Information Measures for Continuous Distributions

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

November 2, 2023

Entropy of a continuous random variable

Question: What is the entropy of a continuous real-valued r.v. X ?

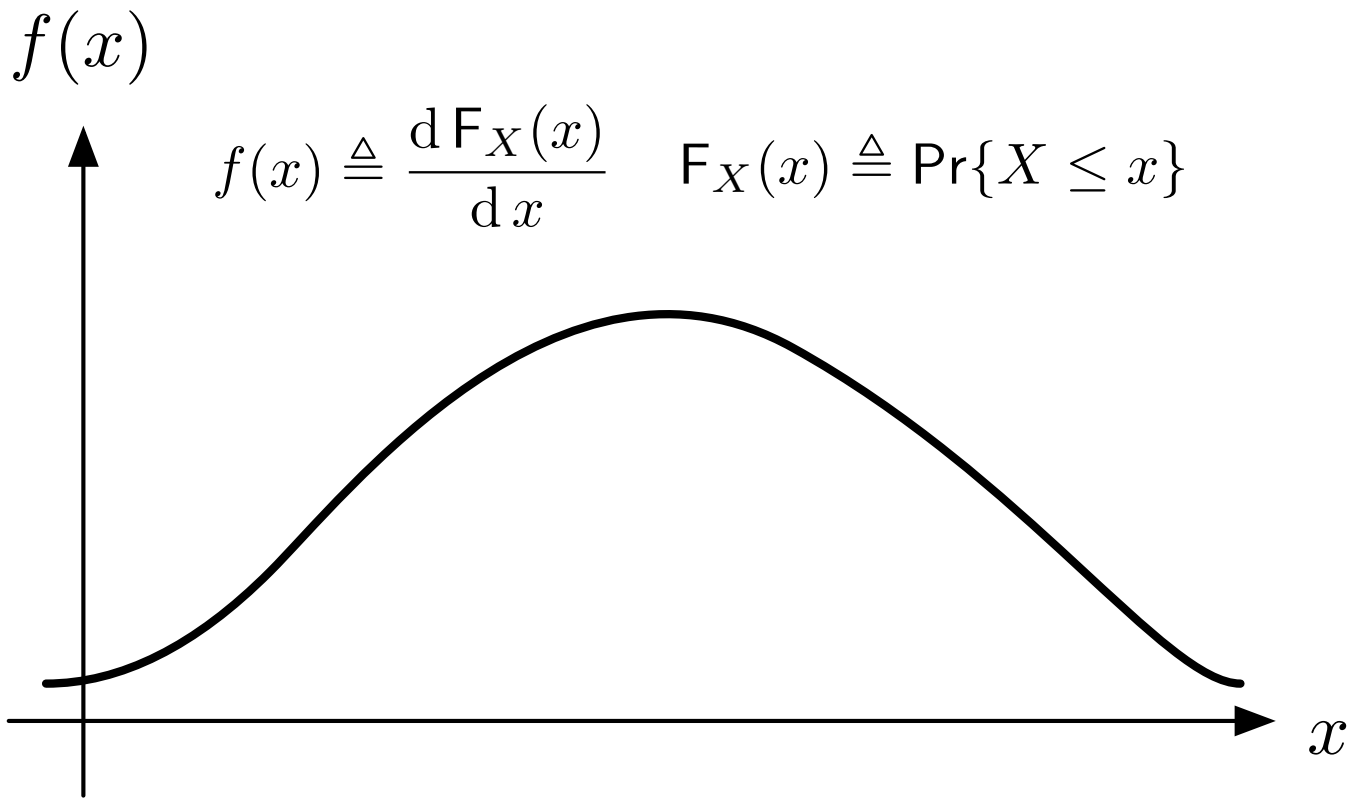
Suppose X has the probability density function (PDF) $\mathsf{f}_X(\cdot)$.

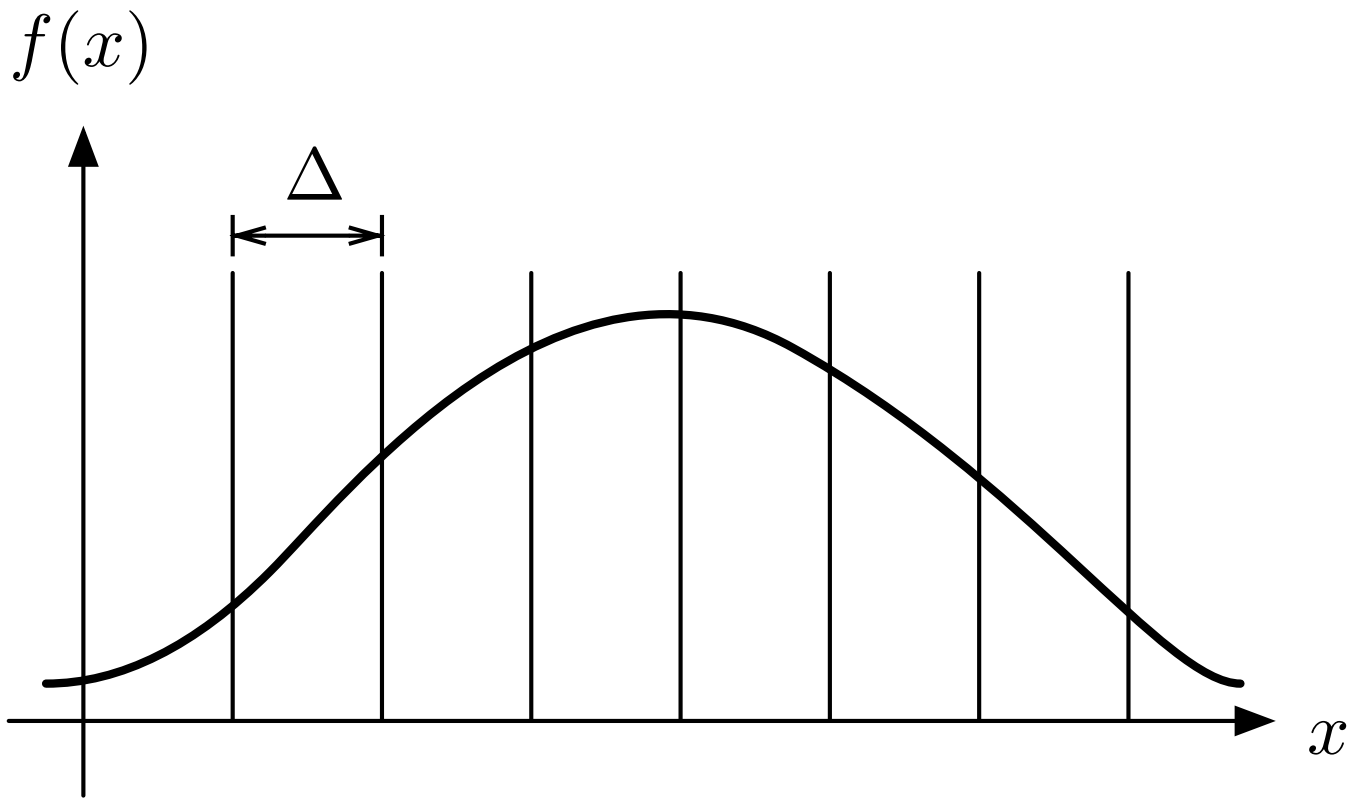
Let us discretize X to answer this question, as follows:

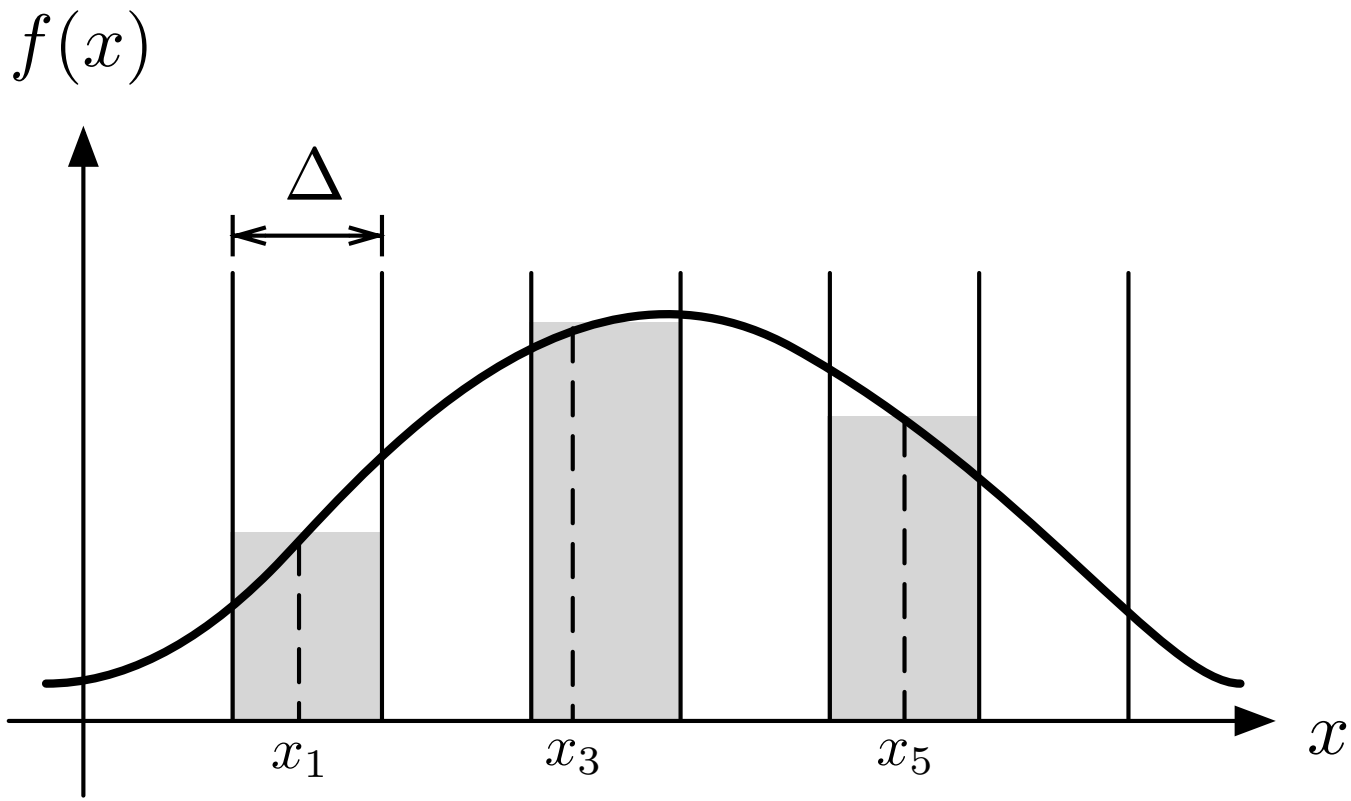
- Partition $\mathbb{R}$ into length-Δ intervals: $\mathbb{R} = \bigcup_{k=-\infty}^{\infty} [k\Delta, (k+1)\Delta)$.
- Suppose that $\mathsf{f}_X(\cdot) \equiv f(\cdot)$ is continuous, then by the mean-value theorem (MVT),

$$\forall\, k \in \mathbb{Z},\ \exists\, x_k \in [k\Delta, (k+1)\Delta) \text{ such that } f(x_k) = \tfrac{1}{\Delta} \int_{k\Delta}^{(k+1)\Delta} f(x)\,\mathrm{d}x.$$

- Set $[X]_\Delta := x_k$ if $X \in [k\Delta, (k+1)\Delta)$, with PMF $\mathsf{P}(x_k) = f(x_k)\Delta$.


$f(x)$
$f(x) \triangleq \frac{\mathrm{d}\,\mathsf{F}_X(x)}{\mathrm{d}\,x}$
$\mathsf{F}_X(x) \triangleq \Pr\{X \le x\}$
x


$f(x)$
Δ
x


$f(x)$
Δ
x_1
x_3
x_5
x

Observation: $\lim_{\Delta\to 0} \mathrm{H}([X]_\Delta) = \mathrm{H}(X)$ (intuitively), while

$$\begin{aligned}
\mathrm{H}([X]_\Delta) &= -\sum_{k=-\infty}^{\infty} (f(x_k)\Delta)\log(f(x_k)\Delta) \\
&= -\Delta \sum_{k=-\infty}^{\infty} f(x_k)\log f(x_k) - \log\Delta \\
&\to -\int_{\infty}^{\infty} f(x)\log f(x)\,\mathrm{d}x + \infty = \infty \qquad \text{as } \Delta \to 0
\end{aligned}$$

Hence, $\mathrm{H}(X) = \infty$ if $-\int_{\infty}^{\infty} f(x)\log f(x)\,\mathrm{d}x = \mathsf{E}\left[\log \frac{1}{f(X)}\right]$ exists.

It is quite intuitive that the entropy of a continuous random variable can be arbitrarily large, because it can take infinitely many possible values.

The term $\mathsf{E}_{X\sim \mathrm{f}_X}[\log 1/\mathrm{f}_X(X)]$ shares a similar form as the Shannon entropy $\mathsf{E}_{X\sim \mathrm{p}_X}[\log 1/\mathrm{p}_X(X)]$, and it is called the *differential entropy* of the density f_X (or the continuous random variable X).

Differential entropy

Definition 1 (Differential Entropy)

The differential entropy of a continuous r.v. X with PDF f_X is defined as

$$\mathrm{h}(X) := \mathsf{E}_{X\sim \mathsf{f}_X}\left[\log \frac{1}{\mathsf{f}_X(X)}\right]$$

if the (improper) integral exists.

The previous quantization argument implies that the Shannon entropy of the quantized version of the continuous r.v. X to n-bit precision ($\Delta = 2^{-n}$) is roughly $n + \mathrm{h}(X)$ bits.

In words, $\mathrm{h}(X)$ is the *extra* number of bits on the average required to describe X to n-bit precision. It can be positive, zero, or negative.

Example 1 (Differential entropy of a uniform r.v.)

For a r.v. $X \sim \mathrm{Unif}([a,b])$, that is, its PDF $\mathsf{f}_X(x) = \frac{1}{b-a}\mathbb{1}\{a \le x \le b\}$, its differential entropy

$$\mathrm{h}(X) = \log(b-a).$$

Interpretation via quantization:

- Suppose $X \sim \mathrm{Unif}([0,1])$. $\mathrm{h}(X) = 0$. To describe X to n-bit precision, n bits are needed.
- Suppose $X \sim \mathrm{Unif}([0,2^{-m}])$, $m > 0$. $\mathrm{h}(X) = -m$. To describe X to n-bit precision, just $n-m$ bits are needed.
- Suppose $X \sim \mathrm{Unif}([0,2^{m}])$, $m > 0$. $\mathrm{h}(X) = m$. To describe X to n-bit precision, an extra m bits are needed.

Remark: if $b-a$ approaches 0, X becomes deterministic, and its differential entropy becomes $-\infty$.

Example 2 (Differential entropy of $\mathrm{N}(0,1)$)

For a r.v. $X \sim \mathrm{N}(0,1)$, that is, its PDF $\mathsf{f}_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, its differential entropy

$$\mathrm{h}(X) = \frac{1}{2} \log(2\pi e).$$

Conditional differential entropy

Suppose X is jointly distributed with another random variable Y, and given $Y = y$, X is still a continuous r.v. with density $\mathsf{f}_{X|Y}(\cdot|y)$. The differential entropy of this density is

$$\mathrm{h}(X|Y=y) = \int_{-\infty}^{\infty} \mathsf{f}_{X|Y}(x|y) \log \frac{1}{\mathsf{f}_{X|Y}(x|y)} \, \mathrm{d}x.$$

Averaging it over Y, the conditional differential entropy emerges:

$$\mathrm{h}(X|Y) = \begin{cases} \sum_{y \in \mathcal{Y}} \mathsf{p}_Y(y) \mathrm{h}(X|Y=y), & \text{if } Y \text{ is discrete with PMF } \mathsf{p}_Y. \\ \int_{-\infty}^{\infty} \mathsf{f}_Y(y) \mathrm{h}(X|Y=y) \, \mathrm{d}y, & \text{if } Y \text{ is continuous with PDF } \mathsf{f}_Y. \end{cases}$$

$$= \mathsf{E}_{X,Y}\left[\log \frac{1}{\mathsf{f}_{X|Y}(X|Y)}\right].$$

Information divergence

Recall that the information divergence of a distribution with density $\mathsf{f}_1(\cdot)$ from another distribution with density $\mathsf{f}_2(\cdot)$ is just

$$\mathrm{D}(\mathsf{f}_1 \| \mathsf{f}_2) := \mathsf{E}_{X\sim \mathsf{f}_1}\left[\log \frac{\mathsf{f}_1(X)}{\mathsf{f}_2(X)}\right] = \int_{x\in \mathsf{supp}_{\mathsf{f}_1}} \mathsf{f}_1(x) \log \frac{\mathsf{f}_1(x)}{\mathsf{f}_2(x)}\, \mathrm{d}x$$

if the (improper) integral exists. Other properties remain to hold such as non-negativity, DPI, chain rule, conditioning increases divergence, etc..

Differential entropy turns out to have a strong connection to KL divergence. In $\mathrm{D}(\mathsf{f}\|\mathsf{g})$, suppose we replace the second PDF g by the "uniform" density $\mathbb{1}\{x \in \mathbb{R}\}$, that is, instead of a probability measure, we use the *Lebesgue measure*. Then,

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \int_{-\infty}^{\infty} \mathsf{f}(x) \log \mathsf{f}(x)\ \mathrm{d}x = -\mathrm{h}(\mathsf{f})\,.$$

Hence, for a probability law that contains a discrete component, its differential entropy is $-\infty$.

Mutual information

The mutual information between two jointly distributed r.v.'s X and Y can be defined via information divergence when they are not discrete:

- If (X, Y) is jointly distributed with PDF $\mathsf{f}_{X,Y}$, then

$$\begin{aligned}\mathrm{I}(X;Y) &= \mathrm{D}(\mathsf{f}_{X,Y} \| \mathsf{f}_X \times \mathsf{f}_Y) \\ &= \mathsf{E}\left[\log \frac{\mathsf{f}_{X,Y}(X,Y)}{\mathsf{f}_X(X)\mathsf{f}_Y(Y)}\right] \\ &= \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) = \mathrm{h}(X) - \mathrm{h}(X|Y) = \mathrm{h}(Y) - \mathrm{h}(Y|X)\,.\end{aligned}$$

- More generally, if $Y|X = x$ has a density for all x, then

$$\mathrm{I}(X;Y) = \mathrm{D}\left(\mathsf{f}_{Y|X} \middle\| \mathsf{f}_Y \middle| \mathsf{P}_X\right) = \mathrm{h}(Y) - \mathrm{h}(Y|X)\,.$$

Properties that extend to continuous r.v.'s

Proposition 1 (Chain rule)

Suppose (X,Y) *is jointly distributed with joint PDF* $\mathsf{f}_{X,Y}$ *and* $\mathrm{h}(X,Y)$ *exists, then* $\mathrm{h}(X,Y)=\mathrm{h}(X)+\mathrm{h}(Y|X)$*. More generally,*

$$\mathrm{h}(X^n)=\sum_{i=1}^n \mathrm{h}\left(X_i\middle|X^{i-1}\right).$$

Proposition 2 (Conditioning reduces differential entropy)

$$\mathrm{h}(X|Y)\le \mathrm{h}(X), \quad \mathrm{h}(X|Y,Z)\le \mathrm{h}(X|Z).$$

New properties of differential entropy

Differential entropy can be negative.

Since $b-a$ can be made arbitrarily small, $\mathrm{h}(X)=\log(b-a)$ can be negative. Hence, the non-negative property of entropy *cannot* be extended to differential entropy.

Scaling will change the differential entropy.

Consider $X\sim\mathrm{Unif}[0,1]$. Then, $2X\sim\mathrm{Unif}[0,2]$. Hence,

$$\mathrm{h}(X)=\log 1=0,\ \mathrm{h}(2X)=\log 2=1\implies \mathrm{h}(X)\neq\mathrm{h}(2X).$$

This is in sharp contrast to entropy: $\mathrm{H}(X)=\mathrm{H}(g(X))$ as long as $g(\cdot)$ is an invertible function.

Scaling and translation

Proposition 3 (Scaling and Translation in the Scaler Case)

Let X be a continuous random variable with differential entropy $\mathrm{h}(X)$.

- *Translation does not change the differential entropy: For a constant c, $\mathrm{h}(X+c) = \mathrm{h}(X)$.*
- *Scaling shifts the differential entropy: For a constant $a \neq 0$, $\mathrm{h}(aX) = \mathrm{h}(X) + \log|a|$.*

Proposition 4 (Scaling and Translation in the Vector Case)

Let $\boldsymbol{X}$ be a continuous random vector with differential entropy $h(\boldsymbol{X})$.

- *For a constant vector $\boldsymbol{c}$, $\mathrm{h}(\boldsymbol{X}+\boldsymbol{c}) = \mathrm{h}(\boldsymbol{X})$.*
- *For an invertible matrix $\mathbf{a} \in \mathbb{R}^{n\times n}$, $\mathrm{h}(\mathbf{a}\boldsymbol{X}) = \mathrm{h}(\boldsymbol{X}) + \log|\det(\mathbf{a})|$.*

The proof of these propositions are left as exercises (simple calculus).

Differential entropy of Gaussian random vectors

Example 3 (Differential entropy of Gaussian random vectors)

For a n-dim random vector $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{m}, \mathbf{k})$, its differential entropy $\mathrm{h}(\boldsymbol{X}) = \frac{1}{2}\log\left((2\pi e)^n \det(\mathbf{k})\right)$.

sol: For an n-dim random vector $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{m}, \mathbf{k})$, we can rewrite $\boldsymbol{X}$ as

$$\boldsymbol{X} = \mathbf{a}\boldsymbol{W} + \boldsymbol{m},$$

where $\mathbf{a}\mathbf{a}^\intercal = \mathbf{k}$ and $\boldsymbol{W}$ consists of $W_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$, $i = 1, ..., n$.

Hence, by the translation and scaling properties of differential entropy:

$$\begin{aligned}\mathrm{h}(\boldsymbol{X}) &= \mathrm{h}(\boldsymbol{W}) + \log|\det(\mathbf{a})| = \textstyle\sum_{i=1}^n \mathrm{h}(W_i) + \frac{1}{2}\log\det(\mathbf{k}) \\ &= \tfrac{n}{2}\log(2\pi e) + \tfrac{1}{2}\log\det(\mathbf{k}) = \tfrac{1}{2}\log\left((2\pi e)^n \det(\mathbf{k})\right).\end{aligned}$$

Maximum differential entropy

Uniform distribution maximizes entropy for r.v. with finite support.

For differential entropy, the maximization problem needs to be associated with constraints on the distribution. (otherwise, it is simple to make it infinite)

It turns out that, under a **second moment constraint**, zero-mean Gaussian maximizes the differential entropy.

Theorem 1 (Maximum Differential Entropy under Covariance Constraint)

Let $\boldsymbol{X}$ be a random vector with mean $\boldsymbol{m}$ and covariance matrix

$$\mathsf{E}\left[(\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^{\intercal}\right]=\mathbf{k},$$

and $\boldsymbol{X}^{\mathrm{G}}$ be Gaussian with the same covariance $\mathbf{k}$. Then,

$$\mathrm{h}(\boldsymbol{X}) \leq \mathrm{h}\left(\boldsymbol{X}^{\mathrm{G}}\right)=\tfrac{1}{2}\log\left((2\pi e)^{n}\det(\mathbf{k})\right).$$

pf: First, we can assume WLOG that both $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{G}}$ are zero-mean, since translation does not change the differential entropy.

Let the PDF of $\boldsymbol{X}$ be $\mathsf{f}(\boldsymbol{x})$ and the PDF of $\boldsymbol{X}^{\mathrm{G}}$ be $\mathsf{f}^{\mathrm{G}}(\boldsymbol{x})$. Hence,

$$0 \le \mathrm{D}\left(\mathsf{f}\middle\|\mathsf{f}^{\mathrm{G}}\right) = \mathsf{E}\left[\log \mathsf{f}(\boldsymbol{X})\right] - \mathsf{E}\left[\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{X})\right] = -\mathrm{h}(\boldsymbol{X}) - \mathsf{E}_{\boldsymbol{X}\sim\mathsf{f}}\left[\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{X})\right].$$

Note that $\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{x})$ is a quadratic function of $\boldsymbol{x}$. $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{G}}$ have the same second moment. Hence,

$$\begin{aligned}
&\mathsf{E}_{\boldsymbol{X}\sim\mathsf{f}}\left[\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{X})\right] = \mathsf{E}_{\boldsymbol{X}\sim\mathsf{f}^{\mathrm{G}}}\left[\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{X})\right] = -\mathrm{h}\left(\boldsymbol{X}^{\mathrm{G}}\right) \\
\Longrightarrow\ & 0 \le \mathrm{D}\left(\mathsf{f}\middle\|\mathsf{f}^{\mathrm{G}}\right) = -\mathrm{h}(\boldsymbol{X}) + \mathrm{h}\left(\boldsymbol{X}^{\mathrm{G}}\right) \\
\Longrightarrow\ & \mathrm{h}(\boldsymbol{X}) \le \mathrm{h}\left(\boldsymbol{X}^{\mathrm{G}}\right).
\end{aligned}$$

□

In general, maximum (differential) entropy problems subject to moment constraints can be solved via the non-negativity of KL divergence.

Summary

- Mutual information between two continuous r.v.'s X and Y with joint density $\mathsf{f}_{X,Y}$: $\mathrm{I}(X;Y) = \mathsf{E}\left[\log \frac{\mathsf{f}_{X,Y}(X,Y)}{\mathsf{f}_X(X)\mathsf{f}_Y(Y)}\right]$.
- Differential entropy and conditional differential entropy: $\mathrm{h}(X) := \mathsf{E}\left[\log \frac{1}{\mathsf{f}_X(X)}\right]$, $\mathrm{h}(X|Y) := \mathsf{E}\left[\log \frac{1}{\mathsf{f}_{X|Y}(X|Y)}\right]$.
- $\mathrm{I}(X;Y) = \mathrm{h}(X) - \mathrm{h}(X|Y) = \mathrm{h}(Y) - \mathrm{h}(Y|X)$.
- Chain rule, conditioning reduces differential entropy, non-negativity of mutual information and KL divergence: remain to hold.
- Differential entropy can be negative; $\mathrm{h}(X) \not\leq \mathrm{h}(X,Y)$.
- Translating a r.v. does not change the differential entropy, while scaling it shifts its differential entropy.
- Under a second moment constraint, zero-mean Gaussian maximizes the differential entropy.