

Say we're working in ML and we want our model to predict a prob distribution P where the model's output is a prob distribution Q . We'd want some kind of scoring function to measure how well our model is fitting against P .

There's multiple possible choices; MSE works for example. But if we impose the (somewhat arbitrary) condition that the scoring function must be local (that is, when we sample an event $x \sim P$, the score only depends on q_x) as well as that all observations carry equal weight (which is a pretty default reasonable assumption), then we get cross-entropy specifically.

We'll find a local scoring function $f(p)$ where: given that an event x occurred and an observer predicted that it would happen with probability p , how well did they predict this would happen? Given that all observations are equal weight, our holistic score is the average of the local score where each observation is weighted equally: $\mathbb{E}_{x \sim P}[f(Q(x))]$.

Given that we know this, f must be defined so that $\mathbb{E}_{x \sim P}[f(Q(x))]$ is proper; in other words, it has a unique global max at $P = Q$. This condition alone narrows down what f can be (we'll work with discrete distributions, but continuous distributions structurally follow the same proof):

- Use Lagrange multipliers: we want to maximize $G(P, Q) = \sum_x (p_x f(q_x))$ with the constraint $\sum_x q_x = 1$.
- $\mathcal{L} = \sum_x (p_x f(q_x)) - \lambda \left(\sum_x q_x - 1 \right)$
- Now we need $\nabla_q \mathcal{L} = 0$; do this component-wise. For each component q_x , $\frac{\partial \mathcal{L}}{\partial q_x} = p_x f'(q_x) - \lambda$. With $\nabla_q \mathcal{L} = 0$, we have $p_x f'(q_x) = \lambda$.
- Our propriety condition says at optimum, $q = p$, so $p_x f'(p_x) = \lambda$. Across all possible distributions p that exist, this must be true; we'll show that λ is constant across distributions to get a nice general $t f'(t) = \lambda$ form (currently λ is a function of P by our setup).
 - Assume ≥ 3 outcomes; construct a "base set" of distributions with probability vectors $(a, b, 1 - a - b)$ for all $a + b < 1$. Denote $g(t) = t f'(t)$ (just shorthand for readability); then, $g(a) = \lambda(P)$ and $g(b) = \lambda(P)$, so $g(a) = g(b)$. So any distribution P' that contains a probability that exists inside a base set distribution satisfies $\lambda(P') = \lambda(P)$.
 - Utilize this to complete the base set: for unconnected (a, b) pairs where $a + b \geq 1$, connect (a, c) and (c, b) where $a + c < 1$ and $c + b < 1$; both are guaranteed to exist from the $a + b < 1$ base set. This shows $g(a) = g(c) = g(b)$, and thus $g(a) = g(b)$.
 - So $g(a) = g(b)$ for all $(a, b) \in (0, 1) \times (0, 1)$, which means all λ across all distributions with ≥ 3 outcomes are equal as well.
 - Note: for 2 outcomes, we can't fully connect all (a, b) pairs, only $(a, 1 - a)$ pairs; log still works for reasons aside from the continued proof below, but is no longer uniquely required.
- So $t f'(t) = \lambda$ for $t \in (0, 1]$; a single λ value for all possible t .
- The ODE $f'(t) = \frac{\lambda}{t}$ gives solutions $f(t) = \lambda \log t + c$. So: all functions f of the form $f(t) = \lambda \log t + c$ gives a critical point at $q = p$ in the function $G(P, Q) = \sum_x (p_x f(q_x))$. But of these functions f , which give a global max at $q = p$ — not just a critical point?
 - If G is concave and $q = p$ is a critical point, then $q = p$ is **the** global max (top of the dome).
 - G is concave iff f is. $f'(t) = \frac{\lambda}{t} \implies f''(t) = -\frac{\lambda}{t^2}$. t^2 is always positive, so for concavity ($f'' < 0$), we need $\lambda > 0$.
- So finally: $\boxed{f(p) = \lambda \log p + c \quad (\lambda > 0)}$.

With $\lambda = 1$ and $c = 0$, we have $f(p) = \log p$ and $G(P, Q) = \sum_x p_x \log(q_x)$. This is just cross-entropy but negative, since we used a convention of higher = better, but loss functions standardize higher = worse. Negating it gives us the standard cross-entropy function:

$$\boxed{H(P, Q) = - \sum_x p_x \log(q_x)}$$