Information, entropy, cross entropy, KL-Divergence

Shannon information

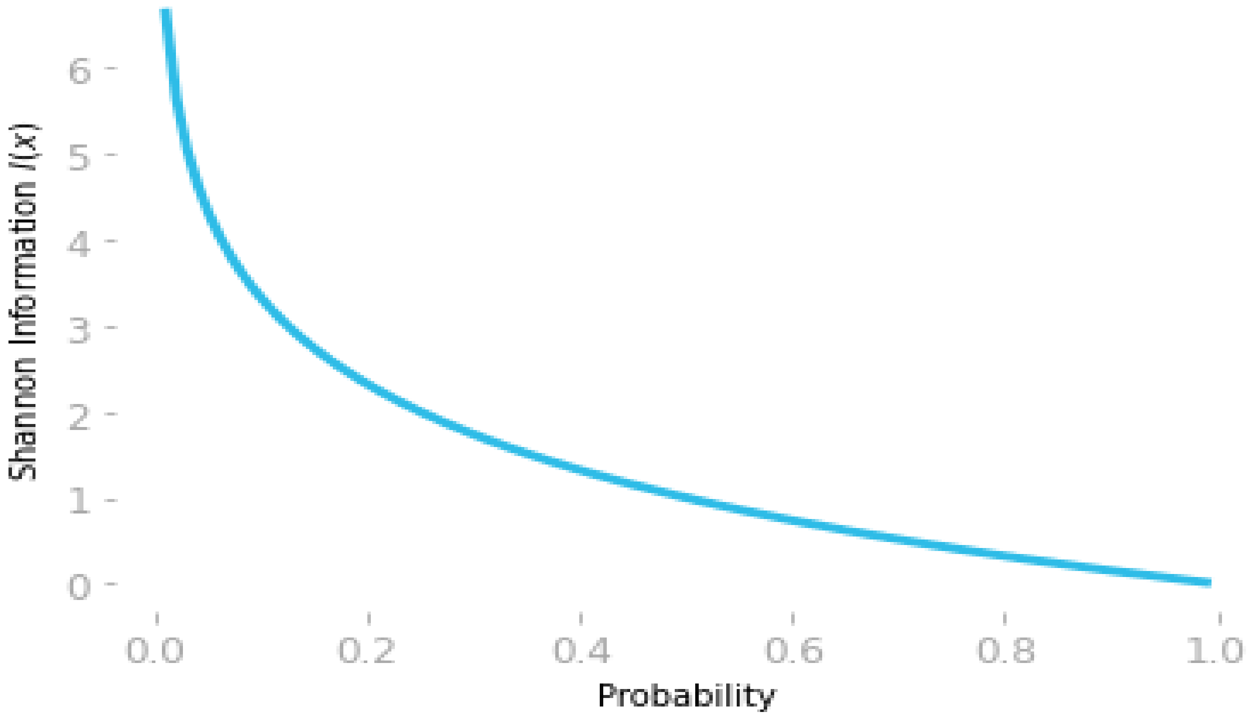
- Message for event with higher probability has less information
- “It is sunny in LA tomorrow”→ not surprise→ less information
- “There will be a Tsunami around Taiwan.” → much surprise→ more information

- Information is a measure for surprise, and high probability implies low information and vice versa.
- Information of two independent events is additive. I(xy)=I(x) +I(y)
- To represent an information as number of bits.

- I(x), Information of x with probability p is represented by

$$I(x) = -\log_2 P(x)$$

- $$I(x,y) = -\log_2 P(x,y) = -(\log_2 P(x) + \log_2 P(y))$$
- = I(x) + I(y), additive


Shannon Information I(x)
6
5
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
Probability

Entropy of a distribution, the expected information

- Consider for instance a biased coin, where you have a probability of 0.8 of getting ‘heads’.
- Here is your distribution: you have a probability of 0.8 of getting ‘heads’ and a probability of *1–0.8 = 0.2* of getting ‘tails’.
- These probabilities are respectively associated with a Shannon information of:

$$-\log_2 0.8 \approx 0.32$$

- And

$$-\log_2 0.2 = 2.32$$

- Entropy of this distribution:

$$0.8 \cdot (-\log_2 0.8) + 0.2 \cdot (-\log_2 0.2) = 0.26 + 0.46 = 0.72$$

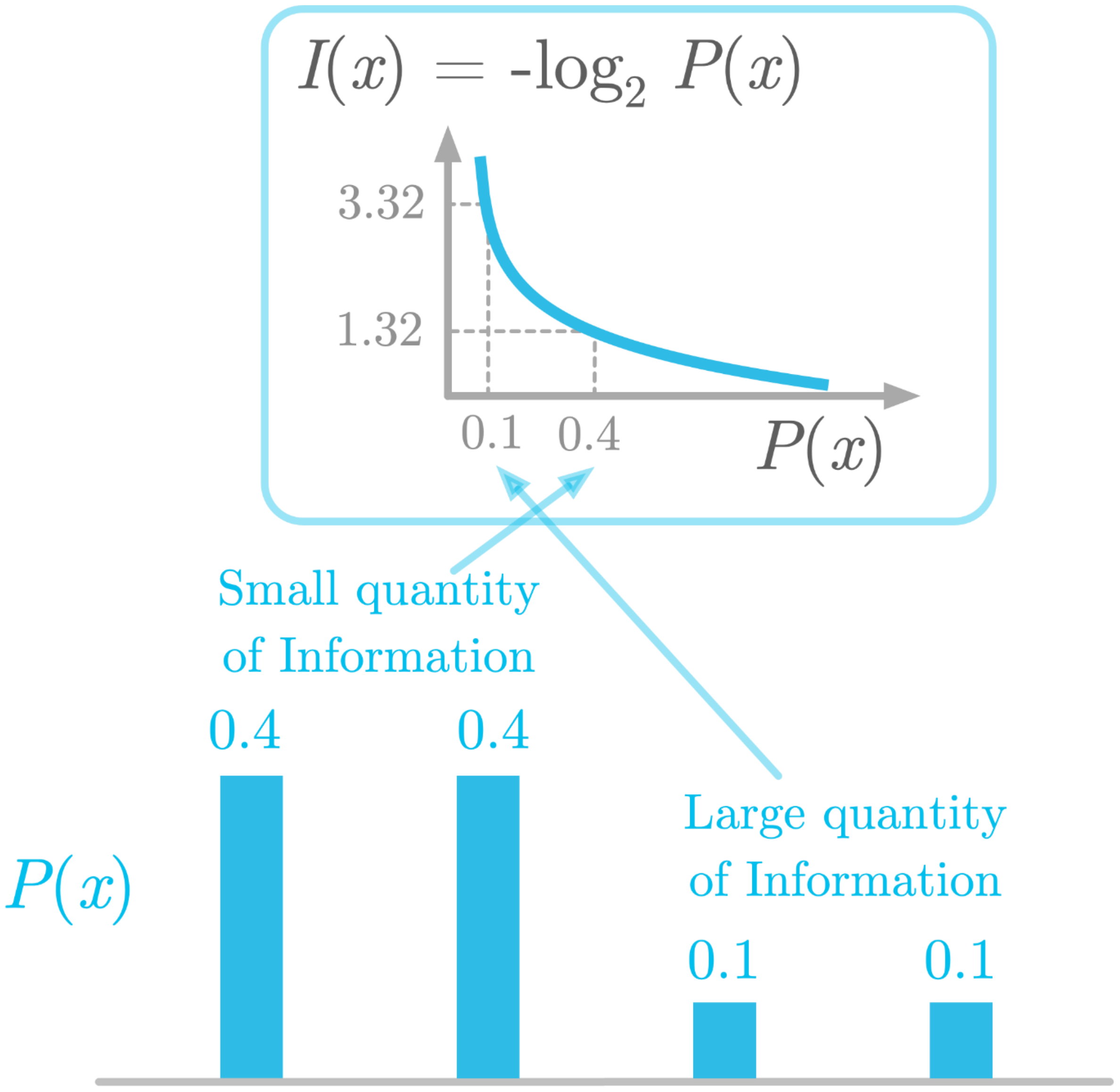
- To summarize, you can consider the entropy as a summary of the information associated with the probabilities of the discrete distribution:
  1. You calculate the Shannon information of each probability of your distribution.
  2. You weight the Shannon information with the corresponding probability.
  3. You sum the weighted results.

Expectation

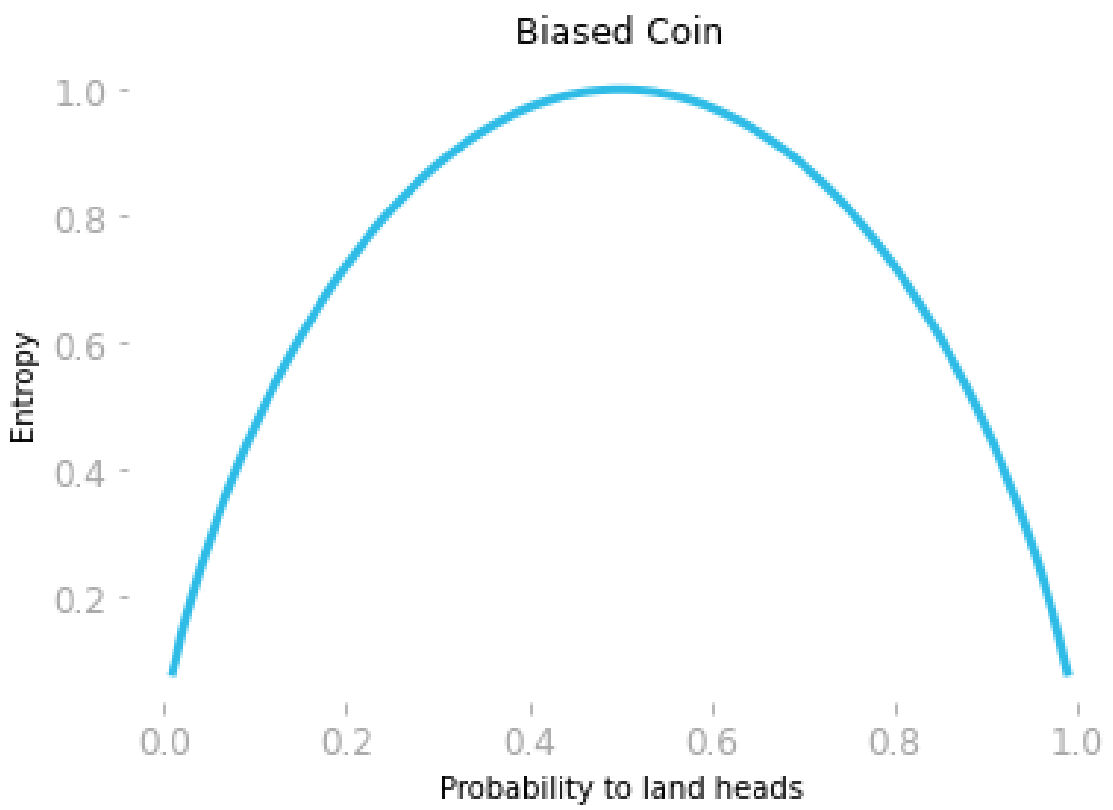
$$\mathbb{E}[\mathrm{X}] = \sum_{i=1}^{n} P(x_i) x_i$$

$$H(\mathrm{X}) = \mathbb{E}[I(x)] = -\sum_{x} P(x) \log_2 P(x)$$

Weighted sum of information, weighted by their corresponding probabilities



$I(x) = -\log_2 P(x)$
3.32
1.32
0.1
0.4
$P(x)$
Small quantity
of Information
Large quantity
of Information
0.4
0.4
0.1
0.1
$P(x)$



- *Figure 3: Entropy as a function of the probability to land "heads".*

Cross-Entropy

- The concept of entropy can be used to compare two probability distributions: this is called the *cross-entropy* between two distributions, which measures how much they differ.

- You can also consider cross-entropy as the expected quantity of information of events drawn from *P(x)* when you use *Q(x)* to encode them.

- I like “Information of distribution Q weighted sum by distribution P(x)”

- Information of Q(x)

$$H(P, Q) = -\sum_{x} P(x) \log_2 Q(x)$$

$$H(P, Q) \neq H(Q, P)$$

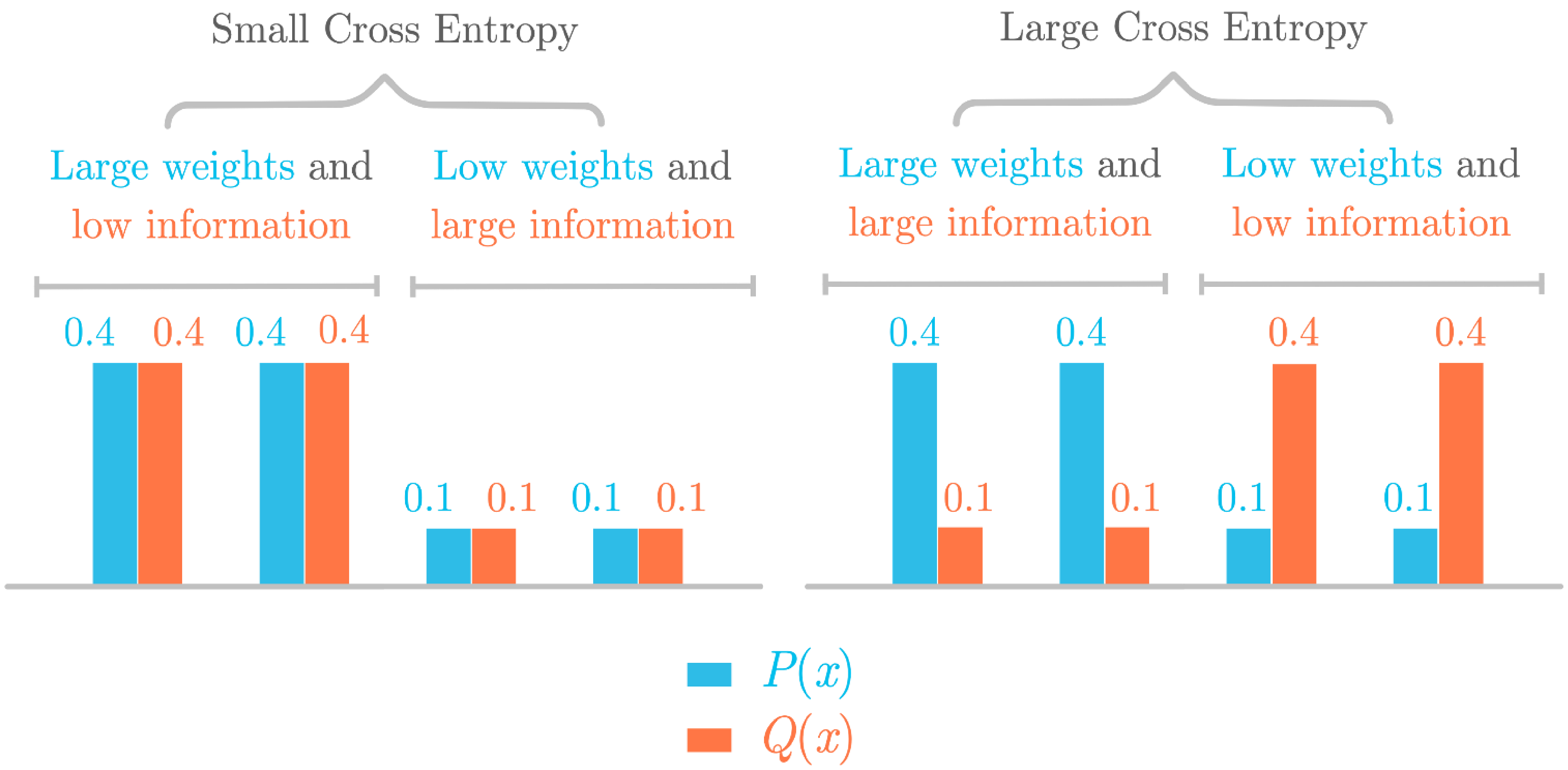
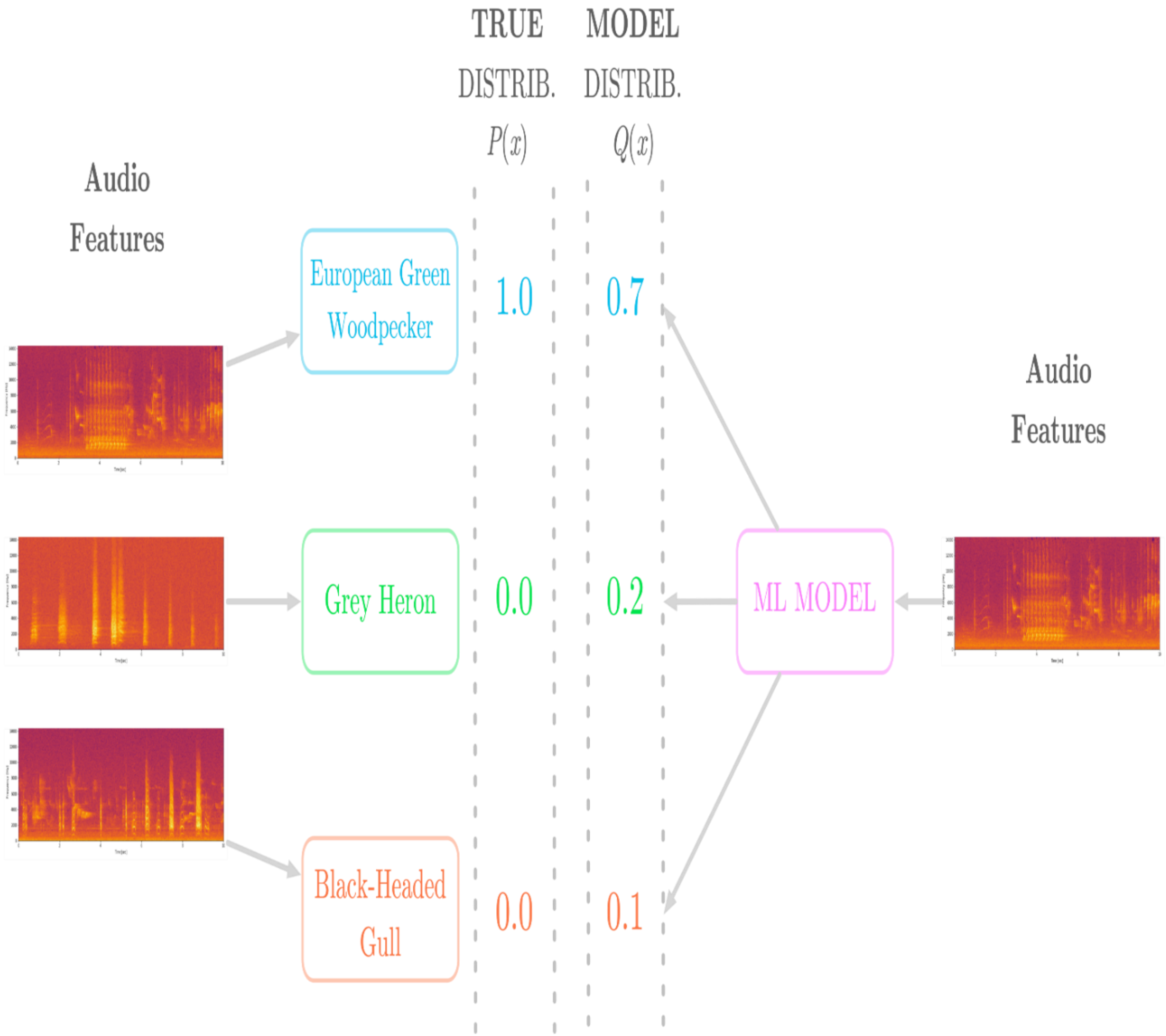
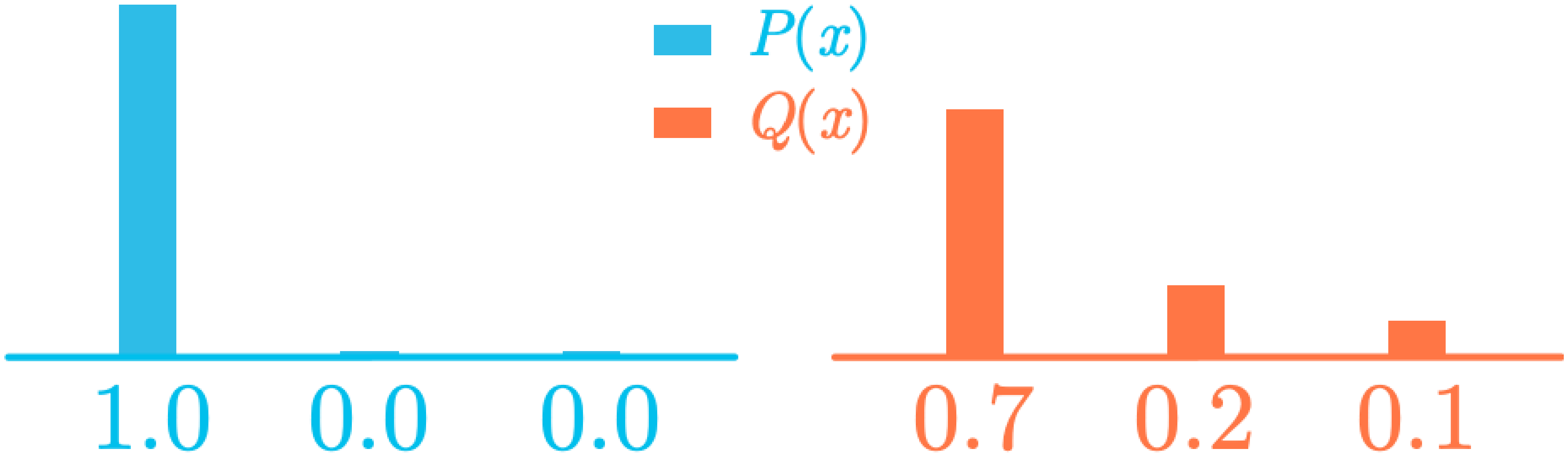


Figure 4: Illustration of the cross-entropy as the Shannon information of Q(x) weighted according to the distribution of P(x).

Classification application




$P(x)$
$Q(x)$
1.0
0.0
0.0
0.7
0.2
0.1

$$
\begin{aligned}
H(P,Q) &= -\sum_{x} P(x)\log Q(x) \\
&= -(1.0\log 0.7 + 0.0\log 0.2 + 0.0\log 0.1) \\
&= -\log 0.7
\end{aligned}
$$

- In machine learning, the cross-entropy is widely used as a loss for binary classification: the log loss

$$H(P,Q) = -\sum_{x} P(x)\log Q(x)$$

$$= -(y\log(\hat{y}) + (1-y)\log(1-\hat{y}))$$

Kullback-Leibler Divergence (KL Divergence)

- Intuitively, the KL divergence is the supplemental amount of information associated with the encoding of the distribution *Q(x)* compared to the true distribution *P(x)*.

- It tells you how different the two distributions are.

$$D_{KL}(P||Q) = H(P,Q) - H(P) \geq 0$$

- Note that when P=Q, H(P,Q) is minimized and it is equal to H(P)

$$
\begin{aligned}
D_{KL}(P||Q) &= H(P,Q) - H(P) \\
&= -\sum_x P(x)\log_2 Q(x) - (-\sum_x P(x)\log_2 P(x)) \\
&= \sum_x P(x)\log_2 P(x) - \sum_x P(x)\log_2 Q(x)
\end{aligned}
$$

- The KL divergence is always non-negative. Since the entropy *H(P)* is identical to the cross-entropy *H(P, P)*, and because the smallest cross-entropy is between identical distributions (*H(P, P)*), *H(P, Q)* is necessarily larger than *H(P)*.
- In addition, the KL divergence is equal to zero when the two distributions are identical.
- In sum, information→measure of supprise of a probability=$-\log_2 P(x)$
- Entropy → weighted sum of the information of a distribution
- Cross-entropy H(P,Q)→ weighted sum of information of Q(x) using P(x) as the weights
- KL divergence $D_{KL}(P||Q) = H(P,Q) - H(P) \geq 0$
- Measure the distance of dist. Q from P