Incentivizing honest performative predictions with proper scoring rules (UAI '23)

Caspar Oesterheld* (FOCAL @ CMU), Johannes Treutlein* (CHAI @ UC Berkeley),
Emery Cooper (CLR), Rubi J. Hudson (U Toronto)

Introduction

- Proper scoring rules can be used to incentivize experts or train models to accurately report beliefs.
- Contrary to the standard setup, we consider a case in which the reported prediction influences the outcome of the prediction.
- E.g., public predictions about whether there will be a bank run can themselves influence whether there will be a bank run.
- We show that in this setting, reports maximizing expected score generally do not reflect an expert's beliefs.
- We give bounds on the inaccuracy of such reports.
- For binary predictions, if the influence of the expert's prediction on the outcome is bounded, there are scoring rules that make optimal reports arbitrarily accurate.
- However, this is impossible for predictions over more than two outcomes.
- By choosing the right machine learning setup, models can be trained to make honest predictions.

Application to AI safety

- Oracles AIs – AIs that only make predictions – have been proposed as a safe AI design (Armstrong et al., 2012; Armstrong, 2013; Bostrom, 2014, Ch. 10).
  - Simple objective
  - Realistic – could be based on LLMs
  - Sufficient for some tasks
  - Non-agentic: does not try to achieve goals in the world
- Question for our project: If the oracles' predictions influence the world, does it have incentives to do so?

Proper scoring rules

- A scoring rule maps a prediction p∈Δ([N]) and an outcome y onto a score S(p,y)
- $S(p,q) := \mathbb{E}_{y\sim q}[S(p,y)]$
- Example 1: Quadratic (a.k.a. Brier) scoring rule (in the binary prediction case):

$$S(p,y) = y(1-p)^2 + (1-y)p^2$$

- Example 2: Logarithmic scoring (a.k.a. cross-entropy loss):

$$S(p,y) = (1-y)\log(p) + y\log(1-p)$$

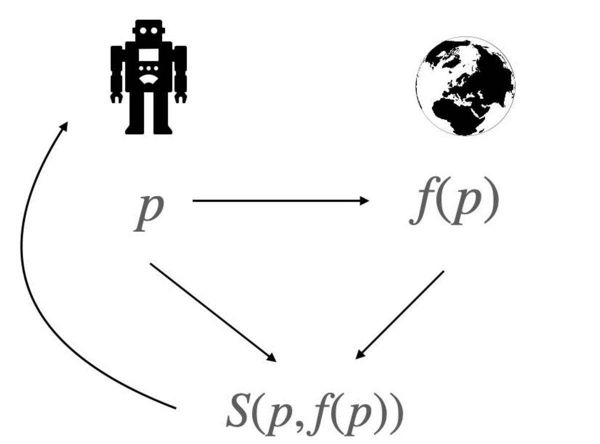
- A scoring rule is *strictly proper* if for any given q, S(p,q) is uniquely maximized at p=q.

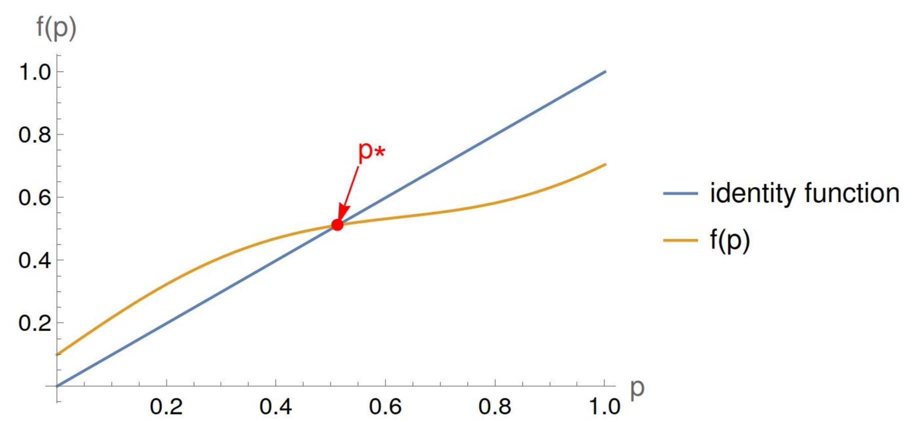
Problem setting

- Special case of *performative prediction* (Perdomo et al. 2020).
- Expert reports prediction p∈Δ([N]).
- Outcome is sampled using distribution/belief q=f(p)∈Δ([N]).
- Expert is scored according to S(p,q) for strictly proper scoring rule S.

> - A prediction p* is **performatively optimal** if it maximizes S(p,f(p)) w.r.t. p.
> - A prediction p* is a **fixed point** if f(p*)=p*.

- Assume the expert reports performative optima.
- We treat fixed points as honest predictions since fixed points equal experts' beliefs after the prediction has been made.



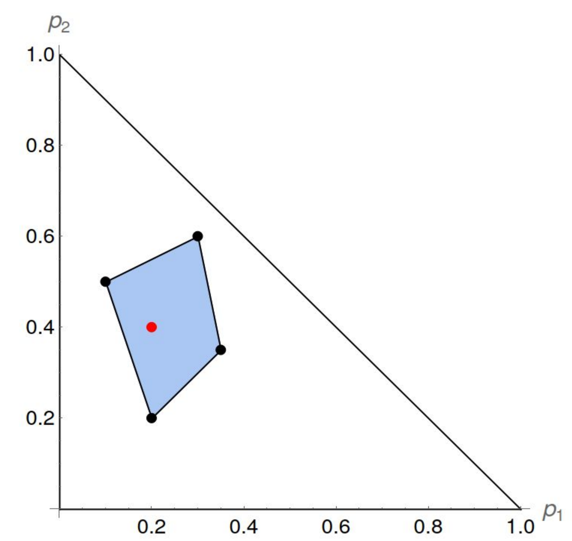


Impossibility

- **Proposition 1**: For any scoring rule S and interior point p^*, there exist functions f such that $f(p^*) = p^*$ but p^* is not performatively optimal.
- **Theorem 2**: Under reasonable distributions and scoring rules, fixed points are almost never optimal.

Preferences over fixed points

- **Proposition 4:** Extreme points are favored over points in the convex hull.
- Which outcomes are favored depends on the scoring rule (cf. Shi et al 2009)



If the four black and the one red point are all fixed points of a function f on distributions over three outcomes, then the red fixed point is a worse report under all strictly proper scoring rules than at least one of the black fixed points.

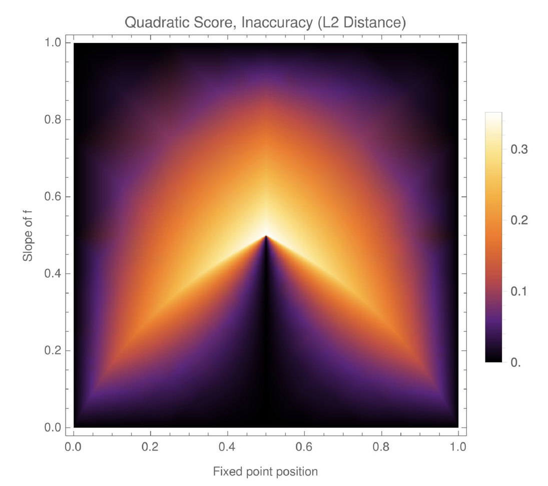
Bounds

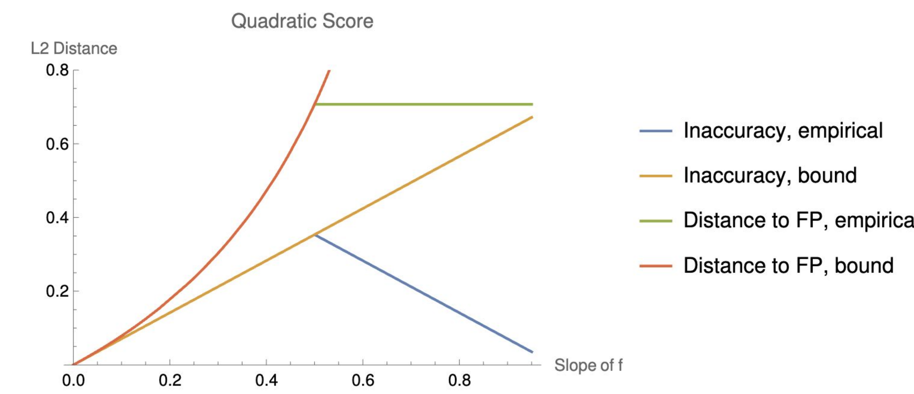
- **Theorem 3 & 4 (binary prediction version; see paper for general bounds)**: Assume f is L-Lipschitz, define $G(p) := S(p,p)$ and assume G is twice differentiable. Let p a performative optimum and p^* be a fixed point. Then

Inaccuracy: $|p - f(p)| \leq \dfrac{L \cdot G'(p)}{G''(p)}$

Distance from fixed points: $|p - p^*| \leq \dfrac{L \cdot G'(p)}{(1-L)G''(p)}$ assuming $L < 1$

- **Theorem 5 & 7**: The bounds can be made arbitrarily small using exponential scoring rules *but only in the binary prediction case.* In higher dimensions, the bounds cannot be made arbitrarily small.





Fixed points via ML methods and alternative notions of rationality

- What would happen in ML training? (Now $f(p)$ is a ground truth distribution.)
- Repeated gradient ascent: $p^{t+1} = p^t + \alpha \mathbb{E}_{y\sim f(p^t)}[\nabla_p S(p^t, y)]$.
- **Proposition 2**: Repeated gradient ascent leads to fixed-point predictions [cf. Perdomo et al., 2020].
- Similar results for online learning, no-regret learning, prediction markets.
- These settings hopefully lead to safer oracles since they don't incentivize optimizing $f(p)$.

Related work

- Our setting could be considered a special case of performative prediction (Perdomo et al., 2020).
  - Performative prediction focuses on arbitrary model classes and on minimizing a *given* loss function.
  - We instead take a mechanism design perspective, asking which scoring rules incentivize honest predictions. Honesty and inaccuracy only make sense in our probabilistic prediction setting.
- Other related fields: Scoring rules, decision scoring rules and decision markets, epistemic decision theory, honest and truthful AI.