# Cross-Entropy & KL Divergence

**Entropy** is a measure of how <u>surprised</u> we are.

⇒ The more unpredictable / uncertain a system is, the more potential for surprise (e.g. surprise of correctly predicting lottery >> coin flip)

Principles for <u>information content</u> of event, $I(E)$:

① Continous function

② Surprise ↑ as probability of event ↓ ⇝ relate to $\frac{1}{p}$

③ Info from 2 independent events should be additive

⇓

$$I(E) = \log\left(\frac{1}{p}\right) = -\log(p)$$

E.g. Flip 2 coins:

- Prob([Head, Head]) $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
- Info([Head, Head]) $= -\log\left(\frac{1}{4}\right) = 2$

} We gain <u>2 bits</u> of info. $E \in \{00, 01, 10, 11\}$

<u>Def</u>: One bit of info = info that reduces uncertainty by $\frac{1}{2}$

- Case 1: 25% HH, 25% HT, 25% TH, 25% TT

  ⇒ Knowing 1st flip is H is 1 bit

  ("Surprise" decreases by: $\frac{-\log 4}{-\log 2} = \frac{2}{1} = 2^1$)

- Case 2: 75% Sunny, 25% Rain

  ⇒ Getting "Sunny" $= -\log_2(0.75) = 0.415$ bits

  ⇒ Getting "Rainy" $= -\log_2(0.25) = 2$ bits

  } A more surprising outcome gives us more info to update our existing beliefs!

<u>Entropy</u>. Average uncertainty / expected info gain of an RV.

$$h(S) = \underbrace{-\Sigma_{i \in S}}_{\text{across all events}} \underbrace{P(i)}_{\text{prob. of event}} \underbrace{\log P(i)}_{\text{info gain from event}}$$

} <u>expected value!!</u>

Greater uncertainty in system → higher entropy → higher expected info gain from each obs

Know nothing abt coin ↓ Full 1 bit info

Know coin is 99% H ↓ Little info gain, v. likely H

**Cross-Entropy:** How surprised are we when we use an estimated distribution $q$ to predict a true distribution $p$?

$$L = -\sum_{k=1}^{K \text{ classes}} \underbrace{p(k)}_{\text{actual prob of observing } k} \underbrace{\log q(k)}_{\text{our surprise when observing } k}$$

⇒ Larger divergence between $p$ and $q$ → obs are more surprising on avg
→ higher entropy → more info gain when updating our beliefs!!
[Perfect for loss function — captures how surprised I am by the correct ans']

**KL Divergence:** $KL(P \| Q) = \underbrace{-\sum_{x \in X} P(x) \log Q(x)}_{\text{surprise using } Q} \underbrace{- \sum_{x \in X} P(x) \log P(x)}_{\text{surprise using } P}$

$= -\sum_{x \in X} P(x) \cdot [\log Q(x) - \log P(x)]$

$= -\sum_{x \in X} P(x) \cdot \log\left(\frac{Q(x)}{P(x)}\right)$

↓
Measure of how diff 2 prob dists are'

↳ Cross-Entropy: $\underbrace{(\text{surprise due to divergence of } P \;\&\; Q)}_{\text{KL divergence isolates this!!}} + (\text{inherent surprice in } P)$