1 Problem

We have a probability distribution P. How do we measure how close Q's probabilities are to P's?

2 Solution

We'll do this by scoring how well Q predicts events that occur according to P; over infinite trials, only distributions equal to P should predict events sampled from P the best. (P scoring P the best is **not** a universal guarantee under all "scoring functions" — only proper ones exhibit that behavior — the point of this writeup is to **find** such a scoring function which makes P predict P the best.)

We'll find a 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? Then, we'll score a distribution based on its **expected value** of f over its possible outcomes (ie, distribution score is $\mathbb{E}_{x\sim P}[f(Q(x))]$ for samples drawn from P, evaluating distribution Q).

Given that we know this, f must satisfy the following conditions:

- **Differentiable**: basically, f is a normal-feeling function. This is required for the next condition.
- **Proper**: if P, Q predict event x occurs with probabilities p_x, q_x respectively, the distribution score $\sum_x (p_x f(q_x))$ must be maximized at only $q = p$.
- **(Arbitrarily imposed) local**: this is the arbitrary assumption that specifically leads us to KL vs. a different divergence like Brier or spherical. We're assuming that the score of a prediction for a single realized outcome doesn't depend on any other predictions.

The propriety condition carries the entire path to narrowing down what f can be (working with discrete distributions, but continuous distributions structurally follow the same proof):

- Use Lagrange multipliers: we want to maximize $G(q) = \sum_x (p_x f(q_x))$ 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, $\dfrac{\partial \mathcal{L}}{\partial q_x} = p_x f'(q_x) - \lambda$. 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 $tf'(t) = \lambda$ 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 $g(t) = tf'(t)$ (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 $tf'(t) = \lambda$ for $t \in (0, 1]$; a single λ value for all possible t.
- The ODE $f'(t) = \dfrac{\lambda}{t}$ 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 $G(q) = \sum_x (p_x f(q_x))$. 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. $f'(t) = \dfrac{\lambda}{t} \implies f''(t) = -\dfrac{\lambda}{t^2}$. t^2 is always positive, so for concavity ($f'' < 0$), we need $\lambda > 0$.
- So finally: $\boxed{f(p) = \lambda \log p + c \quad (\lambda > 0)}$.

Now, $G_p(q)$ (denoting $G(q)$ with p as the probabilities sampled from) is essentially **cross-entropy loss**: except since loss functions standardly have higher = worse, cross-entropy is actually $H(p, q) = -G_p(q)$:

$$\boxed{H(p, q) = -\sum_x p_x \log(q_x)}$$

KL divergence calibrates the best possible score to be zero, by subtracting the best possible cross-entropy score (which is the entropy of P):

$$D_{KL}(P||Q) = \mathbb{E}_{x\sim P}[\log P(x)] - \mathbb{E}_{x\sim P}[\log Q(x)] = \boxed{\mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right]}$$