

ECE 587 / STA 563: Lecture 7 – Differential Entropy

Information Theory
Duke University, Fall 2024

History: Written by Galen Reeves (-2023). Updated by Henry Pfister (2024-).

Last Modified: November 4, 2024

Outline of lecture:

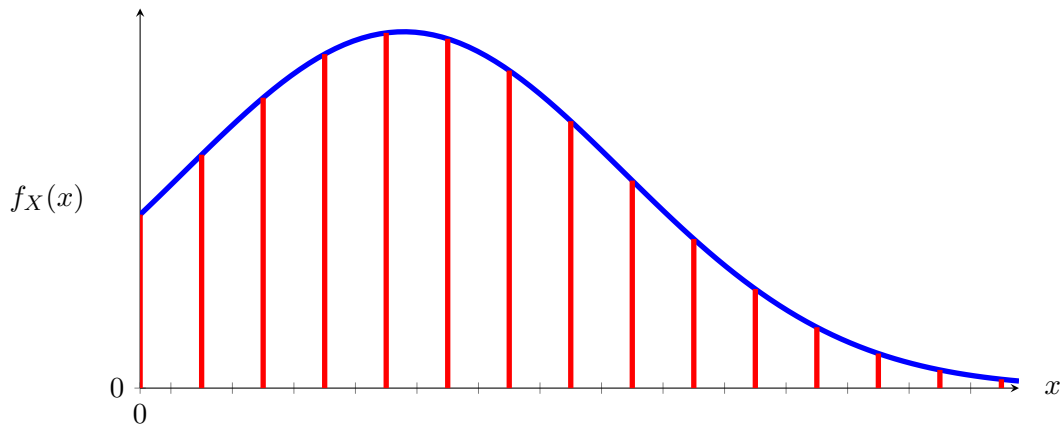
7.1 Entropy of Continuous Variables	1
7.2 Differential Entropy	2
7.3 Properties of Differential Entropy	5
7.4 Entropic Central Limit Theorem	7
7.5 Extension to Abstract Spaces	8

7.1 Entropy of Continuous Variables

- Let X be a continuous real-valued random variable with probability density function (pdf) $f_X(x)$ (often shortened to $f(x)$) defined by

$$\mathbb{P}[X \leq x] = \int_{-\infty}^x f_X(t) dt$$

- Divide range of X into bins of length Δ .



- By mean value theorem, there exists a value x_i in the i th bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x) dx$$

- Consider the quantized random variable X^Δ defined by

$$X^\Delta = x_i \quad \text{if} \quad i\Delta \leq X < (i+1)\Delta$$

- The random variable X^Δ has alphabet $\{x_1, x_2, \dots\}$ and pmf

$$p_{X^\Delta}(x_i) = f(x_i)\Delta$$

- The entropy of the quantized variable X^Δ is

$$\begin{aligned} H(X^\Delta) &= - \sum_i p(x_i) \log p(x_i) \\ &= - \sum_i \Delta f(x_i) \log(f(x_i)\Delta) \\ &= - \sum_i \Delta f(x_i) \log f(x_i) - \sum_i \Delta f(x_i) \log \Delta \\ &= - \sum_i \Delta f(x_i) \log f(x_i) - \log \Delta \end{aligned}$$

- If the function $f_X(x) \log f_X(x)$ is *Riemann integrable* (e.g., this holds by continuity if f_X is continuous and non-negative), then the limit of the first term as Δ becomes small is given by

$$\sum_i \Delta f_X(x_i) \log f_X(x_i) \rightarrow \int f_X(x) \log f_X(x) dx, \quad \text{as } \Delta \rightarrow 0$$

- Thus, for small enough Δ , we have

$$H(X^\Delta) \approx \int f_X(x) \log \left(\frac{1}{f_X(x)} \right) dx + \log \left(\frac{1}{\Delta} \right)$$

- Therefore:

- (1) As $\Delta \rightarrow 0$, the entropy of the quantized version blows up

$$H(X^\Delta) \rightarrow \infty \quad \text{as } \Delta \rightarrow 0$$

This means the entropy of a continuous random variable is *infinite*

- (2) As $\Delta \rightarrow 0$, the difference between the entropy of the quantized version and $\log(1/\Delta)$ satisfies

$$\lim_{\Delta \rightarrow 0} \left(H(X^\Delta) - \log \left(\frac{1}{\Delta} \right) \right) = \int f_X(x) \log \left(\frac{1}{f_X(x)} \right) dx$$

7.2 Differential Entropy

- **Definition:** The *differential entropy* $h(X)$ of a continuous random variable X is

$$h(X) = - \int f_X(x) \log f_X(x) dx = -\mathbb{E}[\log f_X(X)].$$

This is sometimes denoted $h(f)$ and often computed in “nats” by choosing $\log = \ln$.

- **Example:** Uniform distribution:

- The pdf is given by

$$f(x) = 1/a, \quad x \in [0, a]$$

- The differential entropy (in nats) is $h(X) = \int_0^a \frac{1}{a} \ln(a) dx = \ln a$
- Note that for $a < 1$, we have $\ln a < 0$ and so differential entropy can be negative!
- Note that $e^{h(X)} = e^{\log a} = a$ equals the size of the support set.

- **Example:** Normal distribution

- The pdf is given by

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

- The differential entropy (in nats) is

$$\begin{aligned} h(\phi) &= \int_{-\infty}^{\infty} \phi(x) \ln \phi(x) dx \\ &= \mathbb{E}[\ln \phi(X)] \\ &= \mathbb{E}\left[\frac{X^2}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma^2\right] \\ &= \frac{1}{2} \ln e + \frac{1}{2} \ln(2\pi\sigma^2) \\ &= \frac{1}{2} \ln 2\pi e\sigma^2, \quad \text{nats} \end{aligned}$$

- Changing to base 2 gives

$$h(\phi) = \frac{1}{2} \log_2(2\pi e\sigma^2) \quad \text{bits}$$

- By analogy with the discrete case, we can extend all previous definitions to continuous rvs
- The *joint differential entropy* between X and Y is defined by

$$h(X, Y) = \int f_{X,Y}(x, y) \log\left(\frac{1}{f_{X,Y}(x, y)}\right)$$

- The *conditional differential entropy* of X given Y is defined by

$$h(X | Y) = - \int f(x, y) \log f(x | y) dx dy$$

It can also be expressed as

$$h(X|Y) = h(X, Y) - h(Y)$$

- The *Relative entropy* between densities f and g is

$$D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

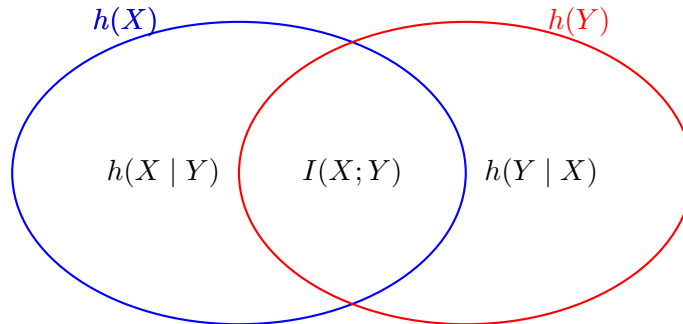
- The *mutual information* between X and Y is

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy$$

- Note that

$$\begin{aligned} I(X; Y) &= h(X) - h(X | Y) \\ &= h(Y) - h(Y | X) \\ &= D(f(x, y) || f(x)f(y)) \end{aligned}$$

- Venn diagram of relationship between mutual information and differential entropy.



- **Example:** (Bivariate Gaussian Distribution) Let $(X, Y) \sim N(0, K)$ be jointly Gaussian with mean zero and covariance K given by

$$K = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

- From the previous example, we know that

$$h(X) = \frac{1}{2} \log(2\pi e \sigma_X^2), \quad h(Y) = \frac{1}{2} \log(2\pi e \sigma_Y^2)$$

- Given $Y = y$, the random variable X has a Gaussian distribution with mean $\mathbb{E}[X|Y = y] = \rho y \sigma_X / \sigma_Y$ and deterministic variance $\text{Var}(X|Y) = (1 - \rho^2) \sigma_X^2$. Thus, we have

$$X|Y = y \sim N\left(\frac{\sigma_X}{\sigma_Y} \rho y, (1 - \rho^2) \sigma_X^2\right)$$

Thus, the conditional entropy is

$$h(X|Y) = \frac{1}{2} \log(2\pi e \sigma_X^2 (1 - \rho^2))$$

- Adding these together yields the joint entropy

$$h(X, Y) = h(X|Y) + h(Y) = \log\left(2\pi e \sqrt{\sigma_X \sigma_Y (1 - \rho^2)}\right)$$

- Taking the difference yields the mutual information

$$I(X; Y) = h(X) - h(X|Y) = -\frac{1}{2} \log(1 - \rho^2) = \frac{1}{2} \log\left(\frac{1}{1 - \rho^2}\right)$$

- Note that if $\rho = \pm 1$ then $X = Y$ and the mutual information is positive infinity!

- **Example:** (Multivariate Gaussian Distribution) Let $X^n \sim N(0, K)$ be an n -dimensional Gaussian vector with mean zero and covariance K . The differential entropy of X is given by

$$h(X^n) = \frac{n}{2} \log\left(2\pi e |K|^{\frac{1}{n}}\right)$$

where $|K|$ denotes the determinant of K . Note that $|K|^{\frac{1}{n}}$ is the *geometric mean* of the eigenvalues of K .

7.3 Properties of Differential Entropy

- **Lemma:** Differential entropy satisfies:

- $h(X + c) = h(X)$
- $h(aX) = h(X) + \log |a|$ for $a \neq 0$.
- $h(AX) = h(X) + \log |\det(A)|$ when A is a square matrix.

- Proof of scaling property for scalar setting.

- The differential entropy of a continuous random variable with density $f_X(x)$ is

$$h(X) = \mathbb{E}[-\log f_X(X)]$$

- For $a > 0$ and $c \in \mathbb{R}$, the cdf of $Y = aX + c$ is given by

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Y \leq y] \\ &= \mathbb{P}[aX + c \leq y] \\ &= F_X((y - c)/a) \end{aligned}$$

and thus the density of Y is

$$f_Y(y) = F'_Y(y) = \frac{d}{dy} F_X((y - c)/a) = \frac{1}{a} f_X((y - c)/a)$$

- As a consequence

$$\begin{aligned} h(aX + c) &= h(Y) \\ &= \mathbb{E}[-\log f_Y(Y)] \\ &= \mathbb{E}\left[-\log\left(\frac{1}{a} f_X((Y - c)/a)\right)\right] \\ &= \mathbb{E}\left[-\log\left(\frac{1}{a} f_X(X)\right)\right] \\ &= \mathbb{E}[-\log f_X(X)] + \log a \\ &= h(X) + \log a \end{aligned}$$

- **Theorem:** (Gaussian distribution maximizes differential entropy under second moment constraints) The differential entropy of an n -dimensional vector X^n with covariance K is upper bounded by the differential entropy of the multivariate Gaussian distribution with the same covariance,

$$h(X^n) \leq \frac{1}{2} \log((2\pi e)^n |K|)$$

Equality holds if and only if $X^n \sim N(0, K)$

- Proof:

- Let Y be Gaussian with $\mathbb{E}[Y] = \mathbb{E}[X] = \mu$, $\text{Cov}(Y) = \text{Cov}(X) = K$, and

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det(K)}} \exp(-(y - \mu)^T K^{-1} (y - \mu)).$$

- Then, the relative entropy (in nats) between f_X and f_Y is given by

$$\begin{aligned}
 D(f_X||f_Y) &= \mathbb{E} \left[\ln \left(\frac{f_X(X)}{f_Y(X)} \right) \right] \\
 &= -h(X) + \mathbb{E} \left[\ln \left(\frac{1}{f_Y(X)} \right) \right] \\
 &= -h(X) + \frac{1}{2} \mathbb{E} [(X - \mu)^T K^{-1} (X - \mu)] + \frac{n}{2} \ln(2\pi|K|^{1/n}) \\
 &= -h(X) + \frac{1}{2} \mathbb{E} [\text{tr}((X - \mu)^T K^{-1} (X - \mu))] + \frac{n}{2} \ln(2\pi|K|^{1/n}) \\
 &= -h(X) + \frac{1}{2} \text{tr}(\mathbb{E}[(X - \mu)(X - \mu)^T] K^{-1}) + \frac{n}{2} \ln(2\pi|K|^{1/n}) \\
 &= -h(X) + \frac{1}{2} \text{tr}(K K^{-1}) + \frac{n}{2} \ln(2\pi|K|^{1/n}) \\
 &= -h(X) + \frac{n}{2} + \frac{n}{2} \ln(2\pi|K|^{1/n}) \\
 &= -h(X) + h(Y),
 \end{aligned}$$

where the trace $\text{tr}(A) = \sum_{i=1}^n A_{i,i}$ of a square matrix $A = BCD$ obeys $\text{tr}(BCD) = \text{tr}(DBC)$.

- Since relative entropy is nonnegative, we conclude that

$$h(X) \leq h(Y)$$

- **Theorem:** If $X \rightarrow Y \rightarrow \hat{X}$ form a Markov chain, then

$$\mathbb{E}[(X - \hat{X})^2] \geq \frac{1}{2\pi e} \exp(2h(X|Y))$$

- **Proof:**

- Conditioned on the event $\{Y = y\}$,

$$\begin{aligned}
 \mathbb{E}[(X - \hat{X})^2 | Y = y] &\geq \text{Var}(X | Y = y) \\
 &\geq \frac{1}{2\pi e} \exp(2h(X | Y = y))
 \end{aligned}$$

where the second inequality follows from the fact that entropy of X conditioned on $Y = y$ is upper bounded by the entropy of Gaussian random variable with the same variance:

$$h(X|Y = y) \leq \frac{1}{2} \log(2\pi e \text{Var}(X|Y = y))$$

- Taking expectation of both sides and applying Jensen's inequality yields the stated result

- **Theorem:** (Entropy Power Inequality) Let X and Y be independent n -dimensional random vectors such that $h(X)$, $h(Y)$ and $h(X + Y)$ exists. Then

$$e^{\frac{2}{n}h(X+Y)} \geq e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(Y)}$$

Moreover, equality holds if and only if X and Y are multivariate Gaussian with proportional covariances.

- The name of the theorem follows from defining the *entropy power* of X to be $\frac{1}{2\pi e} e^{2h(X)/n}$ because this equals the noise power σ^2 when $X \sim N(0, \sigma^2 I)$.

- There are many different proofs of the entropy power inequality, which are interesting in their own right. The following lemma has a simple self-contained proof and implies a special case of the EPI.
- **Lemma:** Let X_1 and X_2 be independent continuous random variables whose distributions are sign invariant (i.e., X_i and $-X_i$ have the same distribution). Then,

$$h\left(\frac{1}{\sqrt{2}}(X_1 + X_2)\right) \geq \frac{1}{2}(h(X_1) + h(X_2))$$

- Proof:

- For any independent random variables X_1 and X_2 , we have

$$\begin{aligned} h(X_1) + h(X_2) &= h(X_1, X_2) \\ &= h\left(\frac{1}{\sqrt{2}}(X_1 + X_2), \frac{1}{\sqrt{2}}(X_1 - X_2)\right) \\ &= h\left(\frac{1}{\sqrt{2}}(X_1 + X_2)\right) + h\left(\frac{1}{\sqrt{2}}(X_1 - X_2)\right) - I\left(\frac{1}{\sqrt{2}}(X_1 + X_2); \frac{1}{\sqrt{2}}(X_1 - X_2)\right) \end{aligned}$$

where the second step holds because the linear transformation applied to the vector (X_1, X_2) has determinant one.

- Because of sign invariance, $(X_1 - X_2)$ and $(X_1 + X_2)$ are equal in distribution and thus $h\left(\frac{1}{\sqrt{2}}(X_1 - X_2)\right) = h\left(\frac{1}{\sqrt{2}}(X_1 + X_2)\right)$. Combining with the above expression and noting that mutual information is non-negative gives the stated result.

- **Special case of EPI:** In the special case, where X_1 and X_2 satisfy the conditions of the lemma and $h(X_1) = h(X_2)$, the following argument establishes the EPI. First apply the map $x \mapsto e^{2x}$ of both sides of the lemma and then use the scaling property of differential entropy to see that

$$e^{h(X_1)+h(X_2)} \leq e^{2h\left(\frac{1}{\sqrt{2}}(X_1+X_2)\right)} = e^{2h(X_1+X_2)+2\ln(1/\sqrt{2})} = \frac{1}{2}e^{2h(X_1+X_2)}.$$

Since $h(X_1) = h(X_2)$, rearranging terms gives $e^{2h(X_1+X_2)} \geq 2e^{h(X_1)+h(X_2)} = e^{2h(X_1)} + e^{2h(X_2)}$.

7.4 Entropic Central Limit Theorem

- Let X_1, X_2, \dots be i.i.d. random variables with mean μ and variance σ^2 and let

$$\begin{aligned} S_n &= \frac{1}{n} \sum_{i=1}^n X_i \\ Z_n &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \end{aligned}$$

denote the average and normalized average of the first n terms.

- The *Law of Large Numbers* (LLN) states that S_n converges almost surely to the mean μ
- The *Central Limit Theorem* (CLT) states that Z_n converges in distribution to Gaussian random variable with mean zero and variance σ^2 . In other words, for all $t \in \mathbb{R}$,

$$\mathbb{P}[Z_n \leq t] \rightarrow \int_{-\infty}^t \frac{1}{\sqrt{2\pi}\sigma^2} e^{-x^2/(2\sigma^2)} dx$$

- Now suppose that the random variables X_1, X_2, \dots are drawn iid from a continuous distribution with finite differential entropy $h(X_i)$. The *entropic CLT* states that the *entropy* of the normalized sum Z_n converges to the entropy of the Gaussian distribution with mean zero and variance σ^2 , i.e.

$$h(Z_n) \rightarrow \frac{1}{2} \log(2\pi e \sigma^2)$$

Furthermore, it has been shown that if X_i is not Gaussian, then $h(Z_n)$ is strictly increasing

$$h(X_1) = h(Z_1) < h(Z_2) < \dots < h(Z_n) < \frac{1}{2} \log(2\pi e \sigma^2).$$

- For the subsequence with $n = 2^m$, we have $Z_{2^m} \stackrel{d}{=} \frac{1}{\sqrt{2}}(Z_{2^{m-1}} + Z'_{2^{m-1}})$ where $Z'_{2^{m-1}}$ is an independent copy of $Z_{2^{m-1}}$ and the proof simplifies because we can apply the lemma recursively to see that $h(Z_{2^m}) \geq h(Z_{2^{m-1}})$. This implies that $h(Z_{2^m}) \uparrow h^*$.
- If $h^* = \frac{1}{2} \log(2\pi e \sigma^2)$, then we have

$$\begin{aligned} D(Z_{2^m} || N(0, \sigma^2)) &= \int f_{Z_{2^m}}(z) \left(\log f_{Z_{2^m}}(z) + \frac{z^2}{2\sigma^2} + \frac{1}{2} \log(2\pi \sigma^2) \right) dz \\ &= \frac{1}{2} \log(2\pi e \sigma^2) - h(Z_{2^m}) \downarrow 0. \end{aligned}$$

This implies that Z_{2^m} converges to $N(0, \sigma^2)$ in total variation and thus in distribution.

- If $h^* \neq \frac{1}{2} \log(2\pi e \sigma^2)$, then since the variance of Z_{2^m} is bounded, we can pass to a subsequence of Z_{2^m} that converges in distribution to some Z^* . From weak convergence along the subsequence and the continuity of convolution, it follows that $(Z^* + Z^{*'})/\sqrt{2} \stackrel{d}{=} Z^*$ and thus $h((Z^* + Z^{*'})/\sqrt{2}) = h(Z^*)$. But, this contradicts the fact that $h(Z_{2^m}) \geq h(Z_{2^{m-1}})$ with equality if and only if $Z_{2^{m-1}}$ is Gaussian.
- A more sophisticated treatment is required, however, to show that $h(Z_n) \rightarrow \frac{1}{2} \log(2\pi e \sigma^2)$ when $n \neq 2^m$.

7.5 Extension to Abstract Spaces

So far, our discussion has focused on random variables where $f_X(x)$ is a continuous function. Kolmogorov extended the definitions of relative entropy (i.e., KL divergence) and mutual information to arbitrary random variables in a particularly nice way. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where the probability measure of $A \subseteq \Omega$ is denoted by $\mathbb{P}[A]$ (and well defined) for all $A \in \mathcal{F}$ where \mathcal{F} is the sigma algebra on Ω of “measurable” sets. A *sigma algebra* on Ω is a collection of subsets of Ω that is closed under complements, countable unions, and countable intersections.

A random variable X is defined by a measurable function $X: \Omega \rightarrow \mathcal{X}$ mapping Ω to the measurable space $(\mathcal{X}, \mathcal{F}_X)$ where \mathcal{F}_X is a sigma algebra on \mathcal{X} . Here, the term *measurable* means that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{F}_X$. In this case, the probability measure of X is defined by $P_X(A) := \mathbb{P}[X^{-1}(A)]$. For a finite measurable partition $\mathcal{A} = \{A_1, \dots, A_n\}$ of \mathcal{X} (i.e., where $X^{-1}(A_1), \dots, X^{-1}(A_n) \in \mathcal{F}$), let $p_i^A = \mathbb{P}[X \in A_i] = P_X(A_i)$. Similarly for another random variable $Z: \Omega \rightarrow \mathcal{X}$, let $q_i^A = \mathbb{P}[Z \in A_i] = P_Z(A_i)$. Then, the relative entropy between the distributions P_X and P_Z is defined to be

$$D(P_X || P_Z) := \sup_{\mathcal{A}} D(p^{\mathcal{A}} || q^{\mathcal{A}}) = \sup_{\mathcal{A}} \sum_{i=1}^{|\mathcal{A}|} p_i^A \log \frac{p_i^A}{q_i^A}.$$

where the supremum runs over all finite measurable partitions of \mathcal{X} . If, for all $A \in \mathcal{F}_X$, the condition $P_X(A) > 0$ implies $P_Z(A) > 0$, then we say that P_X is *absolutely continuous* with respect to P_Z (denoted $P_X \ll P_Z$) and the divergence equals the Lebesgue integral

$$D(P_X || P_Z) = \int_{\mathcal{X}} \log \frac{dP_X}{dP_Z} dP_X.$$

For the mutual information, we define similarly the set \mathcal{Y} , the random variable $Y: \Omega \rightarrow \mathcal{Y}$, and the finite measurable partition \mathcal{B} of \mathcal{Y} . Then, the mutual information between X and Y is defined by

$$I(X; Y) := \sup_{\mathcal{A}, \mathcal{B}} I(X^{\mathcal{A}}, Y^{\mathcal{B}}) = \sup_{\mathcal{A}, \mathcal{B}} \sum_{i=1}^{|\mathcal{A}|} \sum_{j=1}^{|\mathcal{B}|} \mathbb{P}[X \in A_i, Y \in B_j] \log \frac{\mathbb{P}[X \in A_i, Y \in B_j]}{\mathbb{P}[X \in A_i] \mathbb{P}[Y \in B_j]}.$$

where the supremum runs over all finite measurable partitions of \mathcal{X} and \mathcal{Y} . In this case, the joint distribution is given by $P_{XY}(A, B) = \mathbb{P}[X^{-1}(A) \cap Y^{-1}(B)]$ and Lebesgue integral for mutual information is

$$I(X; Y) = \int_{\mathcal{X} \times \mathcal{Y}} \log \frac{dP_{XY}}{d(P_X \times P_Y)} dP_{XY}.$$

Using Kolmogorov's approach, the relative entropy and mutual information can be defined in general using only the discrete definitions of divergence and mutual information.