Let's consider two normal distributions: one centered at 1 (denoted as p) and the other centered at 0.5 (denoted as q). Both have a standard deviation of 0.5. Since we are using grid-approximated distributions, they are essentially vectors of 1000 values each.

The distribution p represents our hypothesis about the distribution of an experiment, while distribution q represents the observed outcome of the experiment. Our goal is to measure how accurate our estimations were. To achieve this, we can employ various distance measures between the distributions.

Thinking of these distributions as vectors allows us to use a simple comparison between corresponding values, followed by summing the results. The Cramer-von Mises, one of the most popular scoring rules, calculates the sum of squared distances between each pair of p and q. The result provides a measure of how accurate or inaccurate our hypothesis p was compared to the true distribution q.

$$D_{\text{CM}}(p, q) = \sum_x |P(x) - Q(x)|^2$$

It's worth noting that the distribution being compared with an inaccuracy score might take the form of a distribution that assigns values to only one point on the x-axis. This is particularly relevant when dealing with definitive beliefs expressed as values of 0 and 1 (or sometimes cumulative values). However, the distance measures still operate similarly, treating it as a distribution.

This idea stays roughly the same for all the cases in the paper, things that vary in the examples are: (1) distribution shapes, (2) inaccuracy score.

The case will be slightly different when it comes to the expected inaccuracy score. In this, we are not comparing a hypothesis with a known outcome, but a hypothesis with the set of all possible outcomes. In the case of binary outputs (like a coin toss), the set of outputs has only two elements:

$$\mathbb{E}_{\text{binary}}(p, q) = I(p, \text{heads})\mathbb{E}q(\text{heads}) + I(p, \text{tails})\mathbb{E}q(\text{tails})$$

But it might be a case that we are talking about an interval of possible outcomes. Let's say we are comparing our earlier distribution p (centered at 1) with a set of possible outcomes that stretches a possible mean of the normal distribution from 0.1 to 2 with a difference of 0.1 (all of them are weighted uniformly).

Our expected inaccuracy score, in such an instance, takes the score of all the possible outcomes to represent how our initial hypothesis performs in all possible worlds.

Let's now take an example with a mystery coin. A rational agent is proposed a bet on a bias of a coin drawn from a bag full of biased coins. Only one coin will be drawn from the bag. In the bag, there are 10 coins, 5 with a normal distribution centered at 0.3 and 5 with a distribution centered at 0.5, both with a standard deviation of 0.05.

One of the possible approaches in scoring an agent's possible credal states is calculating expected inaccuracy for the two possible worlds, so either a coin might have a bias represented with a normal distribution centered at 0.3 or 0.5.