# Information and Communication Theory

## Lecture 4

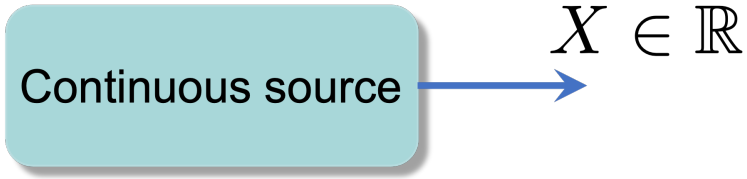
# Continuous Sources

Mário A. T. Figueiredo

DEEC, Instituto Superior Técnico, University of Lisbon, **Portugal**

2023

# Continuous Source




- Probability density function (pdf):

$$f_X(x) \geq 0, \text{ for any } x \in \mathbb{R}; \qquad \int_{\mathbb{R}} f_X(x)\, dx = 1$$

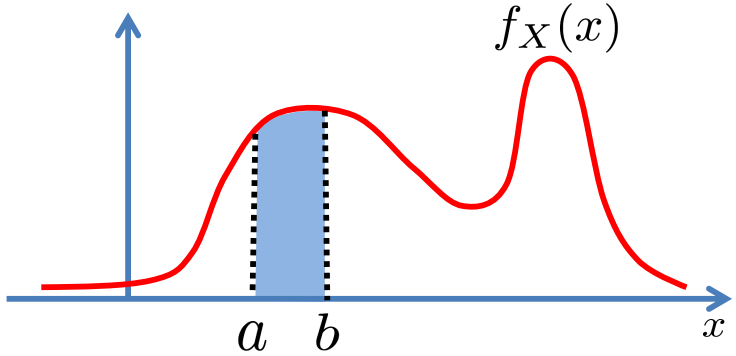
- Probability:

$$\mathbb{P}(X \in [a, b]) = \int_a^b f_X(x)\, dx$$

- Expected value:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\, f_X(x)\, dx$$

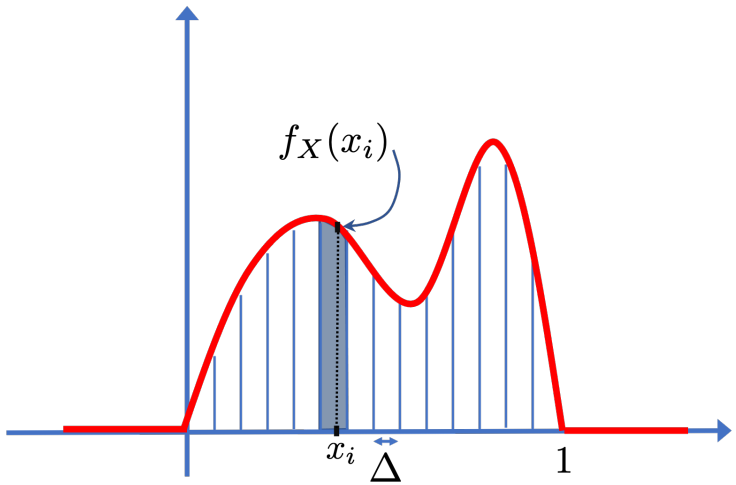
$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\, f_X(x)\, dx$$



# Motivating Example

- Source $X \in [0, 1]$ with some pdf $f_X(x)$
- Discretize $X$ with resolution $\Delta$ to obtain $X^\Delta$

- $\mathbb{P}(X^\Delta = x_i) = \Delta f_X(x_i)$
- $X^\Delta \in \{x_1, ..., x_N\}, \quad N = 1/\Delta$
- $H(X^\Delta) = -\sum_{i=1}^{N} \Delta f_X(x_i) \log\big(\Delta f_X(x_i)\big)$




- $H(X^\Delta) = \underbrace{-\sum_{i=1}^{N} \Delta f_X(x_i) \log\big(f_X(x_i)\big)}_{\stackrel{\Delta\to 0}{\longrightarrow} -\int_0^1 f_X(x) \log f_X(x)\, dx \equiv h(X)} - \log \Delta \;\simeq\; h(X) + \log N$

- Differential entropy:

$$h(X) = -\int_{\mathbb{R}} f_X(x) \log f_X(x) \;=\; \mathbb{E}\big[-\log f_X(X)\big]$$

# Differential Entropy

- Differential entropy:

$$h(X) = -\int_{\mathbb{R}} f_X(x) \log f_X(x) \;=\; \mathbb{E}\big[-\log f_X(X)\big]$$

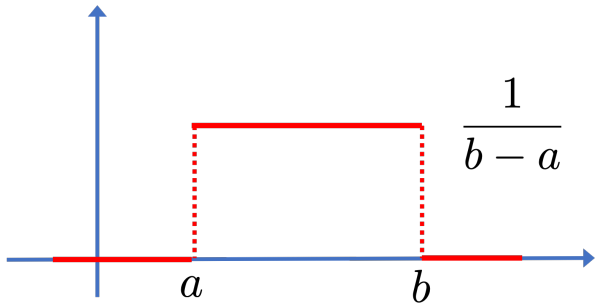
- Qualitatively different from the discrete entropy: it can be negative.

- Example: $X \in [a,\, b]$, with uniform pdf:

$$f_X(x) = 1/(b-a), \text{ if } x \in [a,b];\; 0, \text{otherwise}$$

- Differential entropy (recall $0 \log 0 \equiv 0$):

$$\begin{aligned} h(X) &= -\int_a^b \frac{1}{b-a} \log \frac{1}{a-b}\, dx \\ &= \log(b-a) \end{aligned}$$




- For $b - a < 1$, we have $h(X) < 0$.

# Back to the Example

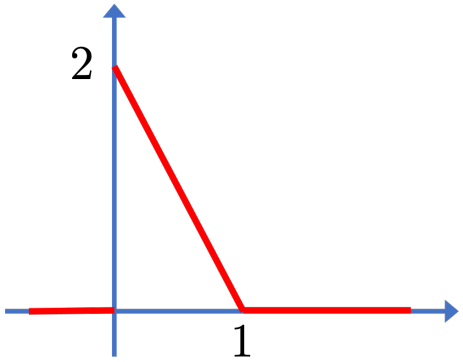
- Source $X \in [0,\, a]$ with uniform pdf.
- Differential entropy: $h(X) = \log a$
- Discretize $X$ with resolution $\Delta$ to obtain $X^\Delta$ ($\Delta \ll a$)
- Expected code length for $X^\Delta$: $L\big(C^{\text{optimal}}\big) \simeq H(X^\Delta)$
- Let $N = a/\Delta$, the number of levels. For $\Delta \ll a$,

$$H(X^\Delta) \simeq h(X) - \log \Delta = \log a - \log \tfrac{a}{N} = \log N.$$

...as expected: we need $\sim \log N$ bits to encode $N$ equiprobable symbols.

- In this case, $h(X)$ is the logarithm of the support size.

# Another Example




- Source $X \in [0,\, 1]$ with pdf $f_X(x) = 2 - 2x$
- Differential entropy: $h(X) \simeq -0.28$
- Discretize $X$ with resolution $\Delta$ to obtain $X^\Delta$ $(\Delta \ll 1)$
- Expected code length for $X^\Delta$: $L\big(C^{\text{optimal}}\big) \simeq H(X^\Delta)$
- Let $N = 1/\Delta$, the number of levels. For $\Delta \ll 1$,

$$H(X^\Delta) \simeq h(X) - \log \Delta \simeq \log N - 0.28 \text{ bits/symbol}$$

...non-uniform probabilities, thus shorter expected code length.

- In this case, $h(X) < 0$ because the density is not uniform.

# Main Properties of Differential Entropy

- Source $X \in \mathbb{R}$ with pdf $f_X(x)$.
- For $Y = X + a$, we have $f_Y(y) = f_X(y - a)$,

$$\begin{aligned} h(Y) &= -\int_{\mathbb{R}} f_X(y-a) \log f_X(y-a)\, dy \\ &= -\int_{\mathbb{R}} f_X(x) \log f_X(x)\, dx \qquad \text{(change of variable } y - a = x\text{)} \\ &= h(X) \end{aligned}$$

- For $Z = cX$, we have $f_Z(z) = \frac{1}{|c|} f_X(\frac{z}{c})$,

$$\begin{aligned} h(Z) &= -\int_{\mathbb{R}} \tfrac{1}{|c|} f_X(\tfrac{z}{c}) \log\Big(\tfrac{1}{|c|} f_X(\tfrac{z}{c})\Big)\, dz \\ &= \log|c| - \int_{\mathbb{R}} f_X(x) \log f_X(x)\, dx \qquad \text{(change of variable } \tfrac{z}{c} = x\text{)} \\ &= \log|c| + h(X) \end{aligned}$$

- Unlike discrete entropy, $h$ is not invariant under injective functions.

# Gaussian Density

- Source $X \in \mathbb{R}$ with pdf $f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Differential entropy:

$$
\begin{aligned}
h(X) &= -\int_{\mathbb{R}} \mathcal{N}(x; \mu, \sigma^2) \log_e \mathcal{N}(x; \mu, \sigma^2)\, dx \\
&= \frac{1}{2\sigma^2} \underbrace{\int_{\mathbb{R}} \mathcal{N}(x; \mu, \sigma^2)(x-\mu)^2\, dx}_{\mathbb{E}[(X-\mu)^2]=\sigma^2} + \log\left(\sqrt{2\pi\sigma^2}\right) \underbrace{\int_{\mathbb{R}} \mathcal{N}(x; \mu, \sigma^2)\, dx}_{1} \\
&= \frac{1}{2} \log\left(2\pi e \sigma^2\right)
\end{aligned}
$$

- As expected (as $h(X + \mu) = h(X)$), the entropy only depends on $\sigma^2$.

# Joint Differential Entropies

- Two sources (random variables) $X, Y \in \mathbb{R}$, with joint pdf $f_{X,Y}(x, y)$

- Joint differential entropy:

$$\begin{aligned} h(X,Y) &= -\iint_{\mathbb{R}^2} f_{X,Y}(x,y) \log f_{X,Y}(x,y)\, dx\, dy \\ &= \mathbb{E}_{X,Y}[-\log f_{X,Y}(X,Y)] \end{aligned}$$

- Independent variables: $X \perp\!\!\!\perp Y$, $f_{X,Y}(x,y) = f_X(x) f_Y(y)$,

$$\begin{aligned} h(X,Y) &= \mathbb{E}_{X,Y}[-\log f_{X,Y}(X,Y)] \\ &= \mathbb{E}_{X,Y}[-\log f_X(X)] + \mathbb{E}_{X,Y}[-\log f_Y(Y)] \\ &= \mathbb{E}_X[-\log f_X(X)] + \mathbb{E}_Y[-\log f_Y(Y)] \\ &= h(X) + h(Y) \end{aligned}$$

...as in the discrete case.

# Conditional Differential Entropy

- Conditional pdf: $f_{Y|X}(y|x) = \dfrac{f_{X,Y}(x,y)}{f_X(x)}$, for $f_X(x) > 0$.
- Naturally, $f_{Y|X}(y|x) \geq 0$ and $\int_{\mathbb{R}} f_{Y|X}(y|x)\, dy = 1$, for any $x$
- Conditioned differential entropy (for a given $x$): $h(Y|X=x)$

$$\begin{aligned} h(Y|X=x) &= -\mathbb{E}\big[\log f_{Y|X}(Y|x)|X=x\big] \\ &= -\int_{\mathbb{R}} f_{Y|X}(y|x) \log f_{Y|X}(y|x)\, dy \end{aligned}$$

...differential entropy of $Y$, given that $X = x$.

- Conditional entropy: expectation (in $X$) of the conditioned entropy:

$$h(Y|X) = \int_{\mathbb{R}} f_X(x)\, h(Y|X=x)\, dx = -\mathbb{E}_{X,Y}[\log f_{Y|X}(Y|X)]$$

# Conditional Differential Entropy

- Conditional entropy: expectation (in $X$) of the conditioned entropy:

$$h(Y|X) = \int_{\mathbb{R}} f_X(x)\, h(Y|X=x)\, dx = -\mathbb{E}_{X,Y}[\log f_{Y|X}(Y|X)]$$

- Exercise: what if $Y = g(X)$, where $g$ is a deterministic function?
  - If $Y = g(X)$, what is $f_{Y|X}(y|x)$? A delta function at $g(x)$

$$f_{Y|X}(y|x) = \delta(y - g(x))$$

  - A delta function has support on a single point: $h(Y|X=x) = -\infty$
  - Conclusion: if $Y = g(X)$, then $h(Y|X) = -\infty$.
  - Very different from the discrete case, where $H(\phi(X)|X) = 0$.

# Bayes for Differential Entropies

- From the conditional pdf definition: $f_{X,Y}(x,y) = f_{X|Y}(x|y)\, f_Y(y)$.
- Joint differential entropy, $h(X,Y) = -\mathbb{E}_{X,Y}\big[\log f_{X,Y}(X,Y)\big]$,

$$\begin{aligned} h(X,Y) &= -\mathbb{E}_{X,Y}\big[\log f_{X,Y}(X,Y)\big] \\ &= -\mathbb{E}_{X,Y}\big[\log f_{X|Y}(X|Y) + \log f_Y(Y)\big] \\ &= -\mathbb{E}_{X,Y}\big[\log f_{X|Y}(X|Y)\big] - \mathbb{E}_{X,Y}\big[\log f_Y(Y)\big] \\ &= h(X|Y) + h(Y) \end{aligned}$$

- By symmetry, $h(X,Y) = h(Y|X) + h(X)$.
- Bayes for differential entropies:

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

- Independent variables: if $X \perp\!\!\!\perp Y$,

$$h(X,Y) = h(X) + h(Y) \;\Rightarrow\; h(X|Y) = h(X) \text{ and } h(Y|X) = h(Y)$$

# Mutual Information

- Recall that $h(X,Y) = h(Y|X) + h(X) = h(X|Y) + h(Y)$.
- Consequently

$$h(X) - h(X|Y) = h(Y) - h(Y|X) \equiv I(X;Y)$$

...named mutual information (MI) (why not **differential** MI?)

- Independent variables: if $X \perp\!\!\!\perp Y$,

$$h(X|Y) = h(X) \Rightarrow I(X;Y) = 0$$

- Deterministically dependent variables:

$$\begin{aligned} Y = g(X) &\Rightarrow h(Y|X) = -\infty \Rightarrow I(X;Y) = +\infty \\ X = g(Y) &\Rightarrow h(X|Y) = -\infty \Rightarrow I(X;Y) = +\infty \end{aligned}$$

# Kullback-Leibler Divergence

- Let $X, X' \in \mathbb{R}$ be two real random variables.
- The Kullback-Leibler divergence (KLD) is defined as

$$D_{\mathsf{KL}}(f_X \parallel f_{X'}) = \int_{\mathbb{R}} f_X(x) \log \frac{f_X(x)}{f_{X'}(x)}\, dx$$
$$= \mathbb{E}_X \left[ \log \frac{f_X(X)}{f_{X'}(X)} \right]$$

- If $f_X(x) = f_{X'}(x)$, for all $x \in \mathcal{X}$, then

$$\log \frac{f_X(x)}{f_{X'}(x)} = 0 \quad \Rightarrow \quad D_{\mathsf{KL}}(f_X \parallel f_{X'}) = 0$$

- In general, $D_{\mathsf{KL}}(f_X \parallel f_{X'}) \neq D_{\mathsf{KL}}(f_{X'} \parallel f_X)$ (not symmetric)

# Kullback-Leibler Divergence

- If, for some $x$, $f_{X'}(x) = 0$ and $f_X(x) > 0$, then

$$D_{\mathsf{KL}}(f_X \parallel f_{X'}) = +\infty$$

- Relationship with the mutual information:

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= \mathbb{E}_X[-\log f_X(X)] + \mathbb{E}_{X,Y}[\log f_{X|Y}(X|Y)] \\ &= \mathbb{E}_{X,Y}[-\log f_X(X) + \log f_{X|Y}(X|Y)] \\ &= \mathbb{E}_{X,Y}[-\log f_X(X) + \log f_{X,Y}(X,Y) - \log f_Y(Y)] \\ &= \mathbb{E}_{X,Y}\left[\log \frac{f_{X,Y}(X,Y)}{f_X(X)\, f_Y(Y)}\right] = D_{\mathsf{KL}}(f_{X,Y} \parallel f_X\, f_Y) \end{aligned}$$

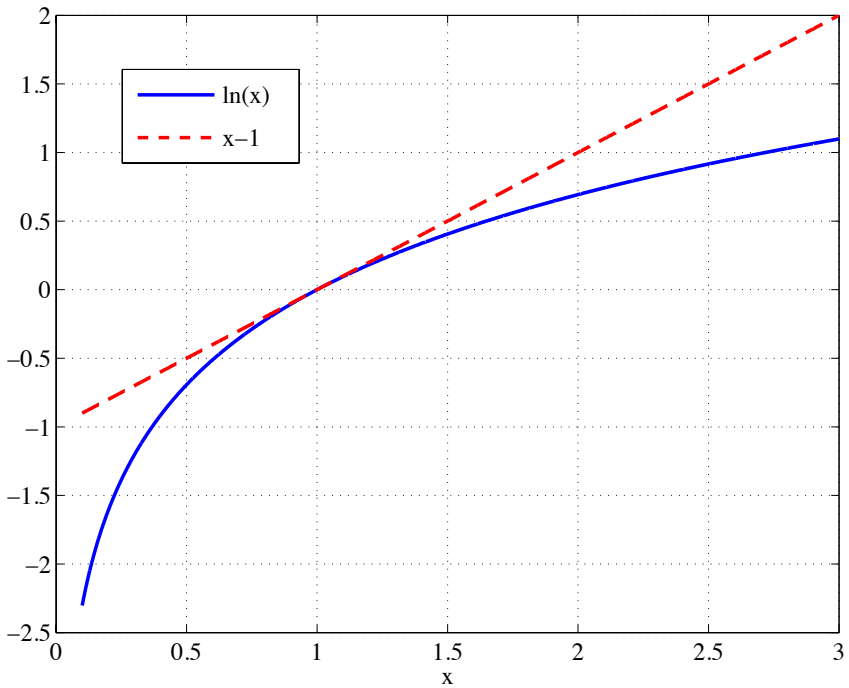
- If $X \perp\!\!\!\perp Y$, then $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$, and $I(X;Y) = 0$.

# An Important Inequality

- Gibbs inequality for the natural logarithm $(\ln \equiv \log_e)$

$$\log_e x \leq x - 1$$

$$\log_e x = x - 1 \quad \Leftrightarrow \quad x = 1$$




- Other bases: $\log_b x = \dfrac{\log_e x}{\log_e b} \leq \dfrac{x-1}{\log_e b} = (x-1)\log_b e$

# Fundamental Inequality of Information Theory

- Fundamental inequality:

$$D_{\mathsf{KL}}(f_X \parallel f_{X'}) \geq 0$$
$$D_{\mathsf{KL}}(f_X \parallel f_{X'}) = 0 \quad \Leftrightarrow \quad f_X = f_{X'} \text{ almost everywhere}$$

- Proof: let $A = \{x \in \mathcal{X} : \ f_X(x) > 0\}$;

$$\begin{aligned} -D_{\mathsf{KL}}(f_X \parallel f_{X'}) &= \int_A f_X(x) \log \frac{f_{X'}(x)}{f_X(x)}\, dx \qquad (0 \log 0 \equiv 0) \\ &\leq \log e \int_A f_X(x) \Big(\frac{f_{X'}(x)}{f_X(x)} - 1\Big)\, dx \quad \text{(previous slide)} \\ &= \log e \Bigg( \underbrace{\int_A f_{X'}(x)\, dx}_{\leq 1} - \underbrace{\int_A f_X(x)\, dx}_{=1} \Bigg) \ \leq \ 0 \end{aligned}$$

...clearly, equality requires $f_{X'}(x) = f_X(x)$, almost everywhere (a.e.)

- Like in the discrete case, $D_{\mathsf{KL}}(f_X \parallel f_{X'}) \geq 0$.

# Corollaries of the Fundamental Inequality

- Non-negativity of the mutual information:

$$I(X;Y) = D_{\mathsf{KL}}(f_{X,Y} \parallel f_X\, f_Y) \geq 0$$
$$I(X;Y) = D_{\mathsf{KL}}(f_{X,Y} \parallel f_X\, f_Y) = 0 \quad \Leftrightarrow \quad f_{X,Y} = f_X\, f_Y, \text{i.e.} \quad X \perp\!\!\!\perp Y$$

- Conditioning reduces entropy; since $I(X;Y) = h(X) - h(X|Y)$,

$$h(X|Y) = h(X) - \overbrace{I(X;Y)}^{\geq 0} \leq h(X)$$
$$h(X|Y) = h(X) \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

- Upper-bound on joint entropy: since $h(X,Y) = h(X|Y) + h(Y)$,

$$h(X,Y) = h(X|Y) + h(Y) \leq h(X) + h(Y)$$
$$h(X,Y) = h(X) + h(Y) \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

# More Corollaries of the Fundamental Inequality

- Upper-bound on joint differential entropy for several variables:

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

$$h(X_1, ..., X_n) = \sum_{i=1}^{n} h(X_i) \Leftrightarrow \text{ all the } X_i \text{ are mutually independent}$$

- Question: in the discrete case, we proved that $H(g(X)) \leq H(X)$; is this true for differential entropy?

- Question: in the discrete case, we proved that, if $X$ and $Y$ are independent variables and $Z = X + Y$, then $H(Z) \geq H(X)$; is this true for differential entropy? Hint: check if $h(Z|X) = h(Y|X)$

# Mixed Continuous-Discrete

- Mixed case: $X$ is continuous and $Y$ is discrete.
- Conditional entropy of $X$ (which is continuous) given $Y$:

$$h(X|Y) = \sum_y h(X|Y=y)\mathbb{P}[Y=y]$$

- Conditional entropy of $Y$ (which is discrete) given $X$:

$$H(Y|X) = \int H(Y|X=x) f_X(x)\, dx$$

- Mutual information:

$$I(X;Y) = h(X) - h(X|Y) = H(Y) - H(Y|X)$$

# More on Mutual Information (MI)

- Unlike the differential entropy, the MI has the same property as in the discrete case: non-negativity.

- Discrete approximation: let $X^\Delta$ be a discretized version of $X \in \mathbb{R}$

- Given another (discrete or continuos) variable $Y$, and for $\Delta \to 0$,

$$I(X^\Delta;Y) = H(X^\Delta) - H(X^\Delta|Y) \simeq h(X) - \log\Delta - \big(h(X|Y) - \log\Delta\big)$$

...thus

$$\lim_{\Delta\to 0} I(X^\Delta, Y) = h(X) - h(X|Y) = I(X;Y)$$

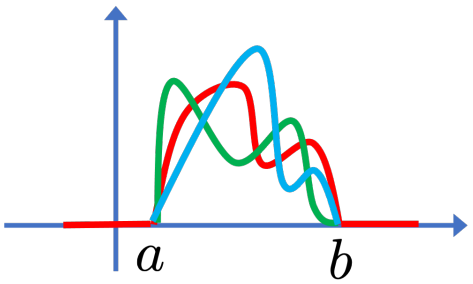
- Unlike the entropy, the (continuous) MI is the limit of the discrete MI.

- Unlike the differential entropy, the MI can be seen as the number of bits of information that the variables have about each other.

# Even More on Mutual Information

- The property in the previous slide suggests the MI is a more fundamental quantity.
- There is a general definition for both discrete and continuous domains.
- Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ (continuous or discrete).
- Let $\mathcal{P}$ and $\mathcal{Q}$ be finite partitions of $\mathcal{X}$ and $\mathcal{Y}$, respectively.
- These partitions induce discrete variables $X^{\mathcal{P}}$ and $Y^{\mathcal{Q}}$.
- Master definition of MI: $I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I\left(X^{\mathcal{P}};Y^{\mathcal{Q}}\right)$
- The discrete and continuous MI can be obtained from this one.

# Maximum Entropy Distributions




- Consider the set of all densities $f_X$ with a given support $[a, b]$
  - Uniform pdf: $U(x; a, b) = 1/(b-a)$, if $x \in [a, b]$, 0, otherwise.
  - From the positivity of the Kullback-Leibler divergence,

$$0 \leq D_{\mathsf{KL}}(f_X \parallel U(\cdot; a, b)) = \int_a^b f_X(x) \log \frac{f_X(x)}{\frac{1}{b-a}}\, dx$$

$$= \underbrace{\int_a^b f_X(x) \log f_X(x)\, dx}_{-h(X)} + \log(b-a) \underbrace{\int_a^b f_X(x)\, dx}_{=1}$$

  - Conclusion: $h(X) \leq \log(b-a)$
    $h(X) = \log(b-a) \Leftrightarrow f_X(x) = U(x; a, b)$.
  - The maximum entropy pdf on a given (finite) support is the uniform.

# Maximum Entropy Distributions

- Consider all densities $f_X$ on $\mathbb{R}$ with variance $\sigma^2$,

$$\int_{\mathbb{R}} \big(x - \mathbb{E}(X)\big)^2 f_X(x)\, dx = \sigma^2$$

- From the positivity of the Kullback-Leibler divergence,

$$0 \leq D_{\mathsf{KL}}(f_X \parallel \mathcal{N}(\cdot; \mathbb{E}(X), \sigma^2)) = \int_a^b f_X(x) \log \frac{f_X(x)}{\mathcal{N}(x; \mathbb{E}(X), \sigma^2)}\, dx$$

$$= \underbrace{\int_{\mathbb{R}} f_X(x) \log f_X(x)\, dx}_{-h(X)} + \tfrac{1}{2\sigma^2} \underbrace{\int_{\mathbb{R}} \big(x - \mathbb{E}(X)\big)^2 f_X(x)\, dx}_{=\sigma^2} + \frac{\log(2\pi\sigma^2)}{2}$$

- Conclusion: $h(X) \leq \frac{1}{2}\log(2\pi e\sigma^2)$
  $h(X) = \frac{1}{2}\log(2\pi e\sigma^2) \Leftrightarrow f_X(x) = \mathcal{N}(x; \mathbb{E}(X), \sigma^2)$.

- The maximum entropy pdf with a given variance is the Gaussian.

# Estimation Error Bound

- Let $X \in \mathbb{R}$, with pdf $f_X$, variance $\sigma^2$, and $\hat{X}$ be an estimator of $X$.
- The minimum mean squared error (MMSE) estimator:

$$\begin{aligned}\hat{X}_{\text{MMSE}} &= \arg\min_{\hat{X}} \mathbb{E}\big[(X-\hat{X})^2\big] \\ &= \arg\min_{\hat{X}} \Big( \underbrace{\mathbb{E}[X^2]}_{\text{indep. of } \hat{X}} - \hat{X}^2 - 2\hat{X}\mathbb{E}[X] \Big) \\ &= \mathbb{E}[X] \qquad \text{(computing derivative and setting to zero)}\end{aligned}$$

- Reverse Gaussian entropy bound $h(X) \leq \frac{1}{2}\log(2\pi e\sigma^2)$:

$$\begin{aligned}\mathbb{E}\big[(X-\hat{X})^2\big] &\geq \min_{\hat{X}} \mathbb{E}\big[(X-\hat{X})^2\big] = \mathbb{E}\big[(X-\mathbb{E}[X])^2\big] = \sigma^2 \\ &\geq \frac{1}{2\pi e} e^{2\,h(X)}\end{aligned}$$

- Equality is achieved if $\hat{X} = \hat{X}_{\text{MMSE}}$ and $f_X$ is Gaussian.

# Estimation Error Bound with Observations

- Let $X \in \mathbb{R}$, with pdf $f_X$, and $Y$ be an observed variable.
- Let $\hat{X}(Y)$ be an estimator of $X$ that is a function of $Y$.
- The minimum mean squared error (MMSE) estimator:

$$\begin{aligned}\hat{X}_{\mathsf{MMSE}}(y) &= \arg\min_{\hat{X}} \mathbb{E}\big[(X-\hat{X})^2 | Y=y\big] \\ &= \arg\min_{\hat{X}} \Big( \underbrace{\mathbb{E}[X^2|Y=y]}_{\text{indep. of } \hat{X}} - \hat{X}^2 - 2\hat{X}\mathbb{E}[X|Y=y] \Big) \\ &= \mathbb{E}[X|Y=y] \qquad \text{(computing derivative and setting to zero)}\end{aligned}$$

- Reverse Gaussian entropy bound $h(X) \le \frac{1}{2}\log(2\pi e\sigma^2)$:

$$\begin{aligned}\mathbb{E}\big[(X-\hat{X})^2|Y=y\big] &\ge \mathbb{E}\big[(X-\mathbb{E}[X|Y=y])^2\big] = \mathsf{var}[X|Y=y] \\ &\ge \frac{1}{2\pi e} e^{2\,h(X|Y=y)}\end{aligned}$$

- Equality achieved if $\hat{X} = \hat{X}_{\mathsf{MMSE}}(y)$ and $f_{X|Y=y}$ is Gaussian, for any $y$.

# Recommended Reading

- T. Cover and J. Thomas, "Elements of Information Theory", John Wiley & Sons, 2006 (Chapter 8).