ECE 587 / STA 563: Lecture 7 – Differential Entropy

Information Theory
Duke University, Fall 2020

Author: Galen Reeves
Last Modified: October 8, 2020

Outline of lecture:

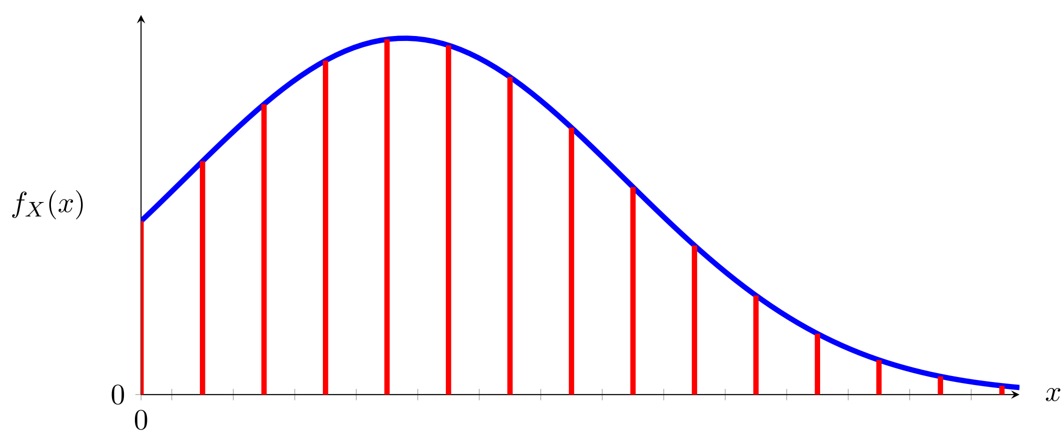
7.1 Entropy of Continuous Variables 1
7.2 Differential Entropy 2
7.3 Properties of Differential Entropy 5
7.4 Entropic Central Limit Theorem 7

7.1 Entropy of Continuous Variables

- Let X be a continuous real-valued random variable with probability density function (pdf) $f_X(x)$ given by

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

- Divide range of X into bins of length Δ.



- By mean value theorem, there exists a valued x_i in the ith 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

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

- The random variable X^Δ has alphabet $\{x_1, x_2, \cdots\}$ 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*, then the limit of the first term as Δ becomes small is given by

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

- Thus, for small 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 $\Delta\to 0$, the entropy of the quantized version blows up

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

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

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

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

7.2 Differential Entropy

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

$$h(X) = -\int f(x)\log f(x)dx$$

  Sometimes denoted $h(f)$.

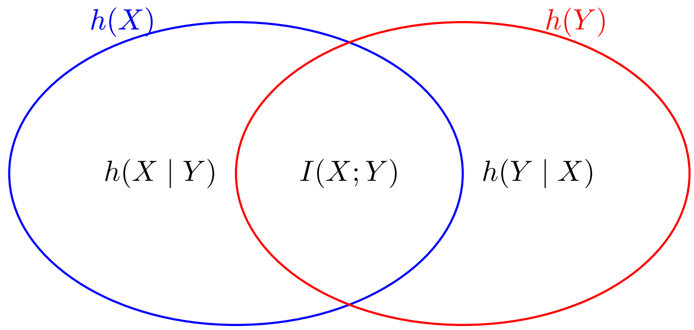
- **Example:** Uniform distribution:
  - The pdf is given by

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

  - The differential entropy is $h(X) = \int_0^a \frac{1}{a}\log(a)dx = \log a$
  - Note that for $a<1$, we have $\log a < 0$ and so differential entropy can be negative!
  - Note that $2^{h(X)} = 2^{\log a} = a$ is 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 measured in nats is
$$\begin{aligned} h(\phi) &= \int_{-\infty}^{\infty} \phi(x) \ln \phi(x) dx \\ &= \mathbb{E}[\ln \phi(X)] \\ &= \mathbb{E}\left[\frac{X^2}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2\right] \\ &= \frac{1}{2}\ln e + \frac{1}{2}\ln(2\pi\sigma^2) \\ &= \frac{1}{2}\ln 2\pi e\sigma^2, \qquad \text{nats} \end{aligned}$$
  - changing the base gives
$$h(\phi) = \frac{1}{2}\log 2\pi e\sigma^2 \qquad \text{bits}$$
  - for $a < 1$, we have $\log a < 0$ and so differential entropy can be negative!
  - note that $2^{h(X)} = 2^{\log a} = a$ is the size of the support set.
- 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
$$h(X \mid Y) = -\int f(x,y) \log f(x \mid y) dx dy$$
  It can also be expressed as
$$h(X|Y) = h(X,Y) - h(Y)$$
- The *Relative entropy* between densities f and g is
$$D(f||q) = \int f(x) \log \frac{f(x)}{g(x)} dx$$
- 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
$$\begin{aligned} I(X;Y) &= h(X) - h(X \mid Y) \\ &= h(Y) - h(Y \mid X) \\ &= D(f(x,y)||f(x)f(y)) \end{aligned}$$

- 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

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

  - From the previous example, we know that

$$h(X) = \frac{1}{2}\log(2\pi e\sigma^2), \qquad h(Y) = \frac{1}{2}\log(2\pi e\sigma^2)$$

  - Conditioned on Y, the random variable X has a Gaussian distribution with mean $\mathbb{E}[X|Y]$ and variance

$$\mathsf{Var}(X|Y) = \mathsf{Var}(X) - \frac{\mathsf{Cov}^2(X,Y)}{\mathsf{Var}(Y)} = (1-\rho^2)\sigma^2$$

    Thus, the conditional entropy is

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

  - Adding these together yields the joint entropy

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

  - 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 $a > 0$, the cdf of $Y = aX$ is given by

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

    and thus the density of Y is

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

  - As a consequence

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

- **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

$$\mathbb{E}[X] = \mathbb{E}[Y], \qquad \mathsf{Cov}(Y) = \mathsf{Cov}(X)$$

◦ The relative entropy between f_X and f_Y obeys

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

◦ Since relative entropy is nonnegative, we conclude that

$$h(X) \le h(Y)$$

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

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

- Proof:

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

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

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:

$$h(X|Y=y) \le \frac{1}{2}\log(2\pi e\,\mathsf{Var}(X|Y=y))$$

◦ 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

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

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

- There are many different proofs of the entropy power inequality, which are interesting in their own right. The following Lemma is a special case of the EPI that has a simple self-contained proof.

- **Lemma:** Let X_1 and X_2 be iid continuous random variables with symmetric distribution (i.e., $X_i = -X_i$ in distribution). Then,

$$h\Big(\tfrac{1}{\sqrt{2}}(X_1+X_2)\Big) \ge \tfrac{1}{2}(h(X_1)+h(X_2))$$

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

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

    where the second step holds because the linear transformation applied to the vector (X_1, X_2) has determinant one.
  - Under the symmetry assumption, we see that $(X_1 - X_2)$ has the same distribution as $(X_1 + X_2)$ and thus $h\Big(\tfrac{1}{\sqrt{2}}(X_1-X_2)\Big) = h\Big(\tfrac{1}{\sqrt{2}}(X_1+X_2)\Big)$. Combining with the above expression and noting that mutual information is non-negative gives the stated result.

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) stats 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 $t \in \mathbb{R}$,

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

- 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.

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

  Furthermore, if $\{X_i\}$ are not Gaussian, then the sequence $h(Z_n)$, is strictly increasing

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