

# Information Theory Basics

Weizi Li

Department of Computer Science  
University of Memphis



- To understand variational autoencoder (which uses variational inference), we need basic information theory
- Shannon, A Mathematical Theory of Communication, 1948

- In computing: 0 or 1
- In information theory: each bit should divide information uncertainty by 2

- Tomorrow weather: rainy 50%, sunny 50%
- Assuming weather station is always true
- Now weather station tells you tomorrow is going to be sunny  
= send you **1** bit of information because the uncertainty is reduced by half **once**:  $50\% = \frac{1}{2^1} = \frac{1}{2}$
- Reversely:  $-\log_2^{\frac{1}{2}} = \log_2^2 = 1$

- Tomorrow weather: rainy 75%, sunny 25%
- Now weather station tells you sunny = send you 2 bits of information because the uncertainty is reduced by half twice:  
 $25\% = \frac{1}{2^2} = \frac{1}{4}$
- Reversely:  $-\log_2^{\frac{1}{4}} = \log_2^4 = 2$

- Tomorrow weather: 8 types of weather with each at 12.5%
- Now weather station tells you 1 of them = send you **3** bits of information because the uncertainty is reduced by half **three times**:  $12.5\% = \frac{1}{2^3} = \frac{1}{8}$
- Reversely:  $-\log_2^{\frac{1}{8}} = \log_2^8 = \mathbf{3}$

- information = number of bits =  $-\log_2^{q(x)}$
- $q(x) \rightarrow 0$ : information increases (knowing odd event = gaining a lot of information)
- $q(x) \rightarrow 1$ : information decreases

- Tomorrow weather: rainy 75%, sunny 25%
- Weather station says rainy:  $-\log_2^{0.75} = 0.41$  (less information)
- Weather station says sunny:  $-\log_2^{0.25} = 2$  (more information)
- Total information from weather station:  
 $75\% \times 0.41 + 25\% \times 2 = 0.81$
- This is also called **entropy**:  
 $H(\mathbf{q}) = -\sum_i q_i \log_2^{q_i}$ ,  $\mathbf{q} = [q_1, q_2, \dots]$  is a distribution



- Entropy: average minimum number of bits needed to encode a string of symbols, based on the frequency of the symbols
- Usage: Coding, Kullback–Leibler (KL) Divergence, ...

- Example: 8 types of weather with probability 35%, 35%, 10%, 10%, 4%, 4%, 1%, 1%
- Entropy:  $-\sum q \log_2 q = 2.23$  bits

- Scheme 1: 000, 001, 010, 011, 100, 101, 110, 111
- Average: 3 bits
- Wasting:  $3 - 2.23 = 0.77$  bits

- Scheme 2: 00, 01, 100, 101, 1100, 1101, 11100, 11101
- Average:  $0.35 \times 2 + 0.35 \times 2 + 0.1 \times 3 + 0.1 \times 3 + 0.04 \times 4 + 0.04 \times 4 + 0.01 \times 5 + 0.01 \times 5 = 2.42$  bits
- Wasting:  $2.42 - 2.23 = 0.19$  bits

- Goal: measure similarity between two distributions
- $KL(\mathbf{q}||\mathbf{p}) = -\sum_i q_i \log p_i - (-\sum_i q_i \log q_i) = -\sum \mathbf{q} \log \mathbf{p} / \mathbf{q}$ 
  - ▶  $-\sum_i q_i \log p_i$ : cross entropy
  - ▶  $-\sum_i q_i \log q_i$ : entropy
  - ▶  $KL = \text{cross entropy} - \text{entropy}$
- When  $\mathbf{q}$  and  $\mathbf{p}$  are similar  $\rightarrow$  cross entropy and entropy are close  $\rightarrow KL(\mathbf{q}||\mathbf{p})$  is small

- $KL(\mathbf{q}||\mathbf{p}) \neq KL(\mathbf{p}||\mathbf{q})$ , not a distance
- $KL \geq 0$ , since it's the sum of {probability  $\times$  information}