Proper Scoring Rules and Prediction

Today we look at the question: How can we incentivize an agent to make an accurate prediction? This is a relatively simple question in that there will only be a single agent (for now). However, it connects to some very beautiful math.

Proper scoring rules. There is a future event or random variable Y with a finite set $\mathcal{Y}$ of possible outcomes. For example, $\mathcal{Y} = \{\text{rain, no rain}\}$. Let $\Delta_{\mathcal{Y}}$ be the set of probability distributions on $\mathcal{Y}$.

1. A single agent, the "expert", reports a probability distribution $p \in \Delta_{\mathcal{Y}}$. This is interpreted as a prediction of the chance of each outcome.
2. The mechanism then observes the true outcome, say y.
3. The mechanism gives the agent a "score" according to a function $S : \Delta_{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$. Their score is $S(p, y)$.

We can interpret the score as a payment that the mechanism will give to the agent. **We assume that the agent's goal is always to maximize expected score.** Suppose that the agent believes the true distribution is q and she reports p. Then let us use the notation $S(p; q)$ for her expected score:

$$S(p; q) := \mathop{\mathbb{E}}_{q} S(p, Y) = \sum_{y} q(y) S(p, y).$$

Definition 1 *A **scoring rule** is a function $S : \Delta_{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$. It is **proper** if truthfulness maximizes expected score: for all beliefs q and all $p \neq q$*

$$S(q; q) \geq S(p; q).$$

We call S *strictly proper* if the inequality is strict for all $p \neq q$; this means that it is strictly better to be truthful than misreport.

Let δ_y be the distribution putting probability one on outcome y. Notice that if an agent believes δ_y, then she will definitely receive $S(p, y)$ for prediction p. So $S(p; \delta_y) = S(p, y)$.

Aside: recapping convexity. We need to review some mathmatical tools here. Consider $\mathbf{R}^d$. For example $\mathbf{R}^1$ would be some points or intervals on the real line; $\mathbf{R}^2$ is the plane; $\mathbf{R}^3$ is 3-d space, etc. A point in $\mathbf{R}^d$ can be written as a vector $x = (x_1, \dots, x_d)$ where each x_i is a real number.

Given two points x, x', we can consider the line segment connecting x and x'. These are the points given by $\alpha x + (1 - \alpha)x'$, for all $\alpha \in (0, 1)$. When $\alpha = 0$, we get x'; when $\alpha = 1$, we get x; and when $\alpha = 0.5$, we get the midpoint between them.

Suppose we have a set C of points in $\mathbf{R}^d$. We say C is *convex* if, for any two points $x, x' \in C$, the line connecting x and x' also completely lies within C. In notation, for all α, $\alpha x + (1 - \alpha)x'$ is in C.

A disk is convex; Pacman is not. A soccer ball is convex; a hockey stick is not.

Now, a function $f : \mathbf{R}^d \to \mathbf{R}$ is called **convex** if the set of points above the function is a convex set. An equivalent definition is that, for any two points on the function's graph, the line connecting them lies above the functions graph: for any x, x' and any $\alpha \in (0, 1)$, if we let $y = \alpha x + (1 - \alpha)x'$, then $f(y) \leq \alpha f(x) + (1 - \alpha)f(x')$. We will assume for simplicity that f is differentiable.

Figure 1: Left: a convex set. Right: a set that is not convex.

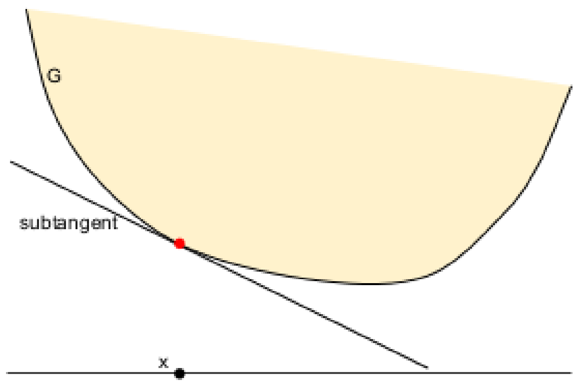


Figure 2: A convex function G. One way to tell it is a convex function is that the yellow region is a convex set. Also shown is the linear function that is tangent at to G at x. The slope of this function is $\nabla G(x)$. Another way to tell that G is convex is that all such tangent lines lie completely below G.

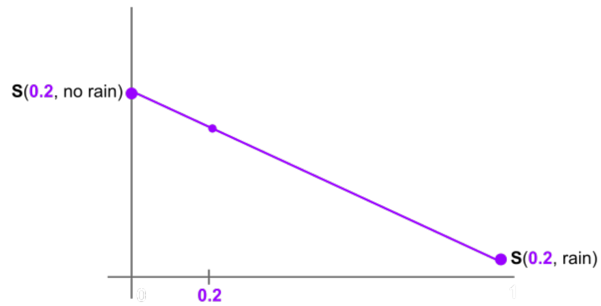
The key property of convex functions is that if we take the derivative (slope) at any point x, this gives a linear function that is tangent to f at x and lies below f everywhere else. i.e. the derivative $\nabla f(x)$ of f at x is the vector that satisfies, for all x',

$$f(x') \geq f(x) + \nabla f(x) \cdot (x' - x).$$

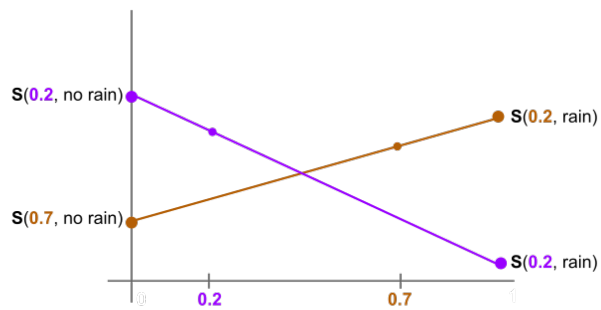
You can focus on the one-dimensional case in the figures where $\nabla f(x)$ is the derivative of f at x, though what we say will extend to higher dimensions where $\nabla f(x)$ is the derivative.

Proper scoring rule characterization. Now the key question: how do we construct proper scoring rules? First, let's see in pictures what happens for a proper scoring rule S when we consider an agent's best response and whether it is to be truthful.

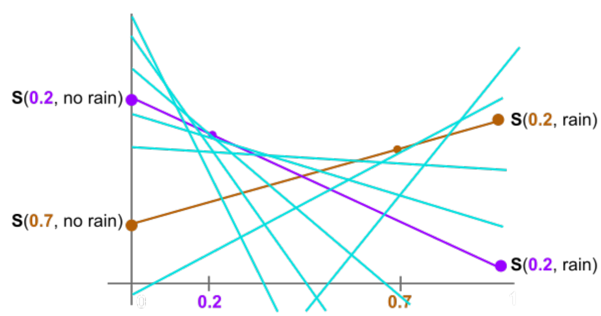
Figure 3: Suppose $\mathcal{Y} = \{\text{rain, no rain}\}$ and the agent reports $p \in [0,1]$ as the probability of rain. In these plots, $q \in [0,1]$ is on the horizontal axis and expected score is on the vertical axis.



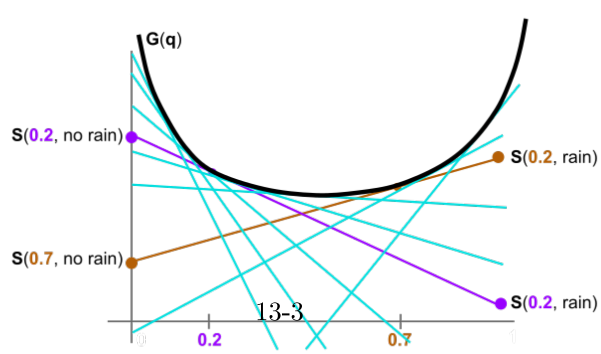
(a) Take for example $p = 0.2$ and plot $S(0.2; q)$ as a function of q. Note that $S(0.2; 0.0) = S(0.2, \text{no rain})$ because the person who believes $q = 0$ believes it will not rain for sure; similarly, $S(0.2; 1.0) = S(0.2, \text{rain})$.



(b) We can do the same with report $p = 0.7$ and $S(07; q)$. Note that with $q = 0.7$, for the scoring rule to be proper (truthful), we should have $S(0.7; q) \geq S(0.2; q)$ and the opposite for $q = 0.2$.



(c) Similarly for all the other reports.



(d) In the end, this traces out a convex function, where for example the linear function $S(0.2; q)$ is tangent at $q = 0.2$ and lies everywhere below the function.

We will prove the following two propositions, which gives a beautiful and complete answer to the key question of how to design proper scoring rules.

Proposition 2 *If $f : \Delta_{\mathcal{Y}} \to \mathbf{R}$ is any convex function, then this is a proper scoring rule:*

$$S(p, y) = f(p) + \nabla f(p) \cdot (\delta_y - p)$$

where δ_y is the probability distribution putting probability 1 on outcome y.

Proof To show that S is proper, we have to show that for any p, q, if someone (call her Pat) has belief q, then she would rather report q than misreport p. First, let us compute the expected score $S(p; q)$ for reporting p with belief q. It is

$$\begin{aligned} S(p; q) &= \sum_y q(y) \left[f(p) + \nabla f(p) \cdot (\delta_y - p) \right] \\ &= f(p) + \nabla f(p) \cdot (q - p). \end{aligned}$$

(You can take my word for it, or work out the details yourself; the key point is that the sum can go into the right side of the dot-product.)

So if Pat reports q, she gets expected score

$$\begin{aligned} S(q; q) &= f(q) + \nabla f(q) \cdot (q - q) \\ &= f(q). \end{aligned}$$

If she reports p, she gets expected score

$$S(p; q) = f(p) + \nabla f(p) \cdot (p - q).$$

Now, recall that because f is a convex function, we have $f(q) \geq f(p) + \nabla f(p) \cdot (p - q)$.

Note: In the proof we assumed that f is differentiable, but actually the same proof works regardless by using what is called a subgradient in place of ∇f. ■

Nice! This tells us how to create a proper scoring rule using a convex function f. To get $S(p, y)$, you take the derivative (or gradient) of f at p and plug in δ_y to get $f(p) + \nabla f(p) \cdot (\delta_y - p)$.

But is this the *only* way to construct proper scoring rules? As it turns out, the answer is yes:

Proposition 3 *For any proper scoring rule S, there exists a convex function f such that*

$$S(p, y) = f(p) + \nabla f(p) \cdot (\delta_y - p).$$

Proof Suppose that S is proper. Define $f(q) = S(q; q)$.

Now, by definition, $S(p; q) = \sum_y q(y) S(p, y)$. Notice this is a *linear* function of q, for any fixed p. Furthermore, this linear function is equal to f at q. Furthermore, we have $S(q; q) \geq S(p; q)$ for all p.

So this linear function $S(p; q)$, as a function of q, lies below f everywhere and is tangent to f at q. So it can be written $S(p; q) = f(p) + \nabla f(p) \cdot (q - p)$.

This holds for every q, and the fact that all of these tangent linear functions lie below f implies that f is a convex function.

Note: again, technically $\nabla f(p)$ above should be some subgradient of f at p. If f is differentiable, then $\nabla f(p)$ is the only subgradient at p. ■

In class, we went through this proof idea in pictures. While I don't expect everyone to remember the mathematical proof, you should have some idea of what the pictures mean and why it is true.

Notice in the theorems above, $f(q) = S(q; q)$. So $f(q)$ is the expected score that the agent gets, when they believe q and truthfully report. Notice that some beliefs have lower expected scores than others! In fact, because f is convex, this says it is generally preferable to be more certain than to be uncertain. We can collect these facts and Propositions 2 and 3 into one big "characterization" theorem.

Theorem 4 (McCarthy 1956; Savage 1971) *A scoring rule S is strictly proper if and only if there exists some convex function f such that:*

1. $S(q;q) = f(q)$, *and*
2. $S(p;q) = f(p) + \nabla f(p) \cdot (q - p)$, *and*
3. $S(p,y) = f(p) + \nabla f(p) \cdot (\delta_y - p)$.

Example. One proper scoring rule is the log scoring rule: $S(p,y) = \log p(y)$.

Let's check that it's proper. The expected score is $S(p;q) = \sum_y q(y) \log p(y)$. The expected score for truthfulness is

$$\begin{aligned} f(q) &= S(q;q) \\ &= \sum_y q(y) \log q(y). \end{aligned}$$

(You may have seen this expression – it's the negative of the Shannon entropy of q!)

Now, we can check the scoring rule characterization: given f, we would define $S(p,y)$ as follows.[1]

$$\begin{aligned} S(p,y) &= f(p) + \nabla f(p) \cdot (\delta_y - p) \\ &= \sum_y p(y) \log p(y) + \nabla f(p) \cdot \delta_y - \nabla f(p) \cdot p \\ &= \sum_y p(y) \log p(y) + 1 + \log p(y) - 1 - \sum_y p(y) \log p(y) \\ &= \log p(y). \end{aligned}$$

Now we can check for ourselves that f is proper (although we know it must be, since it satisfies the characterization theorem since $f(q) = S(q;q)$ is convex). Here is one way. So the difference is

$$\begin{aligned} S(q;q) - S(p;q) &= \sum_y q(y) \left(\log q(y) - \log p(y)\right) \\ &= \sum_y q(y) \log \frac{q(y)}{p(y)}. \end{aligned}$$

This is actually equal to what's called the KL-divergence or relative entropy between q and p, written $KL(q,p)$. It is known to be nonnegative, that is, $KL(p,q) \geq 0$. So $S(q;q) - S(p;q) \geq 0$, so $S(q;q) \geq S(p;q)$. This means S is proper.

You will see another example in the homework.

[1] Note $\nabla f(p)$ is a vector where the entry corresponding to y is $\frac{\partial f}{\partial p(y)} = 1 + \log p(y)$, while δ_y is a vector of all zeros except a 1 at the position corresponding to y.