WIKIPEDIA

Differential entropy

Differential entropy (also referred to as **continuous entropy**) is a concept in information theory that began as an attempt by Shannon to extend the idea of (Shannon) entropy, a measure of average surprisal of a random variable, to continuous probability distributions. Unfortunately, Shannon did not derive this formula, and rather just assumed it was the correct continuous analogue of discrete entropy, but it is not. The actual continuous version of discrete entropy is the limiting density of discrete points (LDDP). Differential entropy (described here) is commonly encountered in the literature, but it is a limiting case of the LDDP, and one that loses its fundamental association with discrete entropy.

Contents

Definition

Let X be a random variable with a probability density function f whose support is a set $\mathcal{X}$. The *differential entropy* $h(X)$ or $h(f)$ is defined as[1]:243

$$\boxed{h(X) = -\int_{\mathcal{X}} f(x) \log f(x)\, dx}$$

For probability distributions which don't have an explicit density function expression, but have an explicit quantile function expression, $Q(p)$, then $h(Q)$ can be defined in terms of the derivative of $Q(p)$ i.e. the quantile density function $Q'(p)$ as [2]:54–59

$$h(Q) = \int_0^1 \log Q'(p)\, dp.$$

As with its discrete analog, the units of differential entropy depend on the base of the logarithm, which is usually 2 (i.e., the units are bits). See logarithmic units for logarithms taken in different bases. Related concepts such as joint, conditional differential entropy, and relative entropy are defined in a similar fashion. Unlike the discrete analog, the differential entropy has an offset that depends on the units used to measure X.[3]:183–184 For example, the differential entropy of a quantity measured in millimeters will be log(1000) more than the same quantity measured in meters; a dimensionless quantity will have differential entropy of log(1000) more than the same quantity divided by 1000.

One must take care in trying to apply properties of discrete entropy to differential entropy, since probability density functions can be greater than 1. For example, the uniform distribution $\mathcal{U}(0,1/2)$ has *negative* differential entropy

$$\int_0^{\frac{1}{2}} -2\log(2)\,dx = -\log(2)\,.$$

Thus, differential entropy does not share all properties of discrete entropy.

Note that the continuous mutual information $I(X;Y)$ has the distinction of retaining its fundamental significance as a measure of discrete information since it is actually the limit of the discrete mutual information of *partitions* of X and Y as these partitions become finer and finer. Thus it is invariant under non-linear homeomorphisms (continuous and uniquely invertible maps), [4] including linear [5] transformations of X and Y, and still represents the amount of discrete information that can be transmitted over a channel that admits a continuous space of values.

For the direct analogue of discrete entropy extended to the continuous space, see limiting density of discrete points.

Properties of differential entropy

- For probability densities f and g, the Kullback–Leibler divergence $D_{KL}(f||g)$ is greater than or equal to 0 with equality only if $f = g$ almost everywhere. Similarly, for two random variables X and Y, $I(X;Y) \geq 0$ and $h(X|Y) \leq h(X)$ with equality if and only if X and Y are independent.
- The chain rule for differential entropy holds as in the discrete case[1]:253

$$h(X_1,\ldots,X_n) = \sum_{i=1}^{n} h(X_i|X_1,\ldots,X_{i-1}) \leq \sum_{i=1}^{n} h(X_i).$$

- Differential entropy is translation invariant, i.e. for a constant c.[1]:253

$$h(X+c) = h(X)$$

- Differential entropy is in general not invariant under arbitrary invertible maps.

In particular, for a constant a

$$h(aX) = h(X) + \log|a|$$

For a vector valued random variable $\mathbf{X}$ and an invertible (square) matrix $\mathbf{A}$

$$h(\mathbf{A}\mathbf{X}) = h(\mathbf{X}) + \log(|\det \mathbf{A}|)$$ [1]:253

- In general, for a transformation from a random vector to another random vector with same dimension $\mathbf{Y} = m(\mathbf{X})$, the corresponding entropies are related via

$$h(\mathbf{Y}) \leq h(\mathbf{X}) + \int f(x) \log \left| \frac{\partial m}{\partial x} \right| dx$$

where $\left| \frac{\partial m}{\partial x} \right|$ is the Jacobian of the transformation m.[6] The above inequality becomes an equality if the transform is a bijection. Furthermore, when m is a rigid rotation, translation, or combination thereof, the Jacobian determinant is always 1, and $h(Y) = h(X)$.

- If a random vector $X \in \mathbb{R}^n$ has mean zero and covariance matrix K, $h(\mathbf{X}) \leq \frac{1}{2} \log(\det 2\pi e K) = \frac{1}{2} \log[(2\pi e)^n \det K]$ with equality if and only if X is jointly gaussian (see below).[1]:254

However, differential entropy does not have other desirable properties:

- It is not invariant under change of variables, and is therefore most useful with dimensionless variables.
- It can be negative.

A modification of differential entropy that addresses these drawbacks is the **relative information entropy**, also known as the Kullback–Leibler divergence, which includes an invariant measure factor (see limiting density of discrete points).

Maximization in the normal distribution

Theorem

With a normal distribution, differential entropy is maximized for a given variance. A Gaussian random variable has the largest entropy amongst all random variables of equal variance, or, alternatively, the maximum entropy distribution under constraints of mean and variance is the Gaussian.[1]:255

Proof

Let $g(x)$ be a Gaussian PDF with mean μ and variance σ^2 and $f(x)$ an arbitrary PDF with the same variance. Since differential entropy is translation invariant we can assume that $f(x)$ has the same mean of μ as $g(x)$.

Consider the Kullback–Leibler divergence between the two distributions

$$0 \leq D_{KL}(f||g) = \int_{-\infty}^{\infty} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx = -h(f) - \int_{-\infty}^{\infty} f(x) \log(g(x)) dx.$$

Now note that

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\log(g(x))dx &= \int_{-\infty}^{\infty} f(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dx \\ &= \int_{-\infty}^{\infty} f(x)\log\frac{1}{\sqrt{2\pi\sigma^2}}dx + \log(e)\int_{-\infty}^{\infty} f(x)\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= -\tfrac{1}{2}\log(2\pi\sigma^2) - \log(e)\frac{\sigma^2}{2\sigma^2} \\ &= -\tfrac{1}{2}\left(\log(2\pi\sigma^2) + \log(e)\right) \\ &= -\tfrac{1}{2}\log(2\pi e\sigma^2) \\ &= -h(g)\end{aligned}$$

because the result does not depend on $f(x)$ other than through the variance. Combining the two results yields

$$h(g) - h(f) \geq 0$$

with equality when $f(x) = g(x)$ following from the properties of Kullback–Leibler divergence.

Alternative proof

This result may also be demonstrated using the variational calculus. A Lagrangian function with two Lagrangian multipliers may be defined as:

$$L = \int_{-\infty}^{\infty} g(x)\ln(g(x))\,dx - \lambda_0\left(1 - \int_{-\infty}^{\infty} g(x)\,dx\right) - \lambda\left(\sigma^2 - \int_{-\infty}^{\infty} g(x)(x-\mu)^2\,dx\right)$$

where *g(x)* is some function with mean μ. When the entropy of *g(x)* is at a maximum and the constraint equations, which consist of the normalization condition $\left(1 = \int_{-\infty}^{\infty} g(x)\,dx\right)$ and the requirement of fixed variance $\left(\sigma^2 = \int_{-\infty}^{\infty} g(x)(x-\mu)^2\,dx\right)$, are both satisfied, then a small variation $\delta g(x)$ about *g(x)* will produce a variation δL about L which is equal to zero:

$$0 = \delta L = \int_{-\infty}^{\infty} \delta g(x)\left(\ln(g(x)) + 1 + \lambda_0 + \lambda(x-\mu)^2\right)\,dx$$

Since this must hold for any small $\delta g(x)$, the term in brackets must be zero, and solving for *g(x)* yields:

$$g(x) = e^{-\lambda_0 - 1 - \lambda(x-\mu)^2}$$

Using the constraint equations to solve for λ_0 and λ yields the normal distribution:

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Example: Exponential distribution

Let X be an exponentially distributed random variable with parameter λ, that is, with probability density function

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0.$$

Its differential entropy is then

$$\begin{aligned} h_e(X) &= -\int_0^\infty \lambda e^{-\lambda x} \log(\lambda e^{-\lambda x})\, dx \\ &= -\left(\int_0^\infty (\log \lambda)\lambda e^{-\lambda x}\, dx + \int_0^\infty (-\lambda x)\lambda e^{-\lambda x}\, dx\right) \\ &= -\log \lambda \int_0^\infty f(x)\, dx + \lambda E[X] \\ &= -\log \lambda + 1\,. \end{aligned}$$

Here, $h_e(X)$ was used rather than $h(X)$ to make it explicit that the logarithm was taken to base *e*, to simplify the calculation.

Relation to estimator error

The differential entropy yields a lower bound on the expected squared error of an estimator. For any random variable X and estimator $\widehat{X}$ the following holds:[1]

$$\operatorname{E}[(X - \widehat{X})^2] \geq \frac{1}{2\pi e} e^{2h(X)}$$

with equality if and only if X is a Gaussian random variable and $\widehat{X}$ is the mean of X.

Differential entropies for various distributions

In the table below $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$ is the gamma function, $\psi(x) = \frac{d}{dx} \ln \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function, $B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ is the beta function, and γ_E is Euler's constant.[7]:219–230

Table of differential entropies

Distribution Name	Probability density function (pdf)	Entropy in nats	Support
Uniform	$f(x)=\frac{1}{b-a}$	$\ln(b-a)$	$[a,b]$
Normal	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\ln\left(\sigma\sqrt{2\,\pi\,e}\right)$	$(-\infty,\infty)$
Exponential	$f(x)=\lambda\exp(-\lambda x)$	$1-\ln\lambda$	$[0,\infty)$
Rayleigh	$f(x)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$	$1+\ln\frac{\sigma}{\sqrt{2}}+\frac{\gamma_E}{2}$	$[0,\infty)$
Beta	$f(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ for $0\leq x\leq 1$	$\ln B(\alpha,\beta)-(\alpha-1)[\psi(\alpha)-\psi(\alpha+\beta)]$ $-(\beta-1)[\psi(\beta)-\psi(\alpha+\beta)]$	$[0,1]$
Cauchy	$f(x)=\frac{\gamma}{\pi}\frac{1}{\gamma^2+x^2}$	$\ln(4\pi\gamma)$	$(-\infty,\infty)$
Chi	$f(x)=\frac{2}{2^{k/2}\Gamma(k/2)}x^{k-1}\exp\left(-\frac{x^2}{2}\right)$	$\ln\frac{\Gamma(k/2)}{\sqrt{2}}-\frac{k-1}{2}\psi\left(\frac{k}{2}\right)+\frac{k}{2}$	$[0,\infty)$
Chi-squared	$f(x)=\frac{1}{2^{k/2}\Gamma(k/2)}x^{\frac{k}{2}-1}\exp\left(-\frac{x}{2}\right)$	$\ln 2\Gamma\left(\frac{k}{2}\right)-\left(1-\frac{k}{2}\right)\psi\left(\frac{k}{2}\right)+\frac{k}{2}$	$[0,\infty)$
Erlang	$f(x)=\frac{\lambda^k}{(k-1)!}x^{k-1}\exp(-\lambda x)$	$(1-k)\psi(k)+\ln\frac{\Gamma(k)}{\lambda}+k$	$[0,\infty)$
F	$f(x)=\frac{n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}}{B(\frac{n_1}{2},\frac{n_2}{2})}\frac{x^{\frac{n_1}{2}-1}}{(n_2+n_1x)^{\frac{n_1+n2}{2}}}$	$\ln\frac{n_1}{n_2}B\left(\frac{n_1}{2},\frac{n_2}{2}\right)+\left(1-\frac{n_1}{2}\right)\psi\left(\frac{n_1}{2}\right)-$ $\left(1+\frac{n_2}{2}\right)\psi\left(\frac{n_2}{2}\right)+\frac{n_1+n_2}{2}\psi\left(\frac{n_1+n_2}{2}\right)$	$[0,\infty)$
Gamma	$f(x)=\frac{x^{k-1}\exp(-\frac{x}{\theta})}{\theta^k\Gamma(k)}$	$\ln(\theta\Gamma(k))+(1-k)\psi(k)+k$	$[0,\infty)$
Laplace	$f(x)=\frac{1}{2b}\exp\left(-\frac{\|x-\mu\|}{b}\right)$	$1+\ln(2b)$	$(-\infty,\infty)$
Logistic	$f(x)=\frac{e^{-x}}{(1+e^{-x})^2}$	2	$(-\infty,\infty)$
Lognormal	$f(x)=\frac{1}{\sigma x\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)$	$\mu+\frac{1}{2}\ln(2\pi e\sigma^2)$	$[0,\infty)$
Maxwell–Boltzmann	$f(x)=\frac{1}{a^3}\sqrt{\frac{2}{\pi}}\,x^2\exp\left(-\frac{x^2}{2a^2}\right)$	$\ln(a\sqrt{2\pi})+\gamma_E-\frac{1}{2}$	$[0,\infty)$
Generalized normal	$f(x)=\frac{2\beta^{\frac{\alpha}{2}}}{\Gamma(\frac{\alpha}{2})}x^{\alpha-1}\exp(-\beta x^2)$	$\ln\frac{\Gamma(\alpha/2)}{2\beta^{\frac{1}{2}}}-\frac{\alpha-1}{2}\psi\left(\frac{\alpha}{2}\right)+\frac{\alpha}{2}$	$(-\infty,\infty)$
Pareto	$f(x)=\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$	$\ln\frac{x_m}{\alpha}+1+\frac{1}{\alpha}$	$[x_m,\infty)$
Student's t	$f(x)=\frac{(1+x^2/\nu)^{-\frac{\nu+1}{2}}}{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}$	$\frac{\nu+1}{2}\left(\psi\left(\frac{\nu+1}{2}\right)-\psi\left(\frac{\nu}{2}\right)\right)+\ln\sqrt{\nu}B\left(\frac{1}{2},\frac{\nu}{2}\right)$	$(-\infty,\infty)$
Triangular			$[0,1]$

	$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \le x \le c, \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c < x \le b, \end{cases}$	$\frac{1}{2} + \ln \frac{b-a}{2}$	
Weibull	$f(x) = \frac{k}{\lambda^k} x^{k-1} \exp\left(-\frac{x^k}{\lambda^k}\right)$	$\frac{(k-1)\gamma_E}{k} + \ln \frac{\lambda}{k} + 1$	$[0, \infty)$
Multivariate normal	$f_X(\vec{x}) = \frac{\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top \Sigma^{-1} \cdot (\vec{x}-\vec{\mu})\right)}{(2\pi)^{N/2} \lvert\Sigma\rvert^{1/2}}$	$\frac{1}{2} \ln\{(2\pi e)^N \det(\Sigma)\}$	$\mathbb{R}^N$

Many of the differential entropies are from.[8]:120–122

Variants

As described above, differential entropy does not share all properties of discrete entropy. For example, the differential entropy can be negative; also it is not invariant under continuous coordinate transformations. Edwin Thompson Jaynes showed in fact that the expression above is not the correct limit of the expression for a finite set of probabilities.[9]:181–218

A modification of differential entropy adds an invariant measure factor to correct this, (see limiting density of discrete points). If $m(x)$ is further constrained to be a probability density, the resulting notion is called relative entropy in information theory:

$$D(p||m) = \int p(x) \log \frac{p(x)}{m(x)} \, dx.$$

The definition of differential entropy above can be obtained by partitioning the range of X into bins of length h with associated sample points ih within the bins, for X Riemann integrable. This gives a quantized version of X, defined by $X_h = ih$ if $ih \le X \le (i+1)h$. Then the entropy of $X_h = ih$ is[1]

$$H_h = -\sum_i hf(ih) \log(f(ih)) - \sum hf(ih) \log(h).$$

The first term on the right approximates the differential entropy, while the second term is approximately $-\log(h)$. Note that this procedure suggests that the entropy in the discrete sense of a continuous random variable should be ∞.

See also

- Information entropy
- Self-information
- Entropy estimation

References

1. Cover, Thomas M.; Thomas, Joy A. (1991). *Elements of Information Theory* (https://archive.org/details/elementsofinform0000cove). New York: Wiley. ISBN 0-471-06259-6.
2. Vasicek, Oldrich (1976), "A Test for Normality Based on Sample Entropy", *Journal of the Royal Statistical Society, Series B*, **38** (1), JSTOR 2984828 (https://www.jstor.org/stable/2984828).
3. Gibbs, Josiah Willard (1902). *Elementary Principles in Statistical Mechanics, developed with especial reference to the rational foundation of thermodynamics*. New York: Charles Scribner's Sons.

4. Kraskov, Alexander; Stögbauer, Grassberger (2004). "Estimating mutual information". *Physical Review E*. **60**: 066138. arXiv:cond-mat/0305641 (https://arxiv.org/abs/cond-mat/0305641). Bibcode:2004PhRvE..69f6138K (https://ui.adsabs.harvard.edu/abs/2004PhRvE..69f6138K). doi:10.1103/PhysRevE.69.066138 (https://doi.org/10.1103%2FPhysRevE.69.066138).
5. Fazlollah M. Reza (1994) [1961]. *An Introduction to Information Theory* (https://books.google.com/books?id=RtzpRAiX6OgC&pg=PA8&dq=intitle:%22An+Introduction+to+Information+Theory%22++%22entropy+of+a+simple+source%22&as_brr=0&ei=zP79Ro7UBovqoQK4g_nCCw&sig=j3lPgyYrC3-bvn1Td42TZgTzj0Q). Dover Publications, Inc., New York. ISBN 0-486-68210-2.
6. "proof of upper bound on differential entropy of f(X)" (https://math.stackexchange.com/q/1745670). *Stack Exchange*. April 16, 2016.
7. Park, Sung Y.; Bera, Anil K. (2009). "Maximum entropy autoregressive conditional heteroskedasticity model" (https://web.archive.org/web/20160307144515/http://wise.xmu.edu.cn/uploadfiles/paper-masterdownload/2009519932327055475115776.pdf) (PDF). *Journal of Econometrics*. Elsevier. Archived from the original (http://www.wise.xmu.edu.cn/Master/Download/..%5C..%5CUploadFiles%5Cpaper-masterdownload%5C2009519932327055475115776.pdf) (PDF) on 2016-03-07. Retrieved 2011-06-02.
8. Lazo, A. and P. Rathie (1978). "On the entropy of continuous probability distributions". *IEEE Transactions on Information Theory*. **24** (1): 120–122. doi:10.1109/TIT.1978.1055832 (https://doi.org/10.1109%2FTIT.1978.1055832).
9. Jaynes, E.T. (1963). "Information Theory And Statistical Mechanics" (http://bayes.wustl.edu/etj/articles/brandeis.pdf) (PDF). *Brandeis University Summer Institute Lectures in Theoretical Physics*. **3** (sect. 4b).

External links

- Hazewinkel, Michiel, ed. (2001) [1994], "Differential entropy" (https://www.encyclopediaofmath.org/index.php?title=p/d031890), *Encyclopedia of Mathematics*, Springer Science+Business Media B.V. / Kluwer Academic Publishers, ISBN 978-1-55608-010-4
- "Differential entropy" (https://planetmath.org/?op=getobj&from=objects&id=1915). *PlanetMath*.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Differential_entropy&oldid=948675131"