sometimes called the *Bayes optimal test* or simply the *Bayes test*).

LEMMA 2.10 (Bayes test) *For distributions* P *and* Q *on a discrete set* $\mathcal{X}$, $P_{\mathtt{err}}(\mathrm{P},\mathrm{Q})$ *is attained by the deterministic test which outputs* 0 *for* $x \in \mathcal{X}$ *such that* $\mathrm{P}(x) \geq \mathrm{Q}(x)$ *and* 1 *otherwise, i.e.,*

$$T^*(0|x) = \begin{cases} 1, & \text{if } \mathrm{P}(x) \geq \mathrm{Q}(x), \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, the minimum average probability of error is given by

$$P_{\mathtt{err}}(\mathrm{P},\mathrm{Q}) = \frac{1}{2}\left[1 - \sum_{x\in\mathcal{X}:\mathrm{P}(x)\geq\mathrm{Q}(x)} (\mathrm{P}(x) - \mathrm{Q}(x))\right].$$

Proof It is easy to verify that

$$P_{\mathtt{err}}(T^*|\mathrm{P},\mathrm{Q}) = \frac{1}{2}\left[1 - \sum_{x\in\mathcal{X}:\mathrm{P}(x)\geq\mathrm{Q}(x)} (\mathrm{P}(x) - \mathrm{Q}(x))\right].$$

Thus, it suffices to show that the right-hand side of the expression above is less than $P_{\mathtt{err}}(T|\mathrm{P},\mathrm{Q})$ for every test T. To this end, note that

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

$$\begin{aligned}
2P_{\mathtt{err}}(T|\mathrm{P},\mathrm{Q}) &= \sum_{x\in\mathcal{X}} \mathrm{P}(x)T(1|x) + \sum_{x\in\mathcal{X}} \mathrm{Q}(x)T(0|x) \\
&= \sum_{x\in\mathcal{X}} \mathrm{P}(x)T(1|x) + \sum_{x\in\mathcal{X}} \mathrm{Q}(x)(1-T(1|x)) \\
&= \sum_{x\in\mathcal{X}} (\mathrm{P}(x)-\mathrm{Q}(x))T(1|x) + \sum_{x\in\mathcal{X}} \mathrm{Q}(x) \\
&\geq \sum_{x\in\mathcal{X}:\mathrm{P}(x)<\mathrm{Q}(x)} (\mathrm{P}(x)-\mathrm{Q}(x)) + 1 \\
&= \sum_{x\in\mathcal{X}} (\mathrm{P}(x)-\mathrm{Q}(x))T^*(1|x) + 1 \\
&= 1 - \sum_{x\in\mathcal{X}:\mathrm{P}(x)\geq\mathrm{Q}(x)} (\mathrm{P}(x)-\mathrm{Q}(x)),
\end{aligned}$$

where the inequality holds since $T(1|x)$ lies in $[0,1]$ and we have dropped only positive terms from the preceding expression. □

The optimal test T^* that emerges from the previous result is a natural one: declare P when you observe x which has higher probability of occurrence under P than under Q. Thus, the difficulty of resolving the hypothesis testing problem above is determined by the quantity $\sum_{x\in\mathcal{X}:\mathrm{P}(x)\geq\mathrm{Q}(x)} \mathrm{P}(x)-\mathrm{Q}(x)$.

We note that the optimal test T^* is a deterministic one. Further, note that a deterministic test can be characterized by the subset $\mathcal{A}=\{x\in\mathcal{X} : T(0|x)=1\}$ of $\mathcal{X}$, and its probability of error is given by

$$P_{\mathtt{err}}(T|\mathrm{P},\mathrm{Q}) = \frac{1}{2}\left[\mathrm{P}(\mathcal{A}^c)+\mathrm{Q}(\mathcal{A})\right] = \frac{1}{2}(1-(\mathrm{P}(\mathcal{A})-\mathrm{Q}(\mathcal{A}))).$$

By comparing with the optimal probability of error for a deterministic test (attained by T^*), we get

$$\sum_{x\in\mathcal{X}:\mathrm{P}(x)>\mathrm{Q}(x)} \mathrm{P}(x)-\mathrm{Q}(x) = \max_{\mathcal{A}\subset\mathcal{X}} \mathrm{P}(\mathcal{A})-\mathrm{Q}(\mathcal{A}),$$

where the maximum is attained by the set $\mathcal{A}^*=\{x\in\mathcal{X} : \mathrm{P}(x)>\mathrm{Q}(x)\}$ (corresponding to T^*).[4]

We would like to treat $\max_{\mathcal{A}\subset\mathcal{X}} \mathrm{P}(\mathcal{A})-\mathrm{Q}(\mathcal{A})$ as a notion of "distance" between the distributions P and Q; our little result above already tells us that the closer P and Q are in this distance, the harder it is to tell them apart using statistical tests. In fact, this quantity is rather well suited to being termed a distance. The next result shows an equivalent form for the same quantity that better justifies its role as a distance.

LEMMA 2.11 *For two distributions* P *and* Q *on a finite set* $\mathcal{X}$*, we have*

$$\max_{A\subset\mathcal{X}} \mathrm{P}(\mathcal{A})-\mathrm{Q}(\mathcal{A}) = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mathrm{P}(x)-\mathrm{Q}(x)|.$$

We leave the proof as an (interesting) exercise (see Problem 2.1).

[4] Alternatively, we can include x with $\mathrm{P}(x)=\mathrm{Q}(x)$ into $\mathcal{A}^*$.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

The expression on the right-hand side of the previous result is (up to normalization) the ℓ_1 distance between finite-dimensional vectors $(\mathrm{P}(x), x \in \mathcal{X})$ and $(\mathrm{Q}(x), x \in \mathcal{X})$. We have obtained the following important notion of distance between P and Q.

DEFINITION 2.12 (Total variation distance) For discrete distributions P and Q on $\mathcal{X}$, the *total variation distance* $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})$ between P and Q is given by[5]

$$d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) := \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathrm{P}(x) - \mathrm{Q}(x)| = \max_{A \subset \mathcal{X}} \mathrm{P}(\mathcal{A}) - \mathrm{Q}(\mathcal{A}).$$

By the foregoing discussion and the definition above, we have

$$P_{\mathtt{err}}(\mathrm{P}, \mathrm{Q}) = \frac{1}{2} \left(1 - d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})\right). \tag{2.5}$$

THEOREM 2.13 (Properties of total variation distance) *For distributions* P, Q, *and* R *on a finite set* $\mathcal{X}$, *the following hold.*

1. (Nonnegativity) $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) \geq 0$ *with equality if and only if* $\mathrm{P} = \mathrm{Q}$.
2. (Triangular inequality) $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) \leq d_{\mathtt{var}}(\mathrm{P}, \mathrm{R}) + d_{\mathtt{var}}(\mathrm{R}, \mathrm{Q})$.
3. (Normalization) $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) \leq 1$ *with equality if and only if* $\mathtt{supp}(\mathrm{P}) \cap \mathtt{supp}(\mathrm{Q}) = \emptyset$.

Proof The first two properties are easy to check; we only show the third one. For that, we note that $\mathrm{P}(\mathcal{A}) - \mathrm{Q}(\mathcal{A}) \leq 1$ for every $\mathcal{A} \subset \mathcal{X}$, whereby $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) \leq 1$. Further, denoting $\mathcal{A}^* = \{x : \mathrm{P}(x) > \mathrm{Q}(x)\}$, we have $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) = \mathrm{P}(\mathcal{A}^*) - \mathrm{Q}(\mathcal{A}^*)$ whereby $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) = 1$ holds if and only if

$$\mathrm{P}(\mathcal{A}^*) - \mathrm{Q}(\mathcal{A}^*) = 1.$$

This in turn is possible if and only if $\mathrm{P}(\mathcal{A}^*) = 1$ and $\mathrm{Q}(\mathcal{A}^*) = 0$, which is the same as $\mathcal{A}^* = \mathtt{supp}(\mathrm{P})$ and $\mathtt{supp}(\mathrm{Q}) \subset \mathcal{A}^c$, completing the proof. □

Next, we note a property which must be satisfied by any reasonable measure of distance between distributions – the *data processing inequality.*

LEMMA 2.14 (Data processing inequality for total variation distance) *Let* P_X *and* Q_X *be distributions on* $\mathcal{X}$ *and* $T \colon \mathcal{X} \to \mathcal{Y}$ *be a channel. Denote by* P_Y *and* Q_Y *the distribution of the output of* T *when the input distribution is* P_X *and* P_Y, *respectively. Then,*

$$d_{\mathtt{var}}(\mathrm{P}_Y, \mathrm{Q}_Y) \leq d_{\mathtt{var}}(\mathrm{P}_X, \mathrm{Q}_X).$$

Proof Define the conditional distribution $W(y|x) := \Pr(T(X) = y \mid X = x)$, $x \in \mathcal{X}, y \in \mathcal{Y}$. Then, the distributions P_Y and Q_Y are given by

$$\mathrm{P}_Y(y) = \sum_{x \in \mathcal{X}} \mathrm{P}_X(x)\, W(y|x), \quad \mathrm{Q}_Y(y) = \sum_{x \in \mathcal{X}} \mathrm{Q}_X(x) W(y|x).$$

[5] Other names used for the distance are *variational distance* and *statistical distance.*

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

Thus, we get

$$\begin{aligned} d_{\mathtt{var}}(\mathrm{P}_Y, \mathrm{Q}_Y) &= \frac{1}{2} \sum_{y \in \mathcal{Y}} |\mathrm{P}_Y(y) - \mathrm{Q}_Y(y)| \\ &= \frac{1}{2} \sum_{y \in \mathcal{Y}} |\sum_{x \in \mathcal{X}} W(y|x)(\mathrm{P}_X(x) - \mathrm{Q}_X(x))| \\ &\leq \frac{1}{2} \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} W(y|x)|\mathrm{P}_X(x) - \mathrm{Q}_X(x)| \\ &= d_{\mathtt{var}}(\mathrm{P}_X, \mathrm{Q}_X), \end{aligned}$$

which completes the proof. □

We can see the property above directly from the connection between hypothesis testing and total variation distance seen earlier. Indeed, we note that $P_{\mathtt{err}}(\mathrm{P}_Y, \mathrm{Q}_Y) \geq P_{\mathtt{err}}(\mathrm{P}_X, \mathrm{Q}_X)$ since the optimal test for P_Y versus Q_Y can be used as a test for P_X versus Q_X as well, by first transforming the observation X to $Y = T(X)$. By (2.5), this yields the data processing inequality above.

We close this section with a very useful property, which will be used heavily in formal security analysis of different protocols later in the book.

LEMMA 2.15 (Chain rule for total variation distance) *For two distributions* P_{XY} *and* Q_{XY} *on* $\mathcal{X} \times \mathcal{Y}$*, we have*

$$d_{\mathtt{var}}(\mathrm{P}_{XY}, \mathrm{Q}_{XY}) \leq d_{\mathtt{var}}(\mathrm{P}_X, \mathrm{Q}_X) + \mathbb{E}_{\mathrm{P}_X}\left[d_{\mathtt{var}}(\mathrm{P}_{Y|X}, \mathrm{Q}_{Y|X})\right],$$

and further,

$$d_{\mathtt{var}}(\mathrm{P}_{X_1 \cdots X_n}, \mathrm{Q}_{X_1 \cdots X_n}) \leq \sum_{i=1}^{n} \mathbb{E}_{\mathrm{P}_{X^{i-1}}}\left[d_{\mathtt{var}}(\mathrm{P}_{X_i|X^{i-1}}, \mathrm{Q}_{X_i|X^{i-1}})\right],$$

where X^i *abbreviates the random variable* $(X_1, \ldots, X_i)$*.*

Proof The proof simply uses the triangular inequality for total variation distance. Specifically, we have

$$d_{\mathtt{var}}(\mathrm{P}_{XY}, \mathrm{Q}_{XY}) \leq d_{\mathtt{var}}(\mathrm{P}_{XY}, \mathrm{P}_X \mathrm{Q}_{Y|X}) + d_{\mathtt{var}}(\mathrm{Q}_{XY}, \mathrm{P}_X \mathrm{Q}_{Y|X}).$$

For the first term on the right-hand side, we have

$$\begin{aligned} d_{\mathtt{var}}(\mathrm{P}_X \mathrm{P}_{Y|X}, \mathrm{P}_X \mathrm{Q}_{Y|X}) &= \frac{1}{2} \sum_{x \in \mathcal{X}} \mathrm{P}_X(x) \sum_{y \in \mathcal{Y}} |\mathrm{P}_{Y|X}(y|x) - \mathrm{Q}_{Y|X}(y|x)| \\ &= \mathbb{E}_{\mathrm{P}_X}\left[d_{\mathtt{var}}(\mathrm{Q}_{Y|X}, \mathrm{P}_{Y|X})\right], \end{aligned}$$

and further, for the second term,

$$\begin{aligned} d_{\mathtt{var}}(\mathrm{Q}_X \mathrm{Q}_{Y|X}, \mathrm{P}_X \mathrm{Q}_{Y|X}) &= \frac{1}{2} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} |\mathrm{Q}_X(x)\mathrm{Q}_{Y|X}(y|x) - \mathrm{P}_X(x)\,\mathrm{Q}_{Y|X}(y|x)| \\ &= \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{Q}_{Y|X}(y|x)|\mathrm{P}_X(x) - \mathrm{Q}_X(x)| \\ &= d_{\mathtt{var}}(\mathrm{P}_X, \mathrm{Q}_X). \end{aligned}$$

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

The claim follows upon combining these expressions with the previous bound; the general proof for $n \geq 2$ follows by applying this inequality repeatedly. □

In the proof above, we noted that

$$d_{\mathtt{var}}(\mathrm{P}_{XY}, \mathrm{P}_X \mathrm{Q}_{Y|X}) = \mathbb{E}_{\mathrm{P}_X}\left[d_{\mathtt{var}}(\mathrm{P}_{Y|X}, \mathrm{Q}_{Y|X})\right],$$

a useful expression of independent interest. Further, for product distributions P_{X^n} and Q_{X^n} (corresponding to independent random variables), the previous result implies the *subadditivity property*

$$d_{\mathtt{var}}(\mathrm{P}_{X^n}, \mathrm{Q}_{X^n}) \leq \sum_{i=1}^{n} d_{\mathtt{var}}(\mathrm{P}_{X_i}, \mathrm{Q}_{X_i}).$$

2.5 Kullback–Leibler Divergence

In this book, we use total variation distance to define our notion of security. However, there is a close cousin of this notion of distance that enjoys great popularity in information theory – the *Kullback–Leibler divergence*.

DEFINITION 2.16 (KL divergence) For two discrete distributions P and Q on $\mathcal{X}$, the Kullback–Leibler (KL) divergence $D(\mathrm{P}\|\mathrm{Q})$ between P and Q is given by

$$D(\mathrm{P}\|\mathrm{Q}) = \begin{cases} \sum_{x \in \mathcal{X}} \mathrm{P}(x) \log \frac{\mathrm{P}(x)}{\mathrm{Q}(x)}, & \mathtt{supp}(\mathrm{P}) \subset \mathtt{supp}(\mathrm{Q}), \\ \infty, & \mathtt{supp}(\mathrm{P}) \not\subset \mathtt{supp}(\mathrm{Q}) \end{cases}$$

where log is to the base 2. This convention will be followed throughout – *all our logarithms are to the base* 2*, unless otherwise stated.*

KL divergence is not a metric, but is a very useful notion of "distance" between distributions. Without giving it an "operational meaning" at the outset, we simply note some of its useful properties in this chapter.

THEOREM 2.17 (Properties of KL divergence) *For distributions P and Q on a finite set $\mathcal{X}$, the following hold.*

1. (Nonnegativity) $D(\mathrm{P}\|\mathrm{Q}) \geq 0$ *with equality if and only if* $\mathrm{P} = \mathrm{Q}$.
2. (Convexity) $D(\mathrm{P}\|\mathrm{Q})$ *is a convex function of the pair* (P, Q) *(over the set of pairs of distributions on* $\mathcal{X}$*).*
3. (Data processing inequality) *For a channel* $T\colon \mathcal{X} \to \mathcal{Y}$*, denote by* P_Y *and* Q_Y *the distribution of the output of* T *when the input distribution is* P_X *and* Q_X*, respectively. Then,* $D(\mathrm{P}_Y\|\mathrm{Q}_Y) \leq D(\mathrm{P}_X\|\mathrm{Q}_X)$.

Proof The proofs of all these properties use the log-sum inequality (see Lemma 2.9). In fact, the first property is equivalent to the log-sum inequality with vectors $(\mathrm{P}(x), x \in \mathcal{X})$ and $(\mathrm{Q}(x), x \in \mathcal{X})$ in the role of $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$, respectively.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

For the second property, consider pairs $(\mathrm{P}_1, \mathrm{Q}_1)$ and $(\mathrm{P}_2, \mathrm{Q}_2)$ of distributions on $\mathcal{X}$. Further, for $\theta \in [0,1]$, consider the pair $(\mathrm{P}_\theta, \mathrm{Q}_\theta)$ of distributions on $\mathcal{X}$ given by $\mathrm{P}_\theta := \theta\mathrm{P}_1 + (1-\theta)\mathrm{P}_2$ and $\mathrm{Q}_\theta := \theta\mathrm{Q}_1 + (1-\theta)\mathrm{Q}_2$. Then, we have

$$\begin{aligned} D(\mathrm{P}_\theta \| \mathrm{Q}_\theta) &= \sum_x \mathrm{P}_\theta(x) \log \frac{\mathrm{P}_\theta(x)}{\mathrm{Q}_\theta(x)} \\ &= \sum_x (\theta\mathrm{P}_1(x) + (1-\theta)\mathrm{P}_2(x)) \log \frac{\theta\mathrm{P}_1(x) + (1-\theta)\mathrm{P}_2(x)}{\theta\mathrm{Q}_1(x) + (1-\theta)\mathrm{Q}_2(x)} \\ &\le \sum_x \theta\mathrm{P}_1(x) \log \frac{\theta\mathrm{P}_1(x)}{\theta\mathrm{Q}_1(x)} + (1-\theta)\mathrm{P}_2(x) \log \frac{(1-\theta)\mathrm{P}_2(x)}{(1-\theta)\mathrm{Q}_2(x)} \\ &= \theta D(\mathrm{P}_1 \| \mathrm{Q}_1) + (1-\theta) D(\mathrm{P}_2 \| \mathrm{Q}_2), \end{aligned}$$

where the inequality is by the log-sum inequality.

Finally, for the data processing inequality, with $W(y|x) := \Pr(T(x) = y \mid X = x)$, we get

$$\begin{aligned} D(\mathrm{P}_Y \| \mathrm{Q}_Y) &= \sum_{y \in \mathcal{Y}} \mathrm{P}_Y(y) \log \frac{\mathrm{P}_Y(y)}{\mathrm{Q}_Y(y)} \\ &= \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \mathrm{P}_X(x) W(y|x) \log \frac{\sum_{x \in \mathcal{X}} \mathrm{P}_X(x) W(y|x)}{\sum_{x \in \mathcal{X}} \mathrm{Q}_X(x) W(y|x)} \\ &\le \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \mathrm{P}_X(x) W(y|x) \log \frac{\mathrm{P}_X(x) W(y|x)}{\mathrm{Q}_X(x) W(y|x)} \\ &= D(\mathrm{P}_X \| \mathrm{Q}_X), \end{aligned}$$

where the inequality uses the log-sum inequality and our convention that $0 \log(0/0) = 0$. □

The convexity of $D(\mathrm{P}\|\mathrm{Q})$ in the pair (P, Q) is a very useful property. In particular, it implies that $D(\mathrm{P}\|\mathrm{Q})$ is convex in P for a fixed Q and convex in Q for a fixed P. Also, later in the book, we will see a connection between KL divergence and probability of error for binary hypothesis testing, and the data processing inequality for KL divergence has similar interpretation to that for total variation distance – adding noise ("data processing") gets distributions closer and makes it harder to distinguish them.

We close this section with a chain rule for KL divergence.

LEMMA 2.18 (Chain rule for KL divergence) *For distributions* $\mathrm{P}_{X_1 \cdots X_n}$ *and* $\mathrm{Q}_{X_1 \cdots X_n}$ *on a discrete set* $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$*, we have*

$$D(\mathrm{P}_{X_1 \cdots X_n} \| \mathrm{Q}_{X_1 \cdots X_n}) = \sum_{i=1}^{n} \mathbb{E}_{\mathrm{P}_{X^{i-1}}} \left[D(\mathrm{P}_{X_i|X^{i-1}} \| \mathrm{Q}_{X_i|X^{i-1}}) \right],$$

where X^i *abbreviates the random variable* $(X_1, \ldots, X_i)$.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

Proof It suffices to show the result for $n = 2$. We have

$$\begin{aligned} D(\mathrm{P}_{XY}\|\mathrm{Q}_{XY}) &= \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{\mathrm{P}_{XY}(X,Y)}{\mathrm{Q}_{XY}(X,Y)}\right] \\ &= \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{\mathrm{P}_X(X)}{\mathrm{Q}_X(X)} + \log \frac{\mathrm{P}_{Y|X}(Y|X)}{\mathrm{Q}_{Y|X}(Y|X)}\right] \\ &= \mathbb{E}_{\mathrm{P}_X}\left[\log \frac{\mathrm{P}_X(X)}{\mathrm{Q}_X(X)}\right] + \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{\mathrm{P}_{Y|X}(Y|X)}{\mathrm{Q}_{Y|X}(Y|X)}\right] \\ &= D(\mathrm{P}_X\|\mathrm{Q}_X) + \mathbb{E}_{\mathrm{P}_X}\left[D(\mathrm{P}_{Y|X}\|\mathrm{Q}_{Y|X})\right]; \end{aligned}$$

the proof for general n is obtained by applying this identity recursively. □

We note that, unlike the chain rule for total variation distance, the chain rule above holds with equality. In particular, KL divergence is seen to be *additive* for product distributions; namely, for $\mathrm{P}_{X_1\cdots X_n} = \prod_{i=1}^n \mathrm{P}_{X_i}$ and $\mathrm{Q}_{X_1\cdots X_n} = \prod_{i=1}^n \mathrm{Q}_{X_i}$, we have

$$D(\mathrm{P}_{X_1\cdots X_n}\|\mathrm{Q}_{X_1\cdots X_n}) = \sum_{t=1}^n D(\mathrm{P}_{X_i}\|\mathrm{Q}_{X_i}).$$

2.6 Shannon Entropy

Probabilistic modeling allows us to capture uncertainty in our knowledge of a quantity (modeled as a random variable). But to build a formal theory for security, we need to quantify what it means to have partial knowledge of a random variable – to have a "bit" of knowledge about a random variable X. Such a quantification of uncertainty is provided by the information-theoretic notion of *Shannon entropy*.

DEFINITION 2.19 (Shannon entropy) For a random variable X, the Shannon entropy of X is given by[6]

$$H(\mathrm{P}_X) := \sum_{x\in\mathcal{X}} \mathrm{P}_X(x) \log \frac{1}{\mathrm{P}_X(x)},$$

where log is to the base 2. For brevity, we often abbreviate $H(\mathrm{P}_X)$ as $H(X)$. However, the reader should keep in mind that H is a function of the distribution P_X of X.

We have not drawn this quantity out of the hat and proposed it as a measure of uncertainty. There is a rich theory supporting the role of entropy as a measure of uncertainty or a measure of randomness. However, the details are beyond the scope of our book. In fact, a heuristic justification for entropy as a measure of randomness comes from the following observation: denoting by $\mathrm{P}_{\mathtt{unif}}$ the uniform distribution on $\mathcal{X}$, we have

$$H(\mathrm{P}) = \log|\mathcal{X}| - D(\mathrm{P}\|\mathrm{P}_{\mathtt{unif}}). \tag{2.6}$$

[6] We follow the convention $0 \log 0 = 0$.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

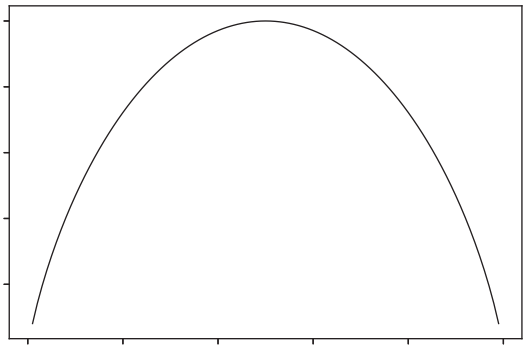

Figure 2.2 The binary entropy function $h(p)$.

That is, Shannon entropy $H(\mathrm{P})$ is a measure of how far P is from a uniform distribution. Heuristically, the uniform distribution is the "most random" distribution on $\mathcal{X}$, and therefore, Shannon entropy is indeed a measure of uncertainty or randomness.

We present the properties of Shannon entropy, which can be derived readily from the properties of KL divergence and (2.6).

THEOREM 2.20 (Shape of the Shannon entropy functional) *Consider a probability distribution* P *on a finite set* $\mathcal{X}$*. Then, the following properties hold.*

1. (Nonnegativity) $H(\mathrm{P}) \geq 0$ *with equality if and only if* $|\mathrm{supp}(\mathrm{P})| = 1$*, namely* P *is the distribution of a constant random variable.*
2. (Boundedness) $H(\mathrm{P}) \leq \log |\mathcal{X}|$ *with equality if and only if* P *is the uniform distribution on* $\mathcal{X}$*.*
3. (Concavity) $H(\mathrm{P})$ *is a concave function of* P*, namely for every* $\theta \in [0,1]$ *and two probability distributions* Q_1 *and* Q_2 *on* $\mathcal{X}$*,*

$$H(\theta \mathrm{Q}_1 + (1-\theta)\mathrm{Q}_2) \geq \theta H(\mathrm{Q}_1) + (1-\theta) H(\mathrm{Q}_2).$$

Proof The nonnegativity property is easy to see: each term in the expression for entropy is nonnegative, whereby $H(\mathrm{P})$ is 0 if and only if each term in the sum is 0. This can only happen if $\mathrm{P}(x) = 1$ for one $x \in \mathcal{X}$ and $\mathrm{P}(x) = 0$ for the rest.

The boundedness property follows from (2.6) using the nonnegativity of KL divergence. Further, the concavity of Shannon entropy also follows from (2.6) using the convexity of $D(\mathrm{P} \| \mathrm{P}_{\mathtt{unif}})$ in P. □

For the special case of binary random variables ($\mathcal{X} = \{0,1\}$), Shannon entropy $H(\mathrm{P})$ depends only on $p := \mathrm{P}(1)$ and is denoted using the function $h\colon [0,1] \to [0,1]$, termed the *binary entropy function*. That is, $h(p) := p \log \frac{1}{p} + (1-p) \log \frac{1}{(1-p)}$, $p \in [0,1]$. We depict $h(p)$ in Figure 2.2.

Next, we seek a notion of residual uncertainty, the uncertainty remaining in X when a correlated random variable Y is revealed. Such a notion is given

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

by the *conditional Shannon entropy*, or simply, conditional entropy, defined below.

DEFINITION 2.21 (Conditional Shannon entropy) For discrete random variables X and Y, the conditional Shannon entropy $H(X|Y)$ is given by $\mathbb{E}_{\mathrm{P}_Y}\left[H(\mathrm{P}_{X|Y})\right]$.

Using the expression for Shannon entropy, it is easy to check that

$$H(Y|X) = \sum_{x,y} \mathrm{P}_{XY}(x,y) \log \frac{1}{\mathrm{P}_{Y|X}(y|x)} = \mathbb{E}\left[-\log \mathrm{P}_{Y|X}(Y|X)\right].$$

Note that the random variable in the expression on the right-hand side is $-\log \mathrm{P}_{Y|X}(Y|X)$, and we take expectation over $(X,Y) \sim \mathrm{P}_{XY}$.

We present now a chain rule for Shannon entropy, which allows us to divide the *joint entropy* $H(X_1,\ldots,X_n)$ of random variables $(X_1,\ldots,X_n)$ into "smaller" components.

LEMMA 2.22 (Chain rule for entropy) *For discrete random variables* $(X_1,\ldots,X_n)$ *and* Y*, we have*

$$H(X_1,\ldots,X_n \mid Y) = \sum_{i=1}^{n} H(X_i|X^{i-1},Y). \tag{2.7}$$

Proof We can derive this using the chain rule for KL divergence and (2.6). Alternatively, we can see it directly as follows. First consider the case $n=2$ and Y is a constant. We have

$$\begin{aligned} H(X_2|X_1) &= \mathbb{E}\left[-\log \mathrm{P}_{X_2|X_1}(X_2|X_1)\right] \\ &= \mathbb{E}[-\log \mathrm{P}_{X_1X_2}(X_1,X_2)] + \mathbb{E}[\log \mathrm{P}_{X_1}(X_1)] \\ &= H(X_1,X_2) - H(X_1), \end{aligned}$$

where we used the Bayes rule in the second identity. The result for general n, but with Y constant, is obtained by applying this result recursively. The more general result when Y is not constant follows from (2.7) with constant Y upon noting that $H(X_1,\ldots,X_n|Y) = H(X_1,\ldots,X_n,Y) - H(Y)$. □

We close this section by commenting on the notation $H(Y|X)$, which is, admittedly, a bit informal. It will be more appropriate to view conditional entropy as a function of (P,W) where P is the distribution of X and W is the channel with X as the input and Y as the output. In particular, we use the notation $H(W|\mathrm{P})$ to denote $H(Y|X)$. Note that

$$H(W|\mathrm{P}) = \sum_{x\in\mathcal{X}} \mathrm{P}(x) \sum_{y\in\mathcal{Y}} W(y|x) \log \frac{1}{W(y|x)},$$

and that $H(W|\mathrm{P})$ is a linear function of P and a concave function of W.

Finally, we note the following important consequence of concavity of $H(P)$.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

LEMMA 2.23 (Conditioning reducing entropy) *For a probability distribution* P *on a finite set* $\mathcal{X}$ *and a channel* $W\colon \mathcal{X} \to \mathcal{Y}$*, we have* $H(W|\mathrm{P}) \le H(\mathrm{P}\circ W)$*. (In our alternative notation,* $H(Y|X) \le H(Y)$ *for all random variables* (X,Y)*.)*

Proof Since $H(\cdot)$ is a concave function, we have

$$H(W|\mathrm{P}) = \sum_x \mathrm{P}(x) H(W_x) \le H\left(\sum_x \mathrm{P}(x) W_x\right) = H(\mathrm{P}\circ W),$$

where we abbreviate the output distribution $W(\cdot|x)$ as W_x. □

2.7 Mutual Information

How much information does a random variable Y reveal about another random variable X? A basic postulate of information theory is that "Information is reduction in Uncertainty." We already saw how to measure uncertainty: $H(X)$ is the uncertainty in a random variable X and $H(X|Y)$ is the uncertainty remaining in X once Y is revealed. Thus, a measure of information provided by the postulate above is $H(X) - H(X|Y)$. This fundamental measure of information is called the *mutual information*.

DEFINITION 2.24 (Mutual information) Given a joint distribution P_{XY}, the mutual information between X and Y is given by $I(X\wedge Y) = H(X) - H(X|Y)$. We will use an alternative definition where we represent the mutual information as a function of the input distribution $\mathrm{P} = \mathrm{P}_X$ and the channel $W = \mathrm{P}_{Y|X}$. Namely, we represent mutual information $I(X\wedge Y)$ as $I(\mathrm{P}, W)$.

Before proceeding, we note several alternative expressions for mutual information.

LEMMA 2.25 (Alternative expressions for mutual information) *For discrete random variables* (X,Y)*, the following quantities are equal to* $I(X\wedge Y)$*:*

1. $H(Y) - H(Y|X)$;
2. $H(X) + H(Y) - H(X,Y)$;
3. $H(X,Y) - H(X|Y) - H(Y|X)$;
4. $D(\mathrm{P}_{XY}\|\mathrm{P}_X \times \mathrm{P}_Y)$.

Proof The equality of $H(X) - H(X|Y)$ with expressions in 1–3 follows upon noting that $H(X|Y) = H(X,Y) - H(Y)$ and $H(Y|X) = H(X,Y) - H(X)$. The equality with $D(\mathrm{P}_{XY}\|\mathrm{P}_X\times\mathrm{P}_Y)$ follows since $H(X,Y) = \mathbb{E}_{\mathrm{P}_{XY}}[-\log \mathrm{P}_{XY}(X,Y)]$, $H(X) = \mathbb{E}_{\mathrm{P}_{XY}}[-\log \mathrm{P}_X(X)]$, and $H(Y) = \mathbb{E}_{\mathrm{P}_{XY}}[-\log \mathrm{P}_Y(Y)]$, whereby

$$H(X) + H(Y) - H(X,Y) = \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{\mathrm{P}_{XY}(X,Y)}{\mathrm{P}_X(X)\,\mathrm{P}_Y(Y)}\right] = D(\mathrm{P}_{XY}\|\mathrm{P}_X \times \mathrm{P}_Y).$$

□

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

The simple mathematical expressions above lead to a rather remarkable theory. First, we observe that the information revealed by Y about X, $I(X \wedge Y)$, coincides with the information revealed by X about Y, $I(Y \wedge X)$. Further, $I(X \wedge Y) = D(\mathrm{P}_{XY} \| \mathrm{P}_X \times \mathrm{P}_Y)$ allows us to interpret mutual information as a measure of how "dependent" X and Y are. In particular, $I(X \wedge Y) = 0$ if and only if X and Y are independent, and equivalently, $H(X) = H(X|Y)$ if and only if X and Y are independent. Also, since KL divergence is nonnegative, it follows that conditioning reduces entropy, a fact we already saw using concavity of entropy.

Next, we define *conditional mutual information* of X and Y given Z as

$$I(X \wedge Y|Z) := H(X|Z) - H(X|Y,Z).$$

It is easy to verify that

$$I(X \wedge Y|Z) = \mathbb{E}_{\mathrm{P}_Z}\left[D(\mathrm{P}_{XY|Z} \| \mathrm{P}_{X|Z} \times \mathrm{P}_{Y|Z})\right],$$

whereby $I(X \wedge Y|Z) = 0$ if and only if X and Y are independent given Z. Further, we can use the chain rule for KL divergence or Shannon entropy to obtain the following chain rule for mutual information.

LEMMA 2.26 (Chain rule for mutual information) *For discrete random variables $(X_1, \ldots, X_n, Y)$, we have*

$$I(X_1, X_2, \ldots, X_n \wedge Y) = \sum_{i=1}^{n} I(X_i \wedge Y \mid X^{i-1}),$$

where $X^{i-1} = (X_1, \ldots, X_{i-1})$ for $1 < i \leq n$ and X^0 is a constant random variable.

Finally, we present a data processing inequality for mutual information which is, in fact, a consequence of the data processing inequality for KL divergence. To present this inequality, we need to introduce the notion of a Markov chain.

DEFINITION 2.27 (Markov chains) Random variables X, Y, Z form a Markov chain if X and Z are independent when conditioned on Y, i.e., when $I(X \wedge Z|Y) = 0$ or, equivalently, $\mathrm{P}_{XYZ} = \mathrm{P}_{X|Y}\mathrm{P}_{Z|Y}\mathrm{P}_Y$. This definition extends naturally to multiple random variables: $X_1, \ldots, X_n$ form a Markov chain if $X^{i-1} = (X_1, \ldots, X_{i-1})$, X_i, and $X_{i+1}^n = (X_{i+1}, \ldots, X_n)$ form a Markov chain for every $1 \leq i \leq n$. We use the notation $X_1 \multimap X_2 \multimap \cdots \multimap X_n$ to indicate that $X_1, \ldots, X_n$ form a Markov chain.[7]

A specific example is when $Z = f(Y)$ for some function f. In this case, for every X we have $X \multimap Y \multimap Z$. Heuristically, if $X \multimap Y \multimap Z$ holds, then Z can contain no more information about X than Y. The following result establishes this bound formally.

LEMMA 2.28 (Data processing inequality for mutual information) *If $X \multimap Y \multimap Z$, then $I(X \wedge Z) \leq I(X \wedge Y)$. Equivalently, $H(X|Y) \leq H(X|Z)$.*

[7] That is, there is no more information about X_{i+1}^n in X^i than that contained in X_i.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

Proof Instead of taking recourse to the data processing inequality for KL divergence, we present an alternative proof. We have

$$\begin{aligned} I(X \wedge Z) &= I(X \wedge Y, Z) - I(X \wedge Y|Z) \\ &\leq I(X \wedge Y, Z) \\ &= I(X \wedge Y) + I(X \wedge Z|Y) \\ &= I(X \wedge Y), \end{aligned}$$

where the inequality holds since conditional mutual information is nonnegative and the final identity holds since $X \multimap Y \multimap Z$. □

2.8 Fano's Inequality

We now prove Fano's inequality – a lower bound for the probability of error for guessing a random variable X using another random variable Y. In particular, Fano's inequality provides a lower bound for the probability of error in terms of the mutual information $I(X \wedge Y)$ between X and Y. Alternatively, we can view Fano's inequality as an upper bound for the conditional entropy $H(X|Y)$, in terms of the probability of error. It is a very useful inequality and is applied widely in information theory, statistics, communications, and other related fields. In fact, the proof of Fano's inequality we present uses nothing more than the data processing inequality.

THEOREM 2.29 (Fano's inequality) *For discrete random variables X and Y, consider $\hat{X}$ such that $X \multimap Y \multimap \hat{X}$.*[8] *Then, we have*

$$H(X|Y) \leq \Pr(\hat{X} \neq X) \log(|\mathcal{X}| - 1) + h\big(\Pr(\hat{X} \neq X)\big),$$

where $h(t) = -t \log t - (1-t) \log(1-t)$ is the binary entropy function.

Proof Instead of P_{XY}, consider the distribution Q_{XY} given by

$$\mathrm{Q}_{XY}(x, y) = \frac{1}{|\mathcal{X}|} \mathrm{P}_Y(y), \quad x \in \mathcal{X}, y \in \mathcal{Y},$$

namely, the distribution when X is uniform and independent of Y. We can treat the estimate $\hat{X}$ as a randomized function of Y, expressed using the channel $\mathrm{P}_{\hat{X}|Y}$. Let $\mathrm{Q}_{XY\hat{X}}$ be given by

$$\mathrm{Q}_{XY\hat{X}}(x, y, \hat{x}) = \mathrm{Q}_{XY}(x, y) \mathrm{P}_{\hat{X}|Y}(\hat{x}|y), \quad \forall\, x, \hat{x} \in \mathcal{X}, y \in \mathcal{Y}.$$

We note that the probability of correctness of the estimate $\hat{X}$ under Q_{XY} is the same as that of a "random guess." Indeed, we have

$$\mathrm{Q}_{XY\hat{X}}(X = \hat{X}) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{Q}_{XY}(x, y) \mathrm{P}_{\hat{X}|Y}(x|y)$$

[8] We can view $\hat{X}$ as an "estimate" of X formed from Y.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

$$= \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{P}_Y(y)\, \mathrm{P}_{\hat{X}|Y}(x|y)$$
$$= \frac{1}{|\mathcal{X}|}. \tag{2.8}$$

The main idea behind our proof is the following. For any distribution P_{XY}, the difference between the performance of the estimator $\hat{X}$ under P_{XY} and Q_{XY}, the independent distribution, is bounded by the "distance" between these distributions. We formalize this using the data processing inequality.

Formally, consider the channel $W : \mathcal{X} \times \mathcal{Y} \times \mathcal{X} \to \{0, 1\}$ given by $W(1|x, y, \hat{x}) = \mathbf{1}[x = \hat{x}]$. Then, by the data processing inequality we get

$$\begin{aligned} D(\mathrm{P}_{XY\hat{X}} \circ W \| \mathrm{Q}_{XY\hat{X}} \circ W) &\le D(\mathrm{P}_{XY\hat{X}} \| \mathrm{Q}_{XY\hat{X}}) \\ &= D(\mathrm{P}_{XY} \| \mathrm{Q}_{XY}) \\ &= \log |\mathcal{X}| - H(X|Y), \end{aligned}$$

where we used the chain rule for KL divergence in the first identity and the second identity can be verified by a direct calculation (see Problem 2.2) . Further, denoting $p = \mathrm{P}_{XY\hat{X}}(X = \hat{X})$ and $q = \mathrm{Q}_{XY\hat{X}}(X = \hat{X})$, we get

$$\begin{aligned} D(\mathrm{P}_{XY\hat{X}} \circ W \| \mathrm{Q}_{XY\hat{X}} \circ W) &= p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \\ &= \log |\mathcal{X}| - (1-p) \log(|\mathcal{X}| - 1) - h(1-p), \end{aligned}$$

where in the second identity we used the expression for probability of correctness q under $\mathrm{Q}_{XY\hat{X}}$ computed in (2.8). Upon combining this expression with the previous bound, we get

$$H(X|Y) \le (1-p) \log(|\mathcal{X}| - 1) + h(1-p),$$

which completes the proof since $\mathrm{P}_{XY\hat{X}}(X \neq \hat{X}) = 1 - p$. □

When X is uniform, it follows from Fano's inequality that

$$\Pr(X \neq \hat{X}) \ge 1 - \frac{I(X \wedge Y) + 1}{\log |\mathcal{X}|},$$

an inequality used popularly for deriving lower bounds in statistics.

2.9 Maximal Coupling Lemma

Earlier in Section 2.4, we saw that $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) = \max_{\mathcal{A} \subset \mathcal{X}} (\mathrm{P}(\mathcal{A}) - \mathrm{Q}(\mathcal{A}))$, a formula that expresses total variation distance as an optimization problem. Such expressions are sometimes called "variational formulae," leading to the alternative name *variational distance* for total variation distance. In fact, we saw in Lemma 2.10 an operational interpretation of this formula, where we can associate with $\mathcal{A}$ the hypothesis test which declares P when $X \in \mathcal{A}$. In this section, we will see yet another variational formula for total variation distance, which

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

is equally important and interesting. This one also gives another operational meaning to the total variation distance, in the context of *optimal transportation cost*.

As a motivation, consider the following optimal transportation problem. A commodity is stored across multiple warehouses labeled by elements of $\mathcal{X}$, with warehouse $x \in \mathcal{X}$ having a fraction $\mathrm{P}(x)$ of it. We need to transfer this commodity to multiple destinations, labeled again by the elements of $\mathcal{X}$, with destination $y \in \mathcal{X}$ seeking $\mathrm{Q}(y)$ fraction of it. Towards that, we assign a fraction $W(y|x)$ of commodity at $x \in \mathcal{X}$ to be shipped to $y \in \mathcal{X}$. Suppose that we incur a cost of 1 when we send a "unit" of the commodity from source x to a destination y that differs from x, and no cost when sending from x to x. What is the minimum possible cost that we can incur?

More precisely, we can represent the input fractions using a random variable X with probability distribution P and the output fractions using a random variable Y with probability distribution Q. While the marginal distributions of X and Y are fixed, our assignment W defines a joint distribution for X and Y. Such a joint distribution is called a *coupling* of distributions P and Q.

DEFINITION 2.30 (Coupling) For two probability distributions P and Q on $\mathcal{X}$, a coupling of P and Q is a joint distribution P_{XY} such that $\mathrm{P}_X = \mathrm{P}$ and $\mathrm{P}_Y = \mathrm{Q}$. The set of all couplings of P and Q is denoted by $\pi(\mathrm{P}, \mathrm{Q})$. Note that $\pi(\mathrm{P}, \mathrm{Q})$ contains $\mathrm{P} \times \mathrm{Q}$ and is, therefore, nonempty

We now express the previous optimal transportation problem using this notion of couplings. An assignment W coincides with a coupling P_{XY} of P and Q, and the cost incurred by this assignment is given by

$$C(X, Y) := \mathbb{E}_{\mathrm{P}_{XY}}[\mathbf{1}[X \neq Y]] = \Pr(X \neq Y)\,.$$

Thus, in the optimal transport problem specified above, the goal is to find the minimum cost[9]

$$C^*(\mathrm{P}, \mathrm{Q}) = \min_{\mathrm{P}_{XY} \in \pi(\mathrm{P},\mathrm{Q})} C(X, Y) = \min_{\mathrm{P}_{XY} \in \pi(\mathrm{P},\mathrm{Q})} \Pr(X \neq Y)\,.$$

Interestingly, $C^*(\mathrm{P}, \mathrm{Q})$ coincides with $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})$.

LEMMA 2.31 (Maximal coupling lemma) *For probability distributions* P *and* Q *on* $\mathcal{X}$*, we have*

$$d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) = C^*(\mathrm{P}, \mathrm{Q}).$$

Proof Consider a coupling $\mathrm{P}_{XY} \in \pi(\mathrm{P}, \mathrm{Q})$. Then, for any x, we have

$$\begin{aligned}
\mathrm{P}(x) &= \Pr(X = x, Y \neq X) + \Pr(X = x, Y = X) \\
&\leq \Pr(X = x, Y \neq X) + \Pr(Y = x) \\
&= \Pr(X = x, Y \neq X) + \mathrm{Q}(x),
\end{aligned}$$

[9] We will see soon that there is a coupling that attains the minimum, justifying the use of min instead of inf in the definition.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

whereby

$$\mathrm{P}(x) - \mathrm{Q}(x) \le \Pr(X = x, Y \neq X)\,, \quad \forall x \in \mathcal{X}.$$

Summing over x such that $\mathrm{P}(x) \ge \mathrm{Q}(x)$, we get

$$d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) \le \sum_{x:\mathrm{P}(x)\ge\mathrm{Q}(x)} \Pr(X = x, Y \neq X) \le \Pr(Y \neq X)\,.$$

Since this bound holds for every coupling $\mathrm{P}_{XY} \in \pi(\mathrm{P}, \mathrm{Q})$, we obtain

$$d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) \le C^*(\mathrm{P}, \mathrm{Q}).$$

For the other direction, let U be a binary random variable, with $\Pr(U = 0) = \sum_{x\in\mathcal{X}} \min\{\mathrm{P}(x), \mathrm{Q}(x)\}$. Noting that

$$\begin{aligned} 1 - \sum_{x\in\mathcal{X}} \min\{\mathrm{P}(x), \mathrm{Q}(x)\} &= \sum_{x\in\mathcal{X}} (\mathrm{P}(x) - \min\{\mathrm{P}(x), \mathrm{Q}(x)\}) \\ &= \sum_{x\in\mathcal{X}:\mathrm{P}(x)\ge\mathrm{Q}(x)} (\mathrm{P}(x) - \mathrm{Q}(x)) \\ &= d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}), \end{aligned}$$

i.e., U is the Bernoulli random variable with parameter $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})$. Conditioned on $U = 0$, we sample $X = x, Y = y$ with probability

$$\Pr(X = x, Y = y | U = 0) = \min\{\mathrm{P}(x), \mathrm{Q}(x)\}\mathbf{1}[x = y]\,/(1 - d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})).$$

Conditioned on $U = 1$, we sample $X = x$ and $Y = y$ with probability

$$\begin{aligned} &\Pr(X = x, Y = y | U = 1) \\ &\quad = \frac{(\mathrm{P}(x) - \mathrm{Q}(x))(\mathrm{Q}(y) - \mathrm{P}(y))}{d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})^2} \mathbf{1}[\mathrm{P}(x) \ge \mathrm{Q}(x), \mathrm{Q}(y) \ge \mathrm{P}(y)]\,. \end{aligned}$$

Thus, $X = Y$ if and only if $U = 0$, whereby

$$\Pr(X \neq Y) = \Pr(U = 1) = d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}).$$

It only remains to verify that $\mathrm{P}_{XY} \in \pi(\mathrm{P}, \mathrm{Q})$. Indeed, note that

$$\begin{aligned} \sum_{y\in\mathcal{Y}} \mathrm{P}_{XY}(x, y) &= \Pr(U = 0) \sum_{y\in\mathcal{X}} \Pr(X = x, Y = y | U = 0) \\ &\quad + \Pr(U = 1) \sum_{y\in\mathcal{X}} \Pr(X = x, Y = y | U = 1) \\ &= \min\{\mathrm{P}(x), \mathrm{Q}(x)\} \sum_{y\in\mathcal{X}} \mathbf{1}[y = x] \\ &\quad + (\mathrm{P}(x) - \mathrm{Q}(x))\mathbf{1}[\mathrm{P}(x) \ge \mathrm{Q}(x)] \sum_{y\in\mathcal{X}:\ \mathrm{Q}(y)\ge\mathrm{P}(y)} \frac{(\mathrm{Q}(y) - \mathrm{P}(y))}{d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})} \\ &= \min\{\mathrm{P}(x), \mathrm{Q}(x)\} + (\mathrm{P}(x) - \mathrm{Q}(x))\mathbf{1}[\mathrm{P}(x) \ge \mathrm{Q}(x)] \\ &= \mathrm{P}(x) \end{aligned}$$

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

for every $x \in \mathcal{X}$, and similarly, $\sum_{x\in\mathcal{X}} \mathrm{P}_{XY}(x,y) = \mathrm{Q}(y)$, which shows that $C^*(\mathrm{P},\mathrm{Q}) \le d_{\mathtt{var}}(\mathrm{P},\mathrm{Q})$ and completes the proof. □

2.10 A Variational Formula for KL Divergence

We have seen two variational formulae for total variation distance. In fact, a very useful variational formula can be given for KL divergence as well.

LEMMA 2.32 (A variational formula for KL divergence) *For probability distributions* P *and* Q *on a set* $\mathcal{X}$ *such that* $\mathrm{supp}(\mathrm{P}) \subset \mathrm{supp}(\mathrm{Q})$*, we have*

$$D(\mathrm{P}\|\mathrm{Q}) = \max_R \sum_{x\in\mathcal{X}} \mathrm{P}(x) \log \frac{R(x)}{\mathrm{Q}(x)},$$

where the max *is over all probability distributions* R *on* $\mathcal{X}$ *such that* $\mathrm{supp}(\mathrm{P}) \subset \mathrm{supp}(R)$*. The* max *is attained by* $R = \mathrm{P}$*.*

Proof Using the expression for KL divergence, we have

$$\begin{aligned} D(\mathrm{P}\|\mathrm{Q}) &= \sum_{x\in\mathcal{X}} \mathrm{P}(x) \log \frac{\mathrm{P}(x)}{\mathrm{Q}(x)} \\ &= \sum_x \mathrm{P}(x) \log \frac{R(x)}{\mathrm{Q}(x)} + D(\mathrm{P}\|R) \\ &\ge \sum_x \mathrm{P}(x) \log \frac{R(x)}{\mathrm{Q}(x)}, \end{aligned}$$

with equality if and only if $\mathrm{P} = R$. □

In fact, a similar formula can be attained by restricting R to a smaller family of probability distributions containing P. A particular family of interest is the "exponentially tilted family" of probability distributions given by $R_f(x) = \mathrm{Q}(x)2^{f(X)}/\mathbb{E}_{\mathrm{Q}}\left[2^{f(X)}\right]$, where $f\colon \mathcal{X} \to \mathbb{R}$, which gives the following alternative variational formula.

LEMMA 2.33 *For probability distributions* P *and* Q *on a set* $\mathcal{X}$ *such that* $\mathrm{supp}(\mathrm{P}) \subset \mathrm{supp}(\mathrm{Q})$*, we have*

$$D(\mathrm{P}\|\mathrm{Q}) = \max_f \mathbb{E}_{\mathrm{P}}[f(X)] - \log \mathbb{E}_{\mathrm{Q}}\left[2^{f(X)}\right],$$

where the maximum is over all functions $f\colon \mathcal{X} \to \mathbb{R}$ *and is attained by* $f(x) = \log(P(x)/Q(x))$*.*

2.11 Continuity of Entropy

Next, we present a bound that relates $H(\mathrm{P}) - H(\mathrm{Q})$ to $d_{\mathtt{var}}(\mathrm{P},\mathrm{Q})$. We will show that $|H(\mathrm{P}) - H(\mathrm{Q})|$ is roughly $\mathcal{O}(d_{\mathtt{var}}(\mathrm{P},\mathrm{Q}) \log 1/d_{\mathtt{var}}(\mathrm{P},\mathrm{Q}))$ with the constant

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

depending on $\log|\mathcal{X}|$. As a consequence, we find that entropy is continuous in P, for distributions P supported on a finite set $\mathcal{X}$.

THEOREM 2.34 (Continuity of entropy) *For probability distributions* P *and* Q *on* $\mathcal{X}$*, we have*

$$|H(\mathrm{P}) - H(\mathrm{Q})| \le d_{\mathtt{var}}(\mathrm{P},\mathrm{Q}) \log(|\mathcal{X}| - 1) + h(d_{\mathtt{var}}(\mathrm{P},\mathrm{Q})),$$

where $h(\cdot)$ *denotes the binary entropy function.*

Proof Consider a coupling P_{XY} of P and Q ($\mathrm{P}_X = \mathrm{P}$ and $\mathrm{P}_Y = \mathrm{Q}$). Then, $H(X) = H(\mathrm{P})$ and $H(Y) = H(\mathrm{Q})$, whereby

$$\begin{aligned} |H(\mathrm{P}) - H(\mathrm{Q})| &= |H(X) - H(Y)| = |H(X|Y) - H(Y|X)| \\ &\le \max\{H(X|Y), H(Y|X)\}. \end{aligned}$$

By Fano's inequality, we have

$$\max\{H(X|Y), H(Y|X)\} \le \Pr(X \neq Y) \log(|\mathcal{X}| - 1) + h(\Pr(X \neq Y)).$$

The bound above holds for every coupling. Therefore, choosing the coupling that attains the lower bound of $d_{\mathtt{var}}(\mathrm{P},\mathrm{Q})$ in the maximal coupling lemma (Lemma 2.31), we get

$$\max\{H(X|Y), H(Y|X)\} \le d_{\mathtt{var}}(P,Q) \log(|\mathcal{X}| - 1) + h(d_{\mathtt{var}}(P,Q)). \qquad \square$$

2.12 Hoeffding's Inequality

Earlier, in Section 2.2, we saw that a sum of i.i.d. random variables $S_n = \sum_{i=1}^n X_i$ takes values close to $n\mathbb{E}[X_1]$ with high probability as n increases. Specifically, in proving the weak law of large numbers, we used Chebyshev's inequality for S_n. We take a short detour in this section and present a "concentration inequality" which often gives better estimates for $|S_n - n\mathbb{E}[X_1]|$ than Chebyshev's inequality. This is a specific instance of the *Chernoff bound* and applies to bounded random variables – it is called *Hoeffding's inequality*. The reason for presenting this bound here is twofold: first, indeed, we use Hoeffding's inequality for our analysis in this book; and second, we will use Hoeffding's lemma to prove Pinsker's inequality in the next section.

Consider a random variable X taking values in a finite set $\mathcal{X} \subset \mathbb{R}$ and such that $\mathbb{E}[X] = 0$. By Markov's inequality applied to the random variable $e^{\lambda X}$, where $\lambda > 0$, we get

$$\begin{aligned} \Pr(X > t) &= \Pr(\lambda X > \lambda t) \\ &= \Pr\left(e^{\lambda X} > e^{\lambda t}\right) \\ &\le \mathbb{E}\left[e^{\lambda(X-t)}\right] \end{aligned}$$

for every $t \in \mathbb{R}$ and every $\lambda > 0$. This very simple bound is, in fact, very powerful, and is called the *Chernoff bound*.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

Of particular interest are *sub-Gaussian* random variables, namely random variables which have similar tail probabilities $\Pr(X > t)$ to Gaussian random variables. It is a standard fact that, for a Gaussian random variable G with zero mean and unit variance, $\Pr(G > t) \leq e^{-\frac{t^2}{2}}$. Roughly speaking, sub-Gaussian random variables are those which have similarly decaying tail probabilities. Using the Chernoff bound given above, we can convert this requirement of quadratically exponential decay of tail probabilities to that for the *log-moment generating function* $\psi_X(\lambda) := \ln \mathbb{E}\left[e^{\lambda X}\right]$, $\lambda \in \mathbb{R}$.

Formally, we have the following definition.

DEFINITION 2.35 (Sub-Gaussian random variables) A random variable X is sub-Gaussian with variance parameter σ^2 if $\mathbb{E}[X] = 0$ and for every $\lambda \in \mathbb{R}$ we have

$$\ln \mathbb{E}\left[e^{\lambda X}\right] \leq \frac{\lambda^2 \sigma^2}{2}.$$

Recall that the log-moment generating function of the standard Gaussian random variable is given by $\psi_G(\lambda) = \lambda^2/2$. Thus, the definition above requires that X has log-moment generating function dominated by that of a Gaussian random variable with variance σ^2.

Using the Chernoff bound provided above, we get a Gaussian-like tail bound for sub-Gaussian random variables.

LEMMA 2.36 (Sub-Gaussian tails) *For a sub-Gaussian random variable X with variance parameter σ^2, we have for every $t > 0$ that*

$$\Pr(X > t) \leq e^{-\frac{t^2}{2\sigma^2}}. \tag{2.9}$$

Proof Since X is sub-Gaussian with variance parameter σ^2, using the Chernoff bound, for every $\lambda > 0$ we have

$$\Pr(X > t) \leq e^{-\lambda t} \mathbb{E}\left[e^{\lambda X}\right] \leq e^{-\lambda t + \lambda^2 \sigma^2/2},$$

which upon minimizing the right-hand side over $\lambda > 0$ gives the desired bound (2.9), which is attained by $\lambda = \frac{t}{\sigma^2}$. □

Next, we make the important observation that the sum of independent sub-Gaussian random variables is sub-Gaussian too. This, when combined with the previous observations, gives a concentration bound for sums of sub-Gaussian random variables.

LEMMA 2.37 (Sum of sub-Gaussian random variables) *Let $X_1, \ldots, X_n$ be independent and sub-Gaussian random variables with variance parameters $\sigma_1^2, \ldots, \sigma_n^2$, respectively. Then, for every $t > 0$, we have*

$$\Pr\left(\sum_{i=1}^{n} X_i \geq t\right) > e^{-\frac{t^2}{2\sum_{i=1}^{n} \sigma_i^2}}.$$

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

Proof First, we note that for every $\lambda > 0$ we have

$$\mathbb{E}\Big[e^{\lambda \sum_{i=1}^n X_i}\Big] = \prod_{i=1}^n \mathbb{E}\big[e^{\lambda X_i}\big] \le e^{\lambda^2 \sum_{i=1}^n \sigma_i^2/2},$$

where we used the independence of the X_i for the identity and the fact that they are sub-Gaussian for the inequality. Thus, $\sum_{i=1}^n X_i$ is sub-Gaussian with variable parameter $\sum_{i=1}^n \sigma_i^2$, and the claim follows from Lemma 2.36. □

Finally, we come to the Hoeffding inequality, which provides a concentration bound for bounded random variable. The main technical component of the proof is to show that a bounded random variable is sub-Gaussian. We show this first.

LEMMA 2.38 (Hoeffding's lemma) *For a random variable X taking finitely many values in the interval $[a, b]$ and such that $\mathbb{E}[X] = 0$, we have*

$$\ln \mathbb{E}\big[e^{\lambda X}\big] \le \frac{(b-a)^2\lambda^2}{8}.$$

Proof As a preparation for the proof, we first note that for any random variable Y taking values in $[a, b]$, we have

$$\mathbb{V}[Y] \le \frac{(b-a)^2}{4}.$$

The first observation we make is that $\mathbb{V}[Y] = \min_{\theta\in[a,b]} \mathbb{E}\big[(Y-\theta)^2\big]$. This can be verified[10] by simply optimizing over θ. It follows that

$$\mathbb{V}[Y] \le \min_{\theta\in[a,b]} \max\{(\theta-a)^2, (b-\theta)^2\} = \frac{(a-b)^2}{4}.$$

Next, we note that the function $\psi(\lambda) = \ln \mathbb{E}\big[e^{\lambda X}\big]$ is a twice continuously differentiable function over $\lambda \in \mathbb{R}$ for a discrete and finite random variable X. Thus, by Taylor's approximation,

$$\psi(\lambda) \le \psi(0) + \psi'(0)\lambda + \max_{c\in(0,\lambda]} \psi''(c)\frac{\lambda^2}{2}.$$

A simple calculation shows that $\psi'(\lambda) = \mathbb{E}\big[Xe^{\lambda X}\big]/\mathbb{E}\big[e^{\lambda X}\big]$, and it follows that

$$\psi(0) = \psi'(0) = 0.$$

Also, differentiating once again, we get

$$\psi''(\lambda) = \frac{\mathbb{E}\big[X^2 e^{\lambda X}\big]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}\big[Xe^{\lambda X}\big]}{\mathbb{E}[e^{\lambda X}]}\right)^2.$$

Denoting by $\mathrm{Q}(x)$ the probability distribution $\mathrm{Q}(x) = \mathrm{P}_X(x)\, e^{\lambda x}/\mathbb{E}\big[e^{\lambda X}\big]$, we note that $\psi''(\lambda)$ is the variance of X under Q. Thus, by our observation earlier,

[10] We only consider discrete random variables, wherein the proofs are technically straightforward.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

$\psi''(\lambda) \leq (b-a)^2/4$. Combining these bounds with the Taylor approximation, we get

$$\psi(\lambda) \leq \frac{(b-a)^2\lambda^2}{8},$$

which completes the proof. □

Thus, a zero-mean random variable taking values in $[a, b]$ is sub-Gaussian with variance parameter $(b-a)^2/4$. We obtain Hoeffding's inequality as a consequence of this fact and Lemma 2.37. We state this final form for random variables which need not be zero-mean. This can be done simply by noting that if $X \in [a, b]$, then even $X - \mathbb{E}[X]$ takes values in an interval of length $b - a$.

THEOREM 2.39 (Hoeffding's inequality) *Consider discrete, independent random variables $X_1, \ldots, X_n$ that take values in the interval $[a_i, b_i]$, $1 \leq i \leq n$. Then, for every $t > 0$, we have*

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}.$$

2.13 Pinsker's Inequality

Returning to the discussion on information-theoretic quantities, we now establish a relation between $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})$ and $D(\mathrm{P}\|\mathrm{Q})$. Roughly, we show that $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})$ is less than $\sqrt{D(\mathrm{P}\|\mathrm{Q})}$.

THEOREM 2.40 (Pinsker's inequality) *For probability distributions* P *and* Q *on $\mathcal{X}$, we have*

$$d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})^2 \leq \frac{\ln 2}{2} D(\mathrm{P}\|\mathrm{Q}).$$

Proof We obtain Pinsker's inequality as a consequence of the variational formula for KL divergence given in Lemma 2.33 and Hoeffding's lemma (Lemma 2.38). Consider the set $\mathcal{A}$ such that $d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) = P(\mathcal{A}) - Q(\mathcal{A})$, and let $f_\lambda(x) = \lambda(\mathbf{1}[\{x \in \mathcal{A}\}] - \mathrm{Q}(\mathcal{A}))$. Then, it is easy to see that $\mathbb{E}_{\mathrm{P}}[f_\lambda(X)] = \lambda d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q})$ and $\mathbb{E}_{\mathrm{Q}}[f_\lambda(X)] = 0$. Using this specific choice of $f = f_\lambda$ in the variation formula for KL divergence given in Lemma 2.33, we get

$$D(\mathrm{P}\|\mathrm{Q}) \geq \lambda d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) - \log \mathbb{E}_{\mathrm{Q}}\left[2^{f_\lambda(X)}\right].$$

Note that the random variable $\mathbf{1}[x \in \mathcal{A}] - \mathrm{Q}(\mathcal{A})$ is zero-mean under Q and takes values between $-\mathrm{Q}(\mathcal{A})$ and $1 - \mathrm{Q}(\mathcal{A})$. Thus, by Hoeffding's lemma,

$$\log \mathbb{E}_{\mathrm{Q}}\left[2^{f_\lambda(X)}\right] = \frac{1}{\ln 2} \ln \mathbb{E}_{\mathrm{Q}}\left[e^{(\ln 2) f_\lambda(X)}\right] \leq \frac{(\ln 2)\,\lambda^2}{8}.$$

Upon combining the two bounds above, we obtain

$$D(\mathrm{P}\|\mathrm{Q}) \geq \lambda d_{\mathtt{var}}(\mathrm{P}, \mathrm{Q}) - \frac{(\ln 2)\,\lambda^2}{8}, \quad \forall\, \lambda > 0,$$

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

which on maximizing the right-hand side over $\lambda > 0$ yields the claimed inequality. □

2.14 Rényi Entropy

In addition to Shannon entropy, in this book, we rely on another related measure of uncertainty and randomness: the Rényi entropy. Below we review some basic properties of this quantity.

DEFINITION 2.41 (Rényi entropy) For a probability distribution P on $\mathcal{X}$ and $\alpha \geq 0, \alpha \neq 1$, the *Rényi entropy* of order α, denoted by $H_\alpha(\mathrm{P})$, is defined as

$$H_\alpha(\mathrm{P}) := \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} \mathrm{P}(x)^\alpha.$$

As for Shannon entropy, we use the notation $H_\alpha(X)$ and $H_\alpha(\mathrm{P})$ to denote the Rényi entropy of order α for a random variable X with distribution P.

THEOREM 2.42 (Properties of $H_\alpha(\mathrm{P})$) *For a probability distribution* P *on a finite set* $\mathcal{X}$ *and* $\alpha > 0, \alpha \neq 1$, *the following properties hold.*

1. $0 \leq H_\alpha(\mathrm{P}) \leq \log |\mathcal{X}|$.
2. $H_\alpha(\mathrm{P})$ *is a nonincreasing function of* α.
3. $\lim_{\alpha \to \infty} H_\alpha(\mathrm{P}) = \min_{x \in \mathcal{X}} - \log \mathrm{P}(x)$.
4. $\lim_{\alpha \to 1} H_\alpha(\mathrm{P}) = H(\mathrm{P})$.

Proof of 1. For $0 < \alpha < 1$ and $\alpha > 1$, respectively, we note that $\sum_{x \in \mathcal{X}} \mathrm{P}(x)^\alpha > 1$ and $\sum_{x \in \mathcal{X}} \mathrm{P}(x)^\alpha \leq 1$. Therefore, $H_\alpha(\mathrm{P}) \geq 0$.

Next, for $\alpha \in (0,1)$, note that

$$\begin{aligned} \sum_x \mathrm{P}(x)^\alpha &= \mathbb{E}_{\mathrm{P}} \left[\left(\frac{1}{\mathrm{P}(X)} \right)^{1-\alpha} \right] \\ &\leq |\mathcal{X}|^{1-\alpha}, \end{aligned}$$

where the inequality uses Jensen's inequality applied to the concave function $t^{1-\alpha}$ for $\alpha \in (0,1)$ and $t > 0$. Thus, $H_\alpha(\mathrm{P}) \leq \log |\mathcal{X}|$ for $\alpha \in (0,1)$. The proof for $\alpha > 1$ can be completed similarly by noting that $t^{1-\alpha}$ is a convex function for $\alpha > 1$ and $t > 0$.

Proof of 2. Consider the function $f(\alpha) = H_\alpha(\mathrm{P})$ for $\alpha > 0$ and $\alpha \neq 1$. Then, denoting $\mathrm{P}_\alpha(x) = \mathrm{P}(x)^\alpha / \sum_{x'} \mathrm{P}(x')^\alpha$, we can verify that

$$f'(\alpha) = -\frac{1}{(1-\alpha)^2} \sum_{x \in \mathcal{X}} \mathrm{P}_\alpha(x) \log \frac{\mathrm{P}_\alpha(x)}{\mathrm{P}(x)} = -\frac{1}{(1-\alpha)^2} D(\mathrm{P}_\alpha \| \mathrm{P}) \leq 0$$

whereby f is a nonincreasing function.

Claims 3 and 4 can be verified by directly computing the limits. □

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

From Claim 4, we can regard the Shannon entropy as the Rényi entropy of order $\alpha = 1$. The Rényi entropy for $\alpha = 0$ is $H_0(\mathrm{P}) = \log|\mathtt{supp}(\mathrm{P})|$, and it is referred to as the max-entropy, denoted $H_{\max}(\mathrm{P})$. On the other hand, the Rényi entropy for $\alpha \to \infty$ is referred to as the min-entropy, denoted $H_{\min}(\mathrm{P})$. We will see operational meanings of the max-entropy and the min-entropy in Chapter 6 and Chapter 7, respectively.

2.15 References and Additional Reading

The content of this chapter concerns many basic quantities in information theory. Many of these appeared in Shannon's seminal work [304] and the relevance of some of them in the context of security appeared in [305]. However, quantities such as total variation distance and Kullback–Leibler divergence are statistical in origin and did not directly appear in Shannon's original work. A good reference for their early use is Kullback's book [207] and references therein. But these notions are now classic and can be accessed best through standard textbooks for information theory such as [75, 88, 151]. In our presentation, some of the proofs are new and not available in these textbooks. In particular, our proof of Fano's inequality based on data processing inequality is a folklore and underlies many generalizations of Fano's inequality; our presentation is closest to that in [150]. Our discussion on the variational formula for Kullback–Leibler divergence, Hoeffding's inequality, and proof of Pinsker's inequality using these tools is based on the presentation in the excellent textbook [40] on concentration inequalities. Our bound for continuity of Shannon entropy using the maximal coupling lemma is from [10, 362] (references we found in [88, Problem 3.10]). Rényi entropy was introduced in [291] and has emerged as an important tool for single-shot results in information theory and information-theoretic cryptography.

Problems

2.1 For two distributions P and Q on a finite set $\mathcal{X}$, prove the following equivalent forms of the total variation distance for discrete distributions P and Q:

$$\begin{aligned} d_{\mathtt{var}}(\mathrm{P},\mathrm{Q}) &= \sup_{A\subset\mathcal{X}} \mathrm{P}(A) - \mathrm{Q}(A) \\ &= \sup_{A\subset\mathcal{X}} |\mathrm{P}(A) - \mathrm{Q}(A)| \\ &= \sum_{x\in\mathcal{X}:\mathrm{P}(x)\geq\mathrm{Q}(x)} \mathrm{P}(x) - \mathrm{Q}(x) \\ &= \frac{1}{2}\sum_{x\in\mathcal{X}} |\mathrm{P}(x) - \mathrm{Q}(x)|. \end{aligned}$$

2.2 For distributions P_{XY} and $\mathrm{Q}_{XY} = \mathrm{P}_{\mathtt{unif}} \times \mathrm{P}_Y$ on $\mathcal{X}\times\mathcal{Y}$, where $\mathrm{P}_{\mathtt{unif}}$ denotes the uniform distribution on $\mathcal{X}$, show that

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

$$D(\mathrm{P}_{XY}\|\mathrm{Q}_{XY}) = \log|\mathcal{X}| - H(X \mid Y).$$

The quantity on the left-hand side was used to define a security index in [90].

2.3 Show the following inequalities for entropies (see [228] for other such inequalities and their application in combinatorics):

1. $H(X_1, X_2, X_3) \leq \frac{1}{2}\left[H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_1)\right]$,
2. $H(X_1, X_2, X_3) \geq \frac{1}{2}\left[H(X_1, X_2|X_3) + H(X_2, X_3|X_1) + H(X_3, X_1|X_2)\right]$.

2.4 In this problem we outline an alternative proof of Fano's inequality. Consider random variables X, Y, and $\hat{X}$ satisfying the Markov relation $X \multimap Y \multimap \hat{X}$ and such that X and $\hat{X}$ both take values in the same finite set $\mathcal{X}$. Denote by E the random variable $\mathbf{1}\left[\hat{X} \neq X\right]$. Show that $H(X \mid \hat{X}) \leq \Pr(E = 1)\log(|\mathcal{X}| - 1) + H(E)$ and conclude that

$$H(X \mid Y) \leq \Pr\left(X \neq \hat{X}\right)\log(|\mathcal{X}| - 1) + h\left(\Pr\left(X \neq \hat{X}\right)\right).$$

2.5 In this problem we outline an alternative proof of Pinsker's inequality. Using the data processing inequality for KL divergence, show that Pinsker's inequality holds if and only if it holds for distributions on $\mathcal{X}$ with $|\mathcal{X}| = 2$. Further, show that Pinsker's inequality for binary alphabet holds, that is, show that for every $p, q \in (0, 1)$ we have

$$p\log\frac{p}{q} + (1-p)\log\frac{(1-p)}{(1-q)} \geq \frac{2}{\ln 2}\cdot(p-q)^2.$$

2.6 Let $(X_1, \ldots, X_n) \in \{0,1\}^n$ be distributed uniformly over all binary sequences with less than np ones, with $0 \leq p \leq 1/2$. Show that $\Pr(X_i = 1) \leq h(p)$ for all $1 \leq i \leq n$, and that $H(X_1, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$. Conclude that for every $t \leq np$,

$$\sum_{i=0}^{t}\binom{n}{i} \leq 2^{nh(p)}.$$

2.7 We now use Problem 2.6 and Pinsker's inequality to derive a Hoeffding-type bound. Consider i.i.d. random variables $X_1, \ldots, X_n$ with common distribution $\texttt{Bernoulli}(p)$, $0 \leq p \leq 1/2$.

1. For any sequence $\mathbf{x} \in \{0,1\}^n$ with $n\theta$ ones, show that

$$\Pr(X^n = \mathbf{x}) = 2^{-nD_2(\theta\|p) - nh(\theta)},$$

   where $D_2(q\|p)$ denotes the KL divergence between Bernoulli distributions with parameters q and p.
2. Use Problem 2.6 to conclude that $\Pr(\sum_{i=1}^n X_i = n\theta) \leq 2^{-nD_2(\theta\|p)}$ and then Pinsker's inequality to conclude that for all $\theta > p$,

$$\Pr\left(\sum_{i=1}^{n} X_i > n\theta\right) \leq ne^{-2n(p-\theta)^2}.$$

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press

2.8 Establish the following variational formula for a discrete distribution P over real numbers. For all $\lambda > 0$,

$$\log \mathbb{E}_P\left[2^{\lambda(X-\mathbb{E}_P[X])}\right] = \max_{Q:\mathtt{supp}(Q)\subset\ \mathtt{supp}(P)} \lambda(\mathbb{E}_Q[X] - \mathbb{E}_P[X]) - D(Q\|P).$$

Show that if $|X| \leq 1$ with probability 1 under P, then $\mathbb{E}_Q[X] - \mathbb{E}_P[X] \leq 2d(P,Q)$. Finally, conclude that if $P(|X| \leq 1) = 1$, then

$$\log \mathbb{E}_P\left[2^{\lambda(X-\mathbb{E}_P[X])}\right] \leq \frac{2\lambda^2}{\ln 2}.$$

2.9 For two pmfs P and Q on a finite set $\mathcal{X}$ and $0 < \theta < 1$, define

$$d_\theta(P,Q) := \max_A P(A) - \frac{1-\theta}{\theta} Q(A).$$

Suppose that $B \sim \mathtt{Ber}(\theta)$ is generated. If $B = 1$, then a sample X from P is generated. If $B = 0$, then a sample X from Q is generated. Consider the minimum probability of error in estimating B from X given by $P_e^*(\theta) = \min_{f:\mathcal{X}\to\{0,1\}} \Pr(B \neq f(X))$. Show that

$$P_e^*(\theta) = \theta(1 - d_\theta(P,Q))$$

and, further, that

$$d_\theta(P,Q) = \frac{1}{2\theta}\sum_x |\theta P(x) - (1-\theta)Q(x)| + \frac{2\theta - 1}{2\theta}.$$

2.10 Show that the Rényi entropy $H_\alpha(\mathrm{P})$ for a distribution P on a finite cardinality set $\mathcal{X}$ is a concave function of P.

https://doi.org/10.1017/9781108670203.003 Published online by Cambridge University Press