Differential Entropy

Continuous Entropy

Radu Trîmbițaș

UBB

November 2012

Outline I

Entropy

- $H(X) = -\sum_x p(x) \log p(x)$
- All entropic quantities we've encountered have been discrete.
- The world is continuous, channels are continuous, noise is continuous,
- We need a theory of compression, entropy, and channel capacity that applies to such continuous domains.
- We explore this next.

Continuous/Differential Entropy

- X continuous RV, F cdf, f pdf, $S := \{x : f(x) > 0\}$ the *support* set of X

Definition 1

The *differential entropy* $h(X)$ of a continuous RV X with pdf f is

$$h(X) = -\int_S f(x) \log f(x) \mathrm{d}\, x. \tag{1}$$

if the integral exists.

- Since we integrate over only the support set, no worries about log 0.
- $h(X)$ depends on f

Continuous Entropy Of Uniform Distribution

- $X \sim U[0, a]$, $a \in \mathbb{R}_+$

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \mathrm{d}\, x = \log a. \tag{2}$$

- For $a < 1$, $\log a < 0$, $h(X) < 0$
- How can entropy (which we know to mean "uncertainty", or "information") be negative?
- In fact, entropy (as we've seen perhaps once or twice) can be interpreted as the exponent of the "volume" of a typical set.
- Example: $2^{H(X)}$ is the number of things that happen, on average, and can have $2^{H(X)} \ll |\mathcal{X}|$.
- Consider a uniform r.v. Y such that $2^{H(X)} = |\mathcal{Y}|$.
- Thus having a negative exponent just means the volume is small.

Continuous Entropy Of Normal Distribution

- Let
$$X \sim N(0, \sigma^2), \qquad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}.$$
- Entropy in nats
$$\begin{aligned} h(x) &= -\int_{\mathbb{R}} f \ln f = -\int_{-\infty}^{\infty} f(x) \left[-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] dx \\ &= \frac{1}{2} \ln 2\pi\sigma^2 + \frac{1}{2} = \frac{1}{2} \ln 2\pi e\sigma^2 \text{ nats} \cdot \frac{1}{\ln 2} \text{bits/nats} \\ &= \frac{1}{2} \log 2\pi e\sigma^2 \text{ bits}. \end{aligned}$$
- So entropy of a normal is monotonically related to the variance.

AEP I

Theorem 2

$X_1, X_2, \ldots, X_n \sim f(x)$ *i.i.d. RVs*

$$-\frac{1}{n} \log f(X_1, X_2, \ldots, X_n) \xrightarrow{p} E[-\log f(X)] = h(X).$$

Proof.

Follows directly from WLLN. □

AEP II

Definition 3

The *typical set* w.r.t $f(x)$ $A_\varepsilon^{(n)}$ is

$$A_\varepsilon^{(n)} = \left\{ (x_1, \ldots, x_n) \in S^n : \left| -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X) \right| \leq \varepsilon \right\},$$

where $f(x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i)$.

- Thus, we have upper/lower bounds on the probability

$$2^{-n(h+\varepsilon)} \leq f(x_{1:n}) \leq 2^{-n(h-\varepsilon)} \tag{3}$$

- The analog of the cardinality of the typical set for the discrete case is the volume of the typical set for continuous random variables.

AEP III

Definition 4

The *volume* $\mathrm{Vol}(A)$ of a set $A \subset \mathbb{R}^n$ is defined as

$$\mathrm{Vol}(A) = \int_A \mathrm{d}\, x_1 \cdots \mathrm{d}\, x_n.$$

Theorem 5 (Properties of $A_\varepsilon^{(n)}$)

1. $P\left(A_\varepsilon^{(n)}\right) > 1 - \varepsilon$ *for* n *sufficiently large;*
2. $\mathrm{Vol}(A_\varepsilon^{(n)}) \leq 2^{n(h(X)+\varepsilon)}$, $\forall n$;
3. $\mathrm{Vol}(A_\varepsilon^{(n)}) \geq (1-\varepsilon)2^{n(h(X)-\varepsilon)}$, $\forall n$.

AEP Proof

Proof of Theorem 5.

1: By Theorem 2, $-\frac{1}{n}\log f\left(x_{1:n}\right) = -\frac{1}{n}\sum \log f\left(X_i\right) \xrightarrow{p} h(X)$.

2:

$$
\begin{aligned}
1 &= \int_{S^n} f\left(x_1,\ldots,x_n\right)dx_1\cdots dx_n \geq \int_{A_\varepsilon^{(n)}} f\left(x_1,\ldots,x_n\right)dx_1\cdots dx_n \\
&\geq \int_{A_\varepsilon^{(n)}} 2^{-n(h(X)+\varepsilon)}dx_1\cdots dx_n = 2^{-n(h(X)+\varepsilon)}\int_{A_\varepsilon^{(n)}} dx_1\cdots dx_n \\
&= 2^{-n(h(X)+\varepsilon)}\mathrm{Vol}(A_\varepsilon^{(n)}).
\end{aligned}
$$

. . .

AEP Proof

Proof of Theorem 5.

3: Using (3)

$$1-\varepsilon \leq P\left(A_\varepsilon^{(n)}\right) = \int_{A_\varepsilon^{(n)}} f(x_1,\ldots,x_n)dx_1\cdots dx_n$$
$$\leq \int_{A_\varepsilon^{(n)}} 2^{-n(h(X)-\varepsilon)}dx_1\cdots dx_n = 2^{-n(h(X)-\varepsilon)}\mathrm{Vol}(A_\varepsilon^{(n)}).$$

□

Theorem 6

The set $A_\varepsilon^{(n)}$ is the smallest volume set with probability $\geq 1-\varepsilon$, to first order in the exponent.

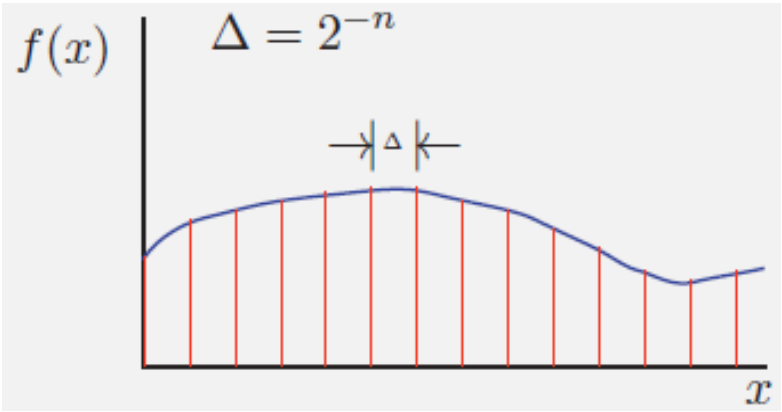
The proof is the same as in the discrete case.

AEP

- Like in the discrete case, $A_\varepsilon^{(n)}$ is the smallest volume that contains, essentially, all of the probability, and that volume is $\approx 2^{nh}$.
- If we look at $(2^{nh})^{1/n}$, we get a "side length" of 2^h.
- So, $-\infty < h < \infty$ is a meaningful range for entropy since it is the exponent of the equivalent side length of the $n - D$ volume.
- Large negative entropy just means small volume.

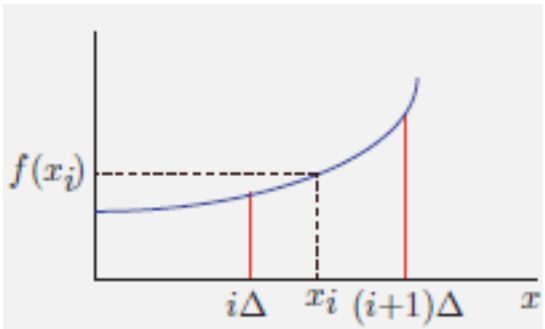
Differential vs. Discrete Entropy

- Let $X \sim f(x)$, and divide the range of X up into bins of length Δ.
- E.g., quantize the range of X using n bits, so that $\Delta = 2^{-n}$.
- We can then view this as follows:



- Mean value theorem, i.e., that if continuous within bin $\exists x_i$ such that

$$f(x_i) = \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x) \mathrm{d}\, x$$



Differential vs. Discrete Entropy I

- Create a quantized random variable X^Δ having those values so that

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

- This gives a discrete distribution

$$P\left(X^\Delta = x_i\right) = p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = \Delta f(x_i)$$

and we can calculate the entropy

$$\begin{aligned} H\left(X^\Delta\right) &= -\sum_{i=-\infty}^{\infty} p_i \log p_i = -\sum_i f(x_i)\Delta \log\left(f(x_i)\Delta\right) \\ &= -\sum_i \Delta f(x_i) \log f(x_i) - \sum_i f(x_i)\Delta \log \Delta \\ &= -\sum_i \Delta f(x_i) \log f(x_i) - \log \Delta \end{aligned}$$

Differential vs. Discrete Entropy II

- This follows since (as expected)

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

- Also, as $\Delta \to 0$, we have $-\log \Delta \to \infty$ and (assuming all is integrable in a Riemannian sense)

$$-\sum_i \Delta f(x_i) \log f(x_i) \to -\int f(x) \log dx$$

- So, $H(X^\Delta) + \log \Delta \to h(f)$ as $\Delta \to 0$.
- Loosely, $h(f) \approx H(X) + \log \Delta$ and for an n-bit quantization with $\Delta = 2^{-n}$, we have

$$H(X^\Delta) \approx h(f) - \log \Delta = h(f) + n.$$

Differential vs. Discrete Entropy III

- This means that as $n \to \infty$, $H(X^\Delta)$ gets larger. Why?
- This makes sense. We start with a continuous random variable X and quantize it at an n-bit accuracy.
- For a discrete representation to represent 2^n values, we expect the entropy to go up with n, and as n gets large so would the entropy, but then adjusted by $h(X)$.
- $H(X^\Delta)$ is the number of bits to describe this n-bit equally spaced quantization of the continuous random variable X.
- $H(X^\Delta) = h(f) + n$ says that it might take either more than n bits to describe X at n-bit accuracy, or less than n bits to describe X at n-bit accuracy.
- If X is very concentrated $h(f) < 0$ then fewer bits. If X is very spread out, then more than n bits.

Examples

1. If $X \sim U[0,1]$ and we let $\Delta = 2^{-n}$, then then $h = 0$, $H(X^{\Delta}) = n$, and n bits suffice to describe X to n bit accuracy.
2. If $X \sim U[0, 1/8]$, the first 3 bits to the right of the decimal point must be 0. To describe X to n-bit accuracy requires only $n-3$ bits, which agrees with $h(X) = -3$.
3. If $X \sim N(0, \sigma^2)$ with $\sigma^2 = 100$, describing X to n bit accuracy would require on the average $n + \frac{1}{2}\log\left(2\pi e \sigma^2\right) = n + 5.37$ bits

- In general, $h(X) + n$ is the number of bits *on the average* required to describe X to n-bit accuracy.
- The differential entropy of a discrete random variable can be considered to be $-\infty$. Note that $2^{-\infty} = 0$, agreeing with the idea that the volume of the support set of a discrete random variable is zero.

Joint and Conditional Differential Entropy

- Like discrete case, we have entropy for vectors of r.v.s
- The joint differential entropy is defined as:

$$h(X_1, \ldots, X_n) = -\int f(x_{1:n}) \log f(x_{1:n}) dx_{1:n}$$

- Conditional differential entropy

$$h(X|Y) = -\int f(x;y) \log f(x|y) dxdy = h(X;Y) - h(Y)$$

Entropy of a Multivariate Gaussian

- When X is distributed according to a multivariate Gaussian distribution, i.e.,

$$X \sim N(\mu, \sigma), \; f(x) = \frac{1}{\left|(2\pi)^n \Sigma\right|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

  then the entropy of X has a nice form, in particular

$$h(X) = \frac{1}{2} \log \left[(2\pi)^n \left|\Sigma\right|\right] \text{ bits}$$

- Notice that the entropy is monotonically related to the determinant of the covariance matrix and is not at all dependent on the mean.
- The determinant is a form of spread, or dispersion of the distribution.

Entropy of a Multivariate Gaussian: Derivation

$$
\begin{aligned}
h(X) &= -\int f(x)\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) - \ln\left(\sqrt{2\pi}\right)^n|\Sigma|^{\frac{1}{2}}\right]dx \\
&= \frac{1}{2}E_f\left[tr\,(x-\mu)^T\Sigma^{-1}(x-\mu)\right] + \frac{1}{2}\ln\left[(2\pi)^n|\Sigma|\right] \\
&= \frac{1}{2}E_f\left[tr\,(x-\mu)^T(x-\mu)\Sigma^{-1}\right] + \frac{1}{2}\ln\left[(2\pi)^n|\Sigma|\right] \\
&= \frac{1}{2}trE_f\left[(x-\mu)^T(x-\mu)\right]\Sigma^{-1} + \frac{1}{2}\ln\left[(2\pi)^n|\Sigma|\right] \\
&= \frac{1}{2}tr\Sigma\Sigma^{-1} + \frac{1}{2}\ln\left[(2\pi)^n|\Sigma|\right] \\
&= \frac{1}{2}trI + \frac{1}{2}\ln\left[(2\pi)^n|\Sigma|\right] = \\
&= \frac{1}{2}n + \frac{1}{2}\ln\left[(2\pi)^n|\Sigma|\right] = \frac{1}{2}\ln\left[(2\pi e)^n|\Sigma|\right]
\end{aligned}
$$

We used $tr\,(ABC) = tr\,(CAB)$.

Relative Entropy and Mutual Information

- The *relative entropy* (or *Kullback-Leibler divergence*) for continuous distributions f and g is

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

- We can, like in the discrete case, use Jensen's inequality to prove $D(f \parallel g) \geq 0$.
- *Mutual Information*:

$$\begin{aligned} I(X;Y) &= D\left(f(x,y) \parallel f(x)f(y)\right) = h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \geq 0. \end{aligned}$$

- Thus, since $I(X;Y) \geq 0$, we have again that conditioning reduces entropy, i.e., $h(Y) \geq h(Y|X)$.

Chain rules and more I

- We still have chain rules

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i | X_{1:i-1}).$$

- And bounds of the form

$$h(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i)$$

- For discrete entropy, we have monotonicity, i.e.,

$$H(X_1, X_2, \ldots, X_k) \leq H(X_1, X_2, \ldots, X_k, X_{k+1}).$$

More generally

$$f(A) = H(X_A)$$

is monotonic non-decreasing in set A (i.e., $f(A) \leq f(B)$; $A \subset B$).

Chain rules and more II

- Is $f(A) = h(X_A)$ monotonic? No, consider Gaussian entropy with diagonal Σ with small diagonal values. So

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

  can get smaller with more random variables.
- Similarly, when some variables independent, adding independent variables with negative entropy can decrease overall entropy.
- Translation does not change entropy

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

-

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

Chain rules and more III

- For vector-valued RVs

$$h(AX) = h(X) + \log|\det(A)|.$$

Example 7 (Mutual information between correlated normal RVs with correlation ρ)

Find $I(X;Y)$, where $(X,Y) \sim N(0,K)$, and

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

Solution.

$h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$ and
$h(X,Y) = \frac{1}{2}\log(2\pi e)^2|K| = \frac{1}{2}\log(2\pi e)^2\sigma^4(1-\rho^2)$, and therefore

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = -\frac{1}{2}\log(1-\rho^2).$$

If $\rho = 0$, X and Y are idependent and $I(X;Y) = 0$. If $\rho = \pm 1$, X and Y are perfectly correlated and the mutual information is infinite. □

Change of Variable

- Let $Y = g(X)$; we wish entropy $h(Y)$
- pdf f_Y

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right|$$

- entropy

$$h(Y) = -E\left(\log f_Y(y)\right) = -E\left(\log(f_X\left(g^{-1}(y)\right)\right) - E\left(\log\left|\frac{dx}{dy}\right|\right)$$
$$= -E\left(\log f_X(x)\right) - E\left(\log\left|\frac{dx}{dy}\right|\right) = h(X) - E\left(\log\left|\frac{dx}{dy}\right|\right)$$

Examples 8

- translation $Y = X + a$, $\frac{dy}{dx} = 1 \Rightarrow h(Y) = h(X)$
- scaling $Y = cX$, $\frac{dy}{dx} = c \Rightarrow h(Y) = h(X) + \log|c|$
- Vector version $Y_{1:n} = AX_{1:n} \Rightarrow h(Y) = h(X) + \log|\det A|$

Hadamard's Inequality

- Using differential entropy, we can sometimes get known results in linear algebra.
- From
$$h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i),$$
consider the case where $X_{1:n}$ is jointly Gaussian $\sim N(0; K)$.
- Then since log is monotonic, we immediately get:
$$|K| \leq \prod_{i=1}^{n} K_{ii},$$
whenever K is positive semi-definite (a result known as Hadamard's inequality).

Moments of of Random Vectors I

- Let $X_{1:n}$ be a continuous random vector.
- The first moment is $\mu = E[X]$.
- The second moment is $C = E[XX^T]$ which is known to be symmetric positive semidefinite.
- Note that $C_{ij} = E[X_i X_j]$.
- There are higher order moments as well, for example, the third order moment has entries of the form $C_{ijk} = E[X_i X_j X_k]$, the forth order moment has entries of the form $C_{ijk\ell} = E[X_i X_j X_k X_\ell]$, and so on.
- Let $C(m)$ be the mth order moment.
- In general, an arbitrary (complex) random vector can have arbitrarily high non-zero mth order moments.

Moments of of Random Vectors II

- It can be shown, moreover, that the multivariate Gaussian only has first and 2nd order moments, but all higher order moments are zero (in fact, the Gaussian can be parameterized exactly via its first and second order moments).
- Now the first and second order moments $C^{(1)}$ (a vector) and $C^{(2)}$ (a matrix) consider all distributions that have these first and second order moments.
- Out of all these distributions, which one would have the highest (differential) entropy?
- Consider what do moments do? They further constrain the possible set of distributions once their value is set.
- Perhaps it then makes sense to intuitively consider that the highest entropy would be granted to the distributions with zero higher order moments.

Moments of of Random Vectors III

- Recall also that the first moment (the mean) doesn't matter for entropy, i.e., $h(X+a) = h(X)$ for any constant a.
- In fact, we have:

Theorem 9

A Gaussian has the maximum entropy over all distributions that have the same first and second moments. That is let $X \in \mathbb{R}^n$ be a vector random variable with $E(X) = 0$ and $E\left(XX^T\right) = K$. Then

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

with equality iff $X \sim N(0, K)$.

Proof of Theorem 9.

- Let $g(X)$ be such that $\int g(x)XX^T dx = K$ (the covariance matrix).
- $X \sim N(0, K)$, $\eta(x)$ its density $\Rightarrow \int \eta(X)XX^T dx = K$.
- But $\log \eta(x)$ has a quadratic form, i.e,

$$\log \eta(x) = -\frac{1}{2}x^T K^{-1}x - \frac{1}{2}\ln\left[(2\pi)^n |K|\right]$$

- Thus, since g and η produce the same results for quadratic forms and by the trace trick, we have

$$0 \leq D(g \parallel \eta) = \int g(x) \log \frac{g(x)}{\eta(x)} dx = -h(g) - \int g(x) \log \eta(x) dx$$
$$= -h(g) - \int \eta(x) \log \eta(x) dx = -h(g) + h(\eta)$$

. . .

Gaussians and Maximum Entropy

Proof of Theorem 9 continued.

- Thus, we get $h(g) \leq h(\eta)$.
- Finally, recall that the Gaussian achieves this entropy ($g = \eta$, from Jensen's inequality).

□

- An instance of a much more general result about maximum entropy.
- Suppose we have a random variable X and a vector of "feature functions" $f(x)$ and consider distributions that satisfy certain constraints $E_p f(X) = \mu$.
- If, over all such distributions that satisfy the constraints, we maximize the entropy, we get a distribution of the form:

$$p(x)\alpha \exp(\lambda f(x)),$$

where λ is a vector of parameters (Lagrange multipliers)

- If $f(X) = (X_1, \ldots, X_n; \{X_i X_j\}_{ij})$, we get back the Gaussian.

Estimation error and differential entropy I

- The multivariate normal (Gaussian) distribution maximizes the entropy over all distributions with the same variance
- This leads to the estimation counterparts of Fano's inequality
- X RV with differential entropy $h(x)$, $\widehat{X}$ an estimate of X and $E\left(X-\widehat{X}\right)^2$ be the expected prediction error (square mean error).

Theorem 10

For any RV X and any estimator $\widehat{X}$,

$$E\left(X-\widehat{X}\right)^2 \geq \frac{1}{2\pi e} e^{2h(X)},$$

with equality iff X is normal and $\widehat{X} = E(X)$.

Estimation error and differential entropy II

Proof.

We have

$$E\left(X-\widehat{X}\right)^2 \geq \min_{\widehat{X}} E\left(X-\widehat{X}\right)^2 \qquad //\text{mean is the best estimator} \tag{4}$$

$$= E\left(X-E(X)\right)^2 \tag{5}$$

$$= var(X) \geq \frac{1}{2\pi e} e^{2h(X)} \; // \text{ maximum entropy for Gaussian} \tag{6}$$

We have equality in (4) iff $\widehat{X} = E(X)$ and in (6) iff X is normal. □

Estimation error and differential entropy III

Corollary 11

Given side information Y and estimator $\widehat{X}(Y)$, it follows that

$$E\left(X-\widehat{X}(Y)\right)^2 \geq \frac{1}{2\pi e}e^{2h(X|Y)}.$$