ECE 5630: Information Theory for Data Transmission, Security and Machine Learning 03/03/20

Lecture 9: Information Measures

Lecturer: Prof. Ziv Goldfeld

Scriber: Hadi AlZayer, *Net ID*: ha366

Assistant Editor: Kia Khezeli

9.1 Shannon Entropy

Definition 9.1 (Shannon Entropy) *Let $\mathcal{X}$ be a countable set and $P \in \mathcal{P}(\mathcal{X})$ with PMF p. The Shannon entropy of $X \sim P$ is*

$$H(X) = H(P) := \mathbb{E}_P\left[\log_2 \frac{1}{p(X)}\right].$$

We henceforth keep the base of the logarithm as 2 and omit it for the simplicity of notation.

Remark 9.1 (Interpretation) *$H(X) = H(P)$ is a measure for the uncertainty or unpredictability of $X \sim P$. One can view $1/\log(p(X))$ as a quantification of surprise from observing X as lower probability outcomes are more surprising. Given this notion, entropy can be interpreted as the expected surprise. Later in this lecture, we will have several examples that will help support this interpretation.*

Remark 9.2 (Shannon Entropy as KL Divergence) *Let $\#$ be the counting measure. Then, $P \ll \#$ and*

$$H(P) = -D_{\mathsf{KL}}\left(P\|\#\right).$$

Furthermore, if $|\mathcal{X}| < \infty$, then $H(P) = \log(|\mathcal{X}|) - D_{\mathsf{KL}}\left(P\|\mathrm{Unif}(\mathcal{X})\right)$.

It follows from Remark 9.2 and non-negativity of KL divergence that for a finite sample space $\mathcal{X}$, the uniform distribution maximizes the Shannon entropy among all $\mathcal{P}(\mathcal{X})$.

Example 9.1

- *Bernoulli:* *Let $X \sim P = \mathsf{Ber}(\alpha)$ where $\alpha \in [0,1]$. Then, $H(X) = H_b(\alpha) := \alpha \log \frac{1}{\alpha} + (1-\alpha)\log\frac{1}{1-\alpha}$, where H_b is called the binary entropy function. Note that the entropy is zero for $\alpha = 0, 1$ and is maximized for $\alpha = \frac{1}{2}$. Note that this matches our interpretation that entropy measures uncertainty as the outcome of a coin toss is the most informative when the coin is fair.*

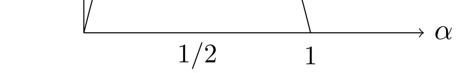

Figure 1: Entropy of a Bernoulli random variable.

- *Uniform: Assume $|\mathcal{X}| < \infty$. Let $P = \mathsf{Unif}(\mathcal{X})$. Then*

$$H(P) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} = \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \log |\mathcal{X}| = \log |\mathcal{X}|.$$

- *Infinite Shannon Entropy: This example demonstrates that Shannon entropy can also be infinite. Let $X \sim P$ with PMF $p(n) = P(X = n) = \frac{c}{n(\log n)^2}$ for all $n \in \{2, 3, \dots\}$ where $c = \sum_{n=2}^{\infty} \frac{1}{n(\log n)^2}$. Then,*

$$H(P) = \sum_{n=2}^{\infty} \frac{c}{n(\log n)^2} \log \left(n(\log n)^2\right) \geq \sum_{n=2}^{\infty} \frac{c}{n \log n} = \infty.$$

Proposition 9.1 (Properties of Shannon Entropy) *The Shannon entropy satisfies:*

1. *Non-negativity: $H(X) \geq 0$ with equality if and only if X is almost surely constant.*
2. *Uniform upper bound: If $|\mathcal{X}| < \infty$, then $H(P) \leq \log |\mathcal{X}|$ with equality if and only if $P = \mathsf{Unif}(\mathcal{X})$*
3. *Invariance to relabling: If f is a bijection, then $H(X) = H(f(X))$.*
4. *Entropy of functions: For any function f, we have $H(X) \geq H(f(X))$ with equality if and only if f is a bijection.*
5. *Concavity: $P \to H(P)$ is concave.*

9.2 Joint Entropy

Definition 9.2 (Joint Entropy) *For a random vector $(X_1, ..., X_n) \sim P_{X_1,...,X_n} \in \mathcal{P}(\mathcal{X}_1 \times ... \times \mathcal{X}_n)$ with PMF $p_{X_1,...,X_n}$, the joint Shannon entropy is $H(X_1, ..., X_n) := \mathbb{E}_{P_{X_1,...,X_n}} \left[\log \frac{1}{p_{X_1,...,X_n}(X_1,...,X_n)}\right]$.*

Definition 9.3 (Conditional Entropy) *Let $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with PMF p_{XY}. The conditional entropy of Y given X is $H(Y|X) := \mathbb{E}_{P_{XY}} \left[\log \frac{1}{P_{Y|X}(Y|X)}\right]$.*

Note that unlike $\mathbb{E}[Y|X]$, which is a random variable, the conditional entropy $H(Y|X)$ is a real number.

Remark 9.3 *Let $P_{Y|X}$ be a transition kernel. For any fixed $x \in \mathcal{X}$, $H(P_{Y|X}(\cdot|x))$ is a regular entropy as given in definition 9.1. We sometimes denote $H(Y|X = x) := H(P_{Y|X}(\cdot|x))$*

Proposition 9.2 (Conditional Entropy) *Let $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Then,*

$$H(Y|X) = \mathbb{E}_{P_X}\left[H(P_{Y|X}(\cdot|X))\right].$$

Proof:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x, y) \log \frac{1}{P_{Y|X}(y|x)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log \frac{1}{P_{Y|X}(y|x)} \\ &= \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log \frac{1}{P_{Y|X}(y|x)} \\ &= \sum_{x \in \mathcal{X}} P_X(x) H(P_{Y|X}(\cdot|x)). \end{aligned}$$

■

Proposition 9.3 (Entropy Properties cont.) *Additional properties of entropy related to conditioning are:*

1. *Conditioning cannot increase entropy:* $H(X) \geq H(X|Y)$ *with equality if and only if* X *and* Y *are independent.*
2. *Chain Rule:*
   - *small chain rule:* $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
   - *full chain rule:* $H(X_1,...,X_n) = H(X_1) + \sum\limits_{i=2}^{n} H(X_i|X_1,...,X_{i-1})$

9.3 Differential Entropy

Differential entropy is the continuous analog of the Shannon entropy.

Definition 9.4 (Differential Entropy) *Let* $P \in \mathcal{P}(\mathbb{R}^d)$ *and* $X \sim P$. *The differential entropy of* $X \sim P$ *is*

$$h(X) = h(P) := -D_{\mathsf{KL}}\left(P \middle\| \mathrm{Leb}(\mathbb{R}^d)\right).$$

If $P \ll \mathrm{Leb}(\mathbb{R}^d)$ *with PDF* p, *then* $h(X) = h(P) = h(p) = \mathbb{E}_P\left[\log\frac{1}{p(X)}\right] = \int_{\mathrm{supp}(P)} p(x)\log\frac{1}{p(x)}\mathsf{d}x$.

Example 9.2

- *Uniform: It is straightforward to show that* $h\big(\mathsf{Unif}([0,a])\big) = \log(a)$. *Note that the differential entropy is negative for* $a < 1$.
- *Gaussian: We have*

$$\begin{aligned} h\big(\mathcal{N}(0,\sigma^2)\big) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \left(\frac{1}{2}\log(2\pi\sigma^2) + \frac{x^2}{2\sigma^2}\right)\mathsf{d}x \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\mathsf{d}x \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} = \frac{1}{2}\log(2\pi e\sigma^2). \end{aligned}$$

  Note that as differential entropy is invariant to labeling, we have $h\big(\mathcal{N}(\mu,\sigma^2)\big) = \frac{1}{2}\log(2\pi e\sigma^2)$ *for all* $\mu \in \mathbb{R}$. *The above argument naturally extends to multivariate Gaussian distributions, showing that* $h\big(\mathcal{N}(m,\Sigma)\big) = \frac{1}{2}\log\big((2\pi e)^d\det(\Sigma)\big)$ *for all* $m \in \mathbb{R}^d$ *and positive definite* $\Sigma \in \mathbb{R}^{d\times d}$.

The definition of $h(X)$ extends to $h(X_1,...,X_m)$ and $h(X|Y)$ similarly to what we saw for Shannon entropy.

Proposition 9.4 (Properties of Differential Entropy) *The differential entropy satisfies the following properties:*

1. *Uniform distribution maximizes h over a bounded domain: If* $\mathrm{supp}(P) \subseteq \mathbb{R}^d$ *is bounded, then* $h(P) \leq h\big(\mathsf{Unif}(\mathrm{supp}(P))\big)$.
2. *Gaussian maximizes h under variance constraint: Let* $\mathcal{X} \subseteq \mathbb{R}^d$ *and consider a positive definite matrix* $\Sigma \in \mathbb{R}^{d\times d}$. *If* $X \sim P$ *has* $\mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right] = \Sigma$, *then* $h(P) \leq h(\mathcal{N}(0,\Sigma))$.
3. *Scaling and shifting: Let* $X \sim P \in \mathcal{P}(\mathbb{R}^d), a \in \mathbb{R}, \mu \in \mathbb{R}^d$, *and* $A \in \mathbb{R}^{d\times d}$ *positive definite. Then,*
   - $h(X + \mu) = h(X)$.
   - $h(aX) = h(X) + d\log|a|$.
   - $h(AX) = h(X) + \log(\det(A))$.

4. *Conditioning cannot increase differential entropy:* $h(X) \geq h(X|Y)$ *with equality if and only if* X *and* Y *are independent.*

5. *Chain rule:*

    - *small chain rule:* $h(X,Y) = h(X) + h(Y|X) = h(Y) + h(X|Y)$
    - *full chain rule:* $h(X_1, ..., X_n) = h(X_1) + \sum_{i=2}^{n} h(X_i|X_1, ..., X_{i-1})$

Remark 9.4 (h **vs.** H**)** *Note that* h *and* H*, while seemingly quantify similar ideas, posses some vastly different behaviors.*

- $H(P) \geq 0$ *while* $h(P)$ *may be negative as seen in Example 9.2.*
- $H(X) \geq H(f(X))$ *while this is not true for* h *(see the scaling property in Proposition 9.4).*
- $H(X+Y) \leq H(X+Y,Y) = H(X,Y) \leq H(X) + H(Y)$*. However, one can find example of continuous distributions for which* h *is not subadditive.*