

Lecture 1: Information measures: Entropy and Divergence

Lecturer: Yihong Wu

Scribe: Don Corleone, Jan 21, 2016

Two methods to describe a random variable (R.V.)  $X$ .

- a function  $X : \Omega \rightarrow \mathcal{X}$  from the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  to a target space  $\mathcal{X}$ .
- a distribution  $P_X$  on some measurable space  $(\mathcal{X}, \mathcal{F})$ .

How to measure the randomness?

## 1.1 Entropy

**Definition 1.1** (Discrete R.V.). Random variable  $X$  — discrete if there exists a countable set  $\mathcal{X} = \{x_j, j = 1, \dots\}$  such that  $\sum_{j=1}^{\infty} P_X(x_j) = 1$ .

**Note:**  $\mathcal{X}$ : alphabet of  $X$ ,  $x \in \mathcal{X}$ : atoms,  $P_X(x_j)$ : probability mass function (pmf).

**Definition 1.2** (Entropy). For a discrete R.V.  $X$  with distribution  $P_X$ :

$$\begin{aligned} H(X) &= \mathbb{E} \left[ \log \frac{1}{P_X(X)} \right] \leftarrow [\text{Measure of randomness of } X \text{ (intuitively)}] \\ &= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}. \end{aligned}$$

Why such definition, why log, why entropy? — later.

**Note:**

- We agree that  $0 \log \frac{1}{0} = 0$  (by continuity of  $x \mapsto x \log \frac{1}{x}$ )
- Also write  $H(P_X)$  instead of  $H(X)$  (abuse of notation, as customary in information theory).
- Basis of log — units

$\log_2 \leftrightarrow$  bits

$\log_e \leftrightarrow$  nats

$\log_{256} \leftrightarrow$  bytes

$\log \leftrightarrow$  arbitrary units, base always matches exp

**Note:** Elementary properties of  $H$ :

1.  $H(X) \geq 0$ , with equality iff  $X = x$  a.s. for some  $x$ .

2.  $H(X) \leq \log |\mathcal{X}|$ , with equality iff  $X$  is uniform on  $\mathcal{X}$ . (*proof: Jensen's inequality.*)

3. Q: Can  $H(X) = +\infty$ ?

A: Yes, example:  $\mathbb{P}[X = k] = \frac{c}{k \ln^2 k}, k = 2, 3, \dots$

**Example 1.1** (Bernoulli).  $X \in \{0, 1\}$ ,  $\mathbb{P}[X = 1] = P_X(1) \triangleq p$

$$H(X) = h(p) \triangleq p \log \frac{1}{p} + \bar{p} \log \frac{1}{\bar{p}}$$

where  $h(\cdot)$  is called the **binary entropy function**.

**Proposition 1.1.**  $h(\cdot)$  is continuous, concave on  $[0, 1]$  and

$$h'(p) = \log \frac{\bar{p}}{p}$$

with infinite slope at 0 and 1.

**Example 1.2** (Geometric).  $X \in \{0, 1, 2, \dots\}$   $\mathbb{P}[X = i] = P_x(i) = p \cdot (\bar{p})^i$

$$\begin{aligned} H(X) &= \sum_{i=0}^{\infty} p \cdot \bar{p}^i \log \frac{1}{p \cdot \bar{p}^i} = \sum_{i=0}^{\infty} p \bar{p}^i \left( i \log \frac{1}{\bar{p}} + \log \frac{1}{p} \right) \\ &= \log \frac{1}{p} + p \cdot \log \frac{1}{\bar{p}} \cdot \frac{1-p}{p^2} = \frac{h(p)}{p} \end{aligned}$$

**Theorem 1.1** (Invariance under relabeling).  $H(X) = H(f(X))$  for any bijective  $f$ .

**Definition 1.3** (Joint entropy).  $X^n = (X_1, X_2, \dots, X_n)$  – a random vector with  $n$  components.

$$H(X^n) = H(X_1, \dots, X_n) = \mathbb{E} \left[ \log \frac{1}{P_{X_1, \dots, X_n}(X_1, \dots, X_n)} \right]$$

For more see [CT06].

# Bibliography

- [CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.