Chapter 28

Shannon Entropy and Kullback-Leibler Divergence

Section 28.1 introduces Shannon entropy and its most basic properties, including the way it measures how close a random variable is to being uniformly distributed.

Section 28.2 describes relative entropy, or Kullback-Leibler divergence, which measures the discrepancy between two probability distributions, and from which Shannon entropy can be constructed. Section 28.2.1 describes some statistical aspects of relative entropy, especially its relationship to expected log-likelihood and to Fisher information.

Section 28.3 introduces the idea of the mutual information shared by two random variables, and shows how to use it as a measure of serial dependence, like a nonlinear version of autocovariance (Section 28.3.1).

Information theory studies stochastic processes as sources of information, or as models of communication channels. It appeared in essentially its modern form with Shannon (1948), and rapidly proved to be an extremely useful mathematical tool, not only for the study of "communication and control in the animal and the machine" (Wiener, 1961), but more technically as a vital part of probability theory, with deep connections to statistical inference (Kullback, 1968), to ergodic theory, and to large deviations theory. In an introduction that's so limited it's almost a crime, we will do little more than build enough theory to see how it can fit in with the theory of inference, and then get what we need to progress to large deviations. If you want to learn more (and you should!), the deservedly-standard modern textbook is Cover and Thomas (1991), and a good treatment, at something more like our level of mathematical rigor, is Gray

(1990).[1]

28.1 Shannon Entropy

The most basic concept of information theory is that of the *entropy* of a random variable, or its distribution, often called Shannon entropy to distinguish it from the many other sorts. This is a measure of the uncertainty or variability associated with the random variable. Let's start with the discrete case, where the variable takes on only a finite or countable number of values, and everything is easier.

Definition 356 (Shannon Entropy (Discrete Case)) *The* Shannon entropy*, or just* entropy*, of a discrete random variable X is*

$$H[X] \equiv -\sum_x \mathbb{P}(X=x) \log \mathbb{P}(X=x) = -\mathbf{E}[\log \mathbb{P}(X)] \tag{28.1}$$

when the sum exists. Entropy has units of bits *when the logarithm has base 2, and* nats *when it has base e.*

The joint entropy of two random variables, $H[X,Y]$, is the entropy of their joint distribution.

The conditional entropy of X given Y, $H[X|Y]$ is

$$\begin{aligned} H[X|Y] &\equiv \sum_y \mathbb{P}(Y=y) \sum_x \mathbb{P}(X=x|Y=y) \log \mathbb{P}(X=x|Y=y) & (28.2)\\ &= -\mathbf{E}[\log \mathbb{P}(X|Y)] & (28.3)\\ &= H[X,Y] - H[Y] & (28.4) \end{aligned}$$

Here are some important properties of the Shannon entropy, presented without proofs (which are not hard).

1. $H[X] \geq 0$
2. $H[X] = 0$ iff $\exists x_0 : \; X = x_0$ a.s.
3. If X can take on $n < \infty$ different values (with positive probability), then $H[X] \leq \log n$. $H[X] = \log n$ iff X is uniformly distributed.
4. $H[X]+H[Y] \geq H[X,Y]$, with equality iff X and Y are independent. (This comes from the logarithm in the definition.)

[1] Remarkably, almost all of the post-1948 development has been either amplifying or refining themes first sounded by Shannon. For example, one of the fundamental results, which we will see in the next chapter, is the "Shannon-Macmillan-Breiman theorem", or "asymptotic equipartition property", which says roughly that the log-likelihood per unit time of a random sequence converges to a constant, characteristic of the data-generating process. Shannon's original version was convergence in probability for ergodic Markov chains; the modern form is almost sure convergence for any stationary and ergodic process. Pessimistically, this says something about the decadence of modern mathematical science; optimistically, something about the value of getting it right the first time.

5. $H[X, Y] \geq H[X]$.
6. $H[X|Y] \geq 0$, with equality iff X is a.s. constant given Y, for almost all Y.
7. $H[X|Y] \leq H[X]$, with equality iff X is independent of Y. ("Conditioning reduces entropy".)
8. $H[f(X)] \leq H[X]$, for any measurable function f, with equality iff f is invertible.

The first three properties can be summarized by saying that $H[X]$ is maximized by a uniform distribution, and minimized, to zero, by a degenerate one which is a.s. constant. We can then think of $H[X]$ as the variability of X, something like the log of the effective number of values it can take on. We can also think of it as how uncertain we are about X's value.[2] $H[X, Y]$ is then how much variability or uncertainty is associated with the pair variable X, Y, and $H[Y|X]$ is how much uncertainty remains about Y once X is known, averaging over Y. Similarly interpretations follow for the other properties. The fact that $H[f(X)] = H[X]$ if f is invertible is nice, because then f just relabels the possible values, meshing nicely with this interpretation.

A simple consequence of the above results is particularly important for later use.

Lemma 357 (Chain Rule for Shannon Entropy) *Let $X_1, X_2, \ldots X_n$ be discrete-valued random variables on a common probability space. Then*

$$H[X_1, X_2, \ldots X_n] = H[X_1] + \sum_{i=2}^{n} H[X_n|X_1, \ldots X_{n-1}] \tag{28.5}$$

PROOF: From the definitions, it is easily seen that $H[X_2|X_1] = H[X_2, X_1] - H[X_1]$. This establishes the chain rule for $n = 2$. A simple argument by induction does the rest. □

For non-discrete random variables, it is necessary to introduce a reference measure, and many of the nice properties go away.

Definition 358 (Shannon Entropy (General Case)) *The* Shannon entropy *of a random variable X with distribution μ, with respect to a reference measure ρ, is*

$$H_\rho[X] \equiv -\mathbf{E}_\mu\left[\log \frac{d\mu}{d\rho}\right] \tag{28.6}$$

[2] This line of reasoning is sometimes supplemented by saying that we are more "surprised" to find that $X = x$ the less probable that event is, supposing that surprise should go as the log of one over that probability, and defining entropy as expected surprise. The choice of the logarithm, rather than any other increasing function, is of course retroactive, though one might cobble together some kind of psychophysical justification, since the perceived intensity of a sensation often grows logarithmically with the physical magnitude of the stimulus. More dubious, to my mind, is the idea that there is any surprise *at all* when a fair coin coming up heads.

when $\mu << \rho$. Joint and conditional entropies are defined similarly. We will also write $H_\rho[\mu]$, with the same meaning. This is sometimes called differential entropy *when ρ is Lebesgue measure on Euclidean space, especially $\mathbb{R}$, and then is written $h(X)$ or $h[X]$.*

It remains true, in the general case, that $H_\rho[X|Y] = H_\rho[X,Y] - H_\rho[Y]$, provided all of the entropies are finite. The chain rule remains valid, conditioning still reduces entropy, and the joint entropy is still $\leq$ the sum of the marginal entropies, with equality iff the variables are independent. However, depending on the reference measure, $H_\rho[X]$ can be negative; e.g., if ρ is Lebesgue measure and $\mathcal{L}(X) = \delta(x)$, then $H_\rho[X] = -\infty$.

28.2 Relative Entropy or Kullback-Leibler Divergence

Some of the difficulties associated with Shannon entropy, in the general case, can be evaded by using relative entropy.

Definition 359 (Relative Entropy, Kullback-Leibler Divergence) *Given two probability distributions, $\nu << \mu$, the* relative entropy of ν with respect to μ, *or* the Kullback-Leibler divergence of ν from μ, *is*

$$D(\mu\|\nu) = -\mathbf{E}_\mu\left[\log \frac{d\nu}{d\mu}\right] \tag{28.7}$$

If ν is not absolutely continuous with respect to μ, then $D(\mu\|\nu) = \infty$.

Lemma 360 *$D(\mu\|\nu) \geq 0$, with equality iff $\nu = \mu$ almost everywhere (μ).*

Proof: From Jensen's inequality, $\mathbf{E}_\mu\left[\log \frac{d\nu}{d\mu}\right] \leq \log \mathbf{E}_\mu\left[\frac{d\nu}{d\mu}\right] = \log 1 = 0$. The second part follows from the conditions for equality in Jensen's inequality. □

Lemma 361 (Divergence and Total Variation) *For any two distributions, $D(\mu\|\nu) \geq \frac{1}{2\ln 2}\|\mu - \nu\|_1^2$.*

Proof: Algebra. See, e.g., Cover and Thomas (1991, Lemma 12.6.1, pp. 300–301). □

Definition 362 *The* conditional relative entropy, *$D(\mu(Y|X)\|\nu(Y|X))$ is*

$$D(\mu(Y|X)\|\nu(Y|X)) \equiv -\mathbf{E}_\mu\left[\log \frac{d\nu(Y|X)}{d\mu(Y|X)}\right] \tag{28.8}$$

Lemma 363 (Chain Rule for Relative Entropy) *$D(\mu(X,Y)\|\nu(X,Y)) = D(\mu(X)\|\nu(X)) + D(\mu(Y|X)\|\nu(Y|X))$*

Proof: Algebra. □

Shannon entropy can be constructed from the relative entropy.

Lemma 364 *The Shannon entropy of a discrete-valued random variable X, with distribution μ, is*

$$H[X] = \log n - D(\mu\|\upsilon) \tag{28.9}$$

where n is the number of values X can take on (with positive probability), and υ is the uniform distribution over those values.

PROOF: Algebra. □

A similar result holds for the entropy of a variable which takes values in a finite subset, of volume V, of a Euclidean space, i.e., $H_\lambda[X] = \log V - D(\mu\|\upsilon)$, where λ is Lebesgue measure and υ is the uniform probability measure on the range of X.

28.2.1 Statistical Aspects of Relative Entropy

From Lemma 361, "convergence in relative entropy", $D(\mu\|\nu_n) \to 0$ as $n \to \infty$, implies convergence in the total variation (L_1) metric. Because of Lemma 360, we can say that KL divergence has some of the properties of a metric on the space of probability distribution: it's non-negative, with equality only when the two distributions are equal (a.e.). Unfortunately, however, it is not symmetric, and it does not obey the triangle inequality. (This is why it's the KL *divergence* rather than the KL *distance*.) Nonetheless, it's *enough* like a metric that it can be used to construct a kind of geometry on the space of probability distributions, and so of statistical models, which can be extremely useful. While we will not be able to go very far into this information geometry[3], it will be important to indicate a few of the connections between information-theoretic notions, and the more usual ones of statistical theory.

Definition 365 (Cross-entropy) *The* cross-entropy of ν and μ, $Q(\mu\|\nu)$, *is*

$$Q_\rho(\mu\|\nu) \equiv -\mathbf{E}_\mu\left[\log \frac{d\nu}{d\rho}\right] \tag{28.10}$$

where ν is absolutely continuous with respect to the reference measure ρ. If the domain is discrete, we will take the reference measure to be uniform and drop the subscript, unless otherwise noted.

Lemma 366 *Suppose ν and μ are the distributions of two probability models, and $\nu << \mu$. Then the cross-entropy is the expected negative log-likelihood of the model corresponding to ν, when the actual distribution is μ. The actual or empirical negative log-likelihood of the model corresponding to ν is $Q_\rho(\nu\|\eta)$, where η is the empirical distribution.*

PROOF: Obvious from the definitions. □

[3]See Kass and Vos (1997) or Amari and Nagaoka (1993/2000). For applications to statistical inference for stochastic processes, see Taniguchi and Kakizawa (2000). For an easier general introduction, Kulhavý (1996) is hard to beat.

Lemma 367 *If $\nu << \mu << \rho$, then $Q_\rho(\mu\|\nu) = H_\rho[\mu] + D(\mu\|\nu)$.*

PROOF: By the chain rule for densities,

$$\frac{d\nu}{d\rho} = \frac{d\mu}{d\rho}\frac{d\nu}{d\mu} \tag{28.11}$$

$$\log\frac{d\nu}{d\rho} = \log\frac{d\mu}{d\rho} + \log\frac{d\nu}{d\mu} \tag{28.12}$$

$$\mathbf{E}_\mu\left[\log\frac{d\nu}{d\rho}\right] = \mathbf{E}_\mu\left[\log\frac{d\mu}{d\rho}\right] + \mathbf{E}_\mu\left[\log\frac{d\nu}{d\mu}\right] \tag{28.13}$$

The result follows by applying the definitions. □

Corollary 368 (Gibbs's Inequality) *$Q_\rho(\mu\|\nu) \geq H_\rho[\mu]$, with equality iff $\nu = \mu$ a.e.*

PROOF: Insert the result of Lemma 360 into the preceding proposition. □

The statistical interpretation of the proposition is this: The log-likelihood of a model, leading to distribution ν, can be broken into two parts. One is the divergence of ν from μ; the other just the entropy of μ, i.e., it is the same for all models. If we are considering the expected log-likelihood, then μ is the actual data-generating distribution. If we are considering the empirical log-likelihood, then μ is the empirical distribution. In either case, to maximize the likelihood is to minimize the relative entropy, or divergence. What we would like to do, as statisticians, is minimize the divergence from the data-generating distribution, since that will let us predict future values. What we *can* do is minimize divergence from the empirical distribution. The consistency of maximum likelihood methods comes down, then, to finding conditions under which a shrinking divergence from the empirical distribution guarantees a shrinking divergence from the true distribution.[4]

Definition 369 *Let $\theta \in \mathbb{R}^k$, $k < \infty$, be the parameter indexing a set $\mathcal{M}$ of statistical models, where for every θ, $\nu_\theta << \rho$, with densities p_θ. Then the* Fisher information matrix *is*

$$I_{ij}(\theta) \equiv \mathbf{E}_{\nu_\theta}\left[\left(\frac{\partial \log p_\theta}{d\theta_i}\right)\left(\frac{\partial \log p_\theta}{d\theta_j}\right)\right] \tag{28.14}$$

Corollary 370 *The Fisher information matrix is equal to the Hessian (second partial derivative) matrix of the relative entropy:*

$$I_{ij}(\theta_0) = \frac{\partial^2}{\partial\theta_i\partial\theta_j} D(\nu_{\theta_0}\|\nu_\theta) \tag{28.15}$$

[4] If we did have a triangle inequality, then we could say $D(\mu\|\nu) \leq D(\mu\|\eta) + D(\eta\|\nu)$, and it would be enough to make sure that both the terms on the RHS went to zero, say by some combination of maximizing the likelihood in-sample, so $D(\eta\|\nu)$ is small, and ergodicity, so that $D(\mu\|\eta)$ is small. While, as noted, there is no triangle inequality, under some conditions this idea is roughly right; there are nice diagrams in Kulhavý (1996).

PROOF: It is a classical result (see, e.g., Lehmann and Casella (1998, sec. 2.6.1)) that $I_{ij}(\theta) = -\mathbf{E}_{\nu_\theta}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p_\theta\right]$. The present result follows from this, Lemma 366, Lemma 367, and the fact that $H_\rho[\nu_{\theta_0}]$ is independent of θ. □

28.3 Mutual Information

Definition 371 (Mutual Information) *The mutual information between two random variables, X and Y, is the divergence of the product of their marginal distributions from their actual joint distribution:*

$$I[X;Y] \equiv D(\mathcal{L}(X,Y) \| \mathcal{L}(X) \times \mathcal{L}(Y)) \tag{28.16}$$

Similarly, the mutual information among n random variables $X_1, X_2, \ldots X_n$ is

$$I[X_1; X_2; \ldots; X_n] \equiv D(\mathcal{L}(X_1, X_2, \ldots X_n) \| \prod_{i=1}^{n} \mathcal{L}(X_i)) \tag{28.17}$$

the divergence of the product distribution from the joint distribution.

Proposition 372 *$I[X;Y] \geq 0$, with equality iff X and Y are independent.*

PROOF: Directly from Lemma 360. □

Proposition 373 *If all the entropies involved are finite,*

$$\begin{aligned} I[X;Y] &= H[X] + H[Y] - H[X,Y] && (28.18)\\ &= H[X] - H[X|Y] && (28.19)\\ &= H[Y] - H[Y|X] && (28.20) \end{aligned}$$

so $I[X;Y] \leq H[X] \wedge H[Y]$.

PROOF: Calculation. □

This leads to the interpretation of the mutual information as the reduction in uncertainty or effective variability of X when Y is known, averaging over their joint distribution. Notice that in the discrete case, we can say $H[X] = I[X;X]$, which is why $H[X]$ is sometimes known as the *self-information*.

28.3.1 Mutual Information Function

Just as with the autocovariance function, we can define a mutual information function for one-parameter processes, to serve as a measure of serial dependence.

Definition 374 (Mutual Information Function) *The* mutual information function *of a one-parameter stochastic process X is*

$$\iota(t_1, t_2) \equiv I[X_{t_1}; X_{t_2}] \tag{28.21}$$

which is symmetric in its arguments. If the process is stationary, it is a function of $|t_1 - t_2|$ alone.

Notice that, unlike the autocovariance function, ι includes *nonlinear* dependencies between X_{t_1} and X_{t_2}. Also notice that $\iota(\tau) = 0$ means that the two variables are strictly independent, not just uncorrelated.

Theorem 375 *A stationary process is mixing if* $\iota(\tau) \to 0$.

Proof: Because then the total variation distance between the joint distribution, $\mathcal{L}(X_{t_1} X_{t_2})$, and the product of the marginal distributions, $\mathcal{L}(X_{t_1})\mathcal{L}(X_{t_2})$, is being forced down towards zero, which implies mixing (Definition 338). □