Information

- For a probability distribution $p(X)$ for a rv X, define the information of outcome x to be ($\log$ = nat log)

$$I(x) = -\log p(x)$$

- This is 0 if p is 1 (no information for a certain outcome) and is large if p is near 0 (lots of information if the event is not likely).
- **Additivity:** If X and Y are indep, then info is additive:

$$I(x,y) = -\log p(x,y) = -\log p(x)p(y) = -\log p(x) - \log p(y)$$
$$= I(x) + I(y)$$

PURDUE UNIVERSITY

Entropy

- Entropy is the expected information of a rand var:

$$
\begin{aligned}
H(X) &= E(I(X)) \\
&= E(-\log(X)) \\
&= \int_W -p(x) \log p(x) dx
\end{aligned}
$$

  - Note that 0 log 0 = 0
- Entropy is a measure of *unpredictability* of a random variable. For a given set of states, equal probability gives maximum entropy.

PURDUE UNIVERSITY

Cross-entropy

- Compare one distribution to another.
- Suppose we have distribution p,q on same set W. Then

$$\begin{aligned} H(p,q) &= E_{X\sim p}(I(X \sim q)) \\ &= E_{X\sim p}(-\log q(X)) \\ &= \int_W -p(x)\log q(x)dx \end{aligned}$$

- In the discrete case,

$$H(p,q) = \sum_i -p(x_i)\log q(x_i)$$

Cross entropy as loss function

- Question: given p(x), what q(x) minimizes the cross entropy (in the discrete case)?
- Constrained optimization:

$$\min_q -\sum_i p_i \log q_i \text{ subject to}$$

$$\sum_i q_i = 1 \text{ and } q_i \geq 0 \text{ for all } i$$

Constrained optimization

- More general constrained optimization:

$$\min_x f(x) \text{ subject to}$$

$$g_i(x) = 0 \text{ and } h_j(x) \geq 0 \text{ for all } i, j$$

  - f is the objection function (loss)
  - g_i are the equality constraints
  - h_j are the inequality constraints
- If no constraints: look for a point where gradient of f vanishes. But we need to include constraints.

PURDUE UNIVERSITY