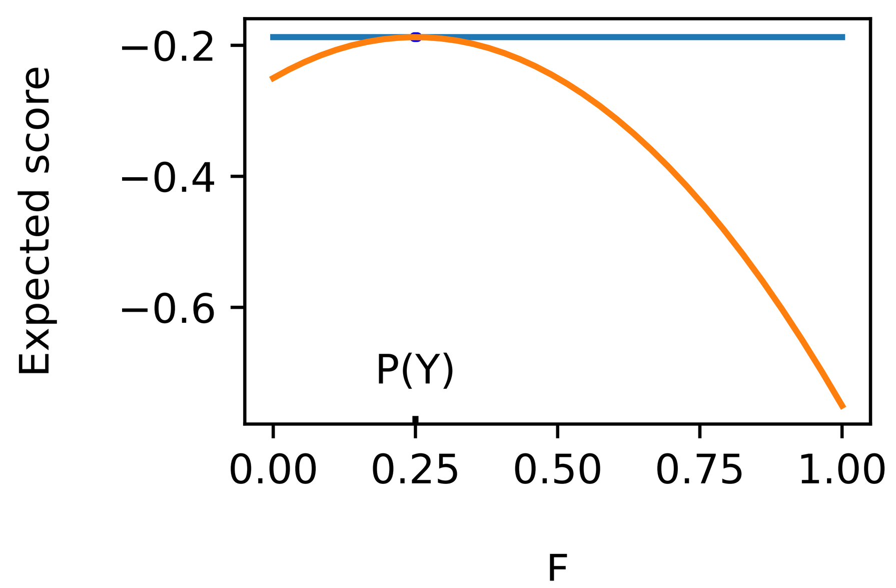
[1] University of Amsterdam

# Conditional Forecasts and Proper Scoring Rules for Reliable and Accurate Performative Predictions

[2] Booking.com

Philip Boeken[1, 2] Onno Zoeter[2] Joris M. Mooij[1]

## Scoring rules




*Example with Brier score* [1]:

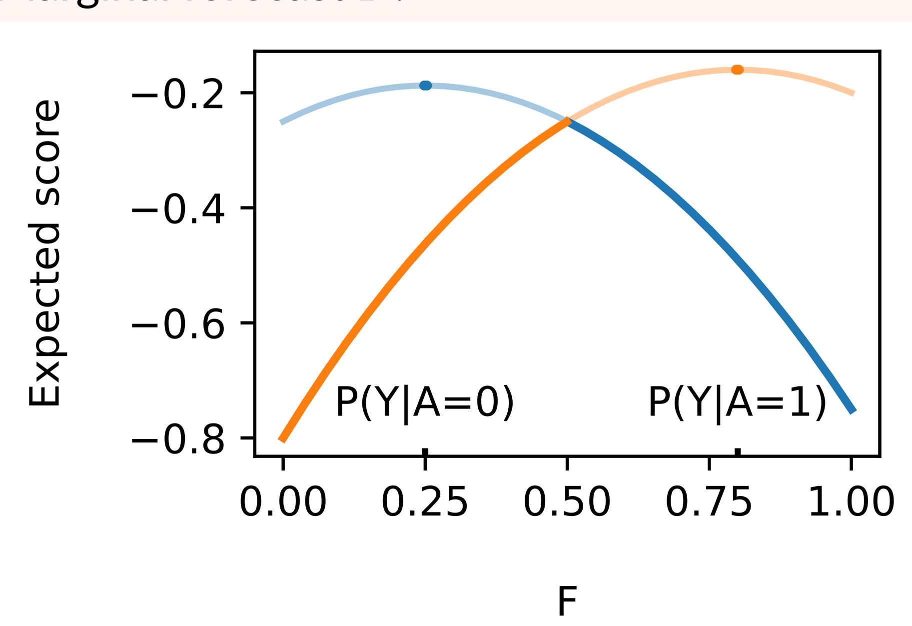
$$\mathcal{Y} = \{0,1\}$$
$$F \in [0,1]$$
$$P(Y=1) = 0.25$$
$$S(F,y) = -(F-y)^2$$
$$\mathbb{E}_P[S(F,Y)] = -(P-F)^2 - P(1-P)$$

$S$ is *proper* if $\mathbb{E}_P[S(P,Y)] \geq \mathbb{E}_P[S(F,Y)]$ for all $F \neq P$, and *strictly proper* if the inequality is strict.

## Problem: Scoring rules are not proper in performative setting

Performative setting: $F \to A \to Y$

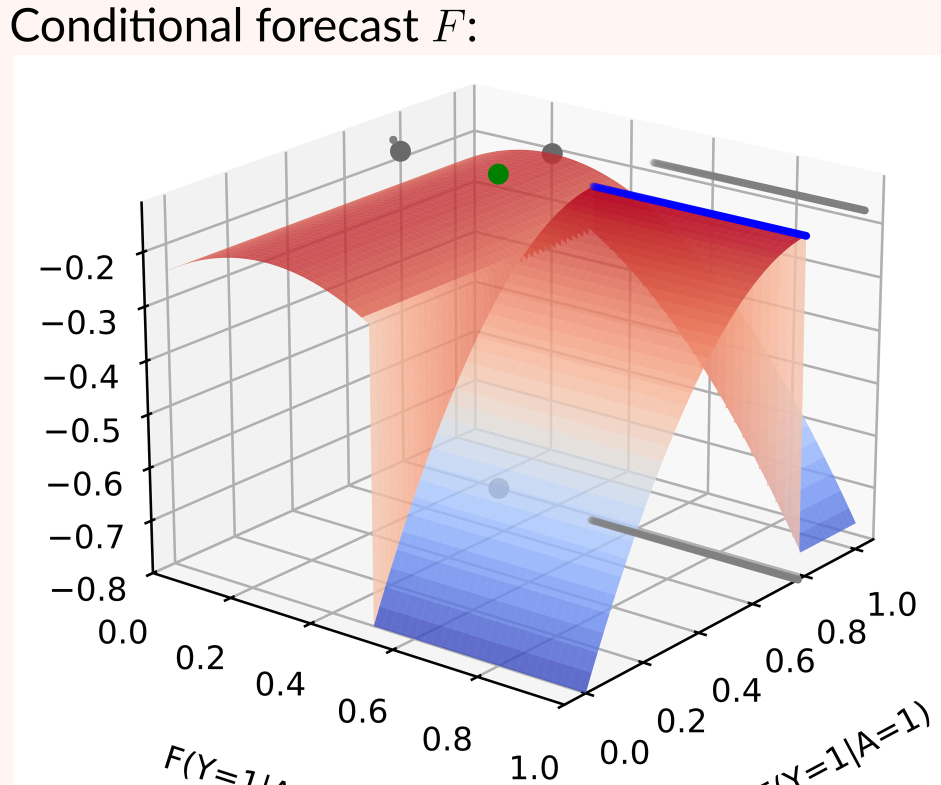
Marginal forecast $F$:




$$F \in [0,1]$$
$$A = \begin{cases} 0 & \text{if } F < 0.5 \\ 1 & \text{otherwise} \end{cases}$$
$$P(Y=1 \mid A) = \begin{cases} 0.25 & \text{if } A = 0 \\ 0.8 & \text{if } A = 1 \end{cases}$$

Correct forecast does not exist!

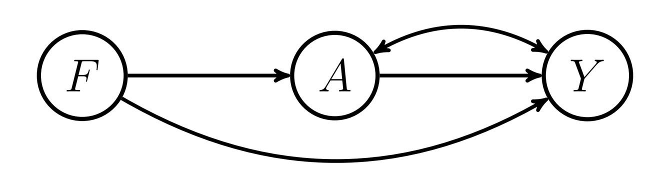
Conditional forecast $F$:




$$F = \begin{pmatrix} F(Y=1 \mid A=0) \\ F(Y=1 \mid A=1) \end{pmatrix} \in [0,1]^2$$
$$A = \begin{cases} 0 & \text{if } F_0 < 0.5 \text{ or } F_1 > 0.775 \\ 1 & \text{otherwise} \end{cases}$$
$$P(Y=1 \mid A) = \begin{cases} 0.25 & \text{if } A = 0 \\ 0.8 & \text{if } A = 1 \end{cases}$$

Correct forecast exists, but is not optimal!

## The formal framework




**Definition**: Let $\mathcal{M}$ be a set of SCMs such that $G_{[F,A,Y]}(M)$ is a subgraph of the graph above for all $M \in \mathcal{M}$.

**Theorem**: Let $\mathcal{M}$ be *the* set of SCMs with common graph $G$ satisfying the above definition. There exists a correct forecast for $Y|A$ for all $M \in \mathcal{M}$ iff $Y \perp^d_G F \mid A$.
This is assumed in all subsequent results.

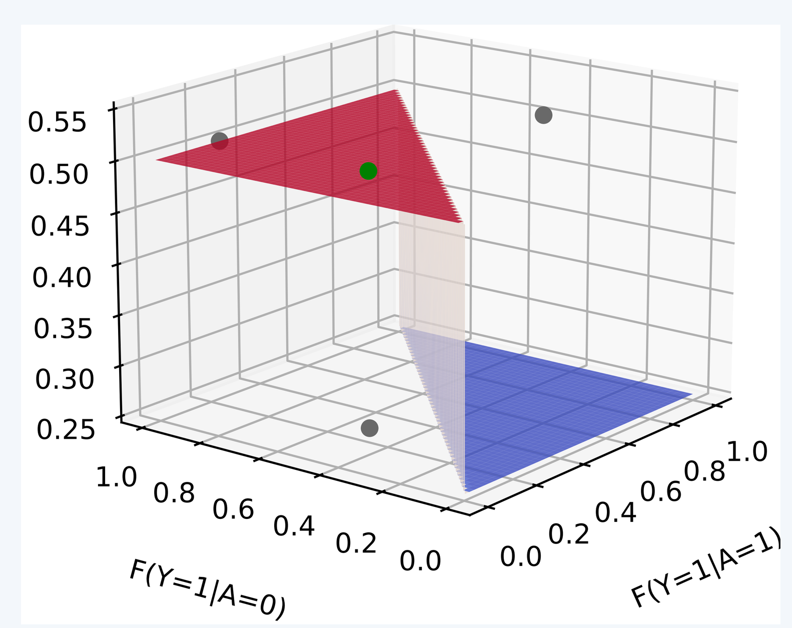
**Definition**: We call a *conditional scoring rule* $S(F,a,y)$

- *observationally (strictly) proper* relative to $\mathcal{M}$ if for all $M \in \mathcal{M}$, $\mathbb{E}_M[S \mid \mathrm{do}(F)]$ is (only) maximised at $F$ with $F(Y \mid A=a) = P_M(Y \mid A=a, \mathrm{do}(F))$ for all $a \in \mathcal{A}$ such that $P_M(A=a \mid \mathrm{do}(F)) > 0$;
- *counterfactually (strictly) proper* relative to $\mathcal{M}$ if for all $M \in \mathcal{M}$, $\mathbb{E}_M[S \mid \mathrm{do}(F)]$ is (only) maximised at $F$ with $F(Y \mid A=a) = P_M(Y \mid A=a, \mathrm{do}(F))$ for all $a \in \mathcal{A}$ such that $P_M(A=a \mid \mathrm{do}(F)) = 0$;
- *(strictly) proper* relative to $\mathcal{M}$ if for all $M \in \mathcal{M}$, $\mathbb{E}_M[S \mid \mathrm{do}(F)]$ is (only) maximised at $F$ with $F(Y \mid A=a) = P_M(Y \mid A=a, \mathrm{do}(F))$ for all $a \in \mathcal{A}$.

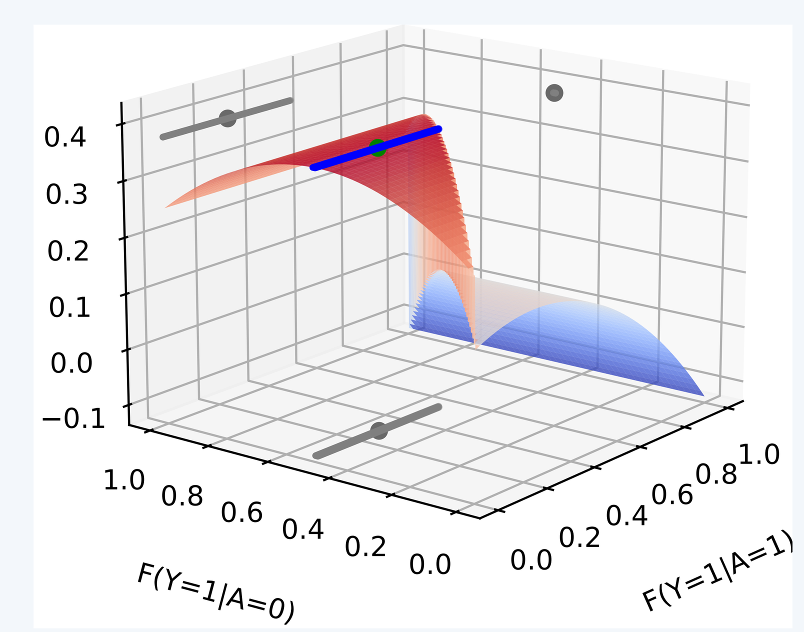
**Theorem**: Let $\mathcal{M}$ be the set of models such that $Y \perp^d_G F \mid A$. Let $S$ be a classically strictly proper scoring rule, then $S$ is not strictly proper for predicting $Y|A$ in $\mathcal{M}$.

## Solution 1: Decision-theoretic setting

**Setting**: Utility $U : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, deterministic decision rule $F \mapsto A$ [2, 3, 4].




**Theorem** [3]:
Utility score $S(F,y) := U(a_F, y)$
observationally proper,
counterfactually proper,
incentive compatible.




**Theorem**:
$S(F,y) = U(a_F,y) + \Delta \cdot S'(F_{a_F}, y)$
observationally *strictly* proper,
counterfactually proper,
incentive compatible.

**Definition**: Conditional scoring rule $S$ is *incentive compatible* with $U$ if

$$\arg\max_F \mathbb{E}_M[S \mid \mathrm{do}(F)] = \arg\max_F \mathbb{E}_M[U \mid \mathrm{do}(F)].$$

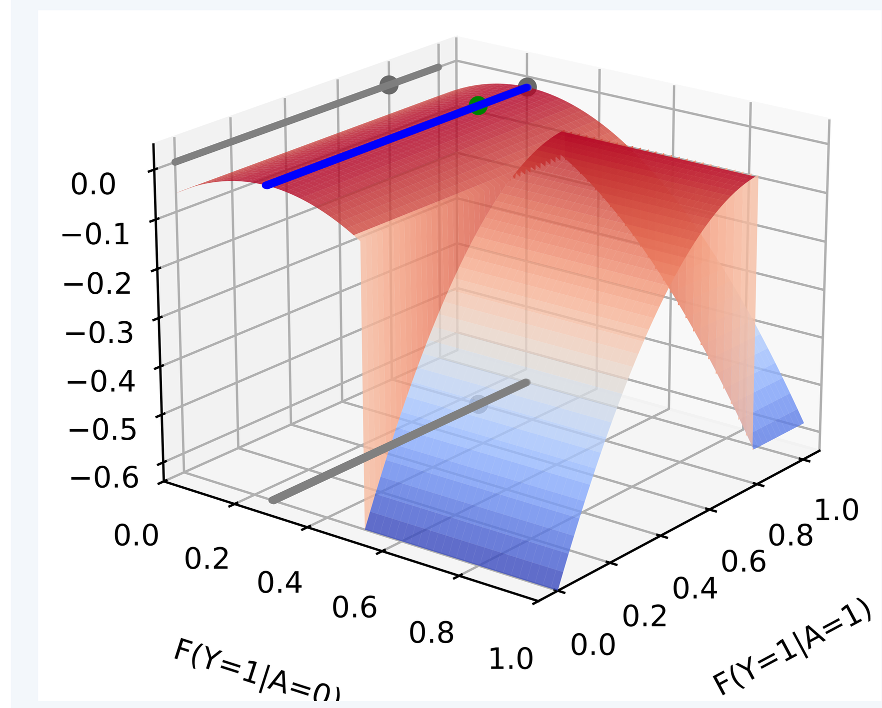
## Solution 2: Performative divergence

In the classical setting, divergence is given by

$$D(F,P) := \mathbb{E}_P[S(F,Y)] - \mathbb{E}_P[S(P,Y)].$$

This generalises to the *performative divergence*

$$D(F,M) := \mathbb{E}_M[S(F,Y) \mid \mathrm{do}(F)] - \mathbb{E}_M[S(P_M(Y \mid A, \mathrm{do}(F)), Y) \mid \mathrm{do}(F)].$$




**Theorem**
Performative divergence $D$:
observationally strictly proper,
counterfactually proper.

**Corollary**
$(A_1,Y_1),\ldots,(A_n,Y_n) \sim P_M(A,Y \mid \mathrm{do}(F))$
Unbiased estimator $\hat{D}(A_1,\ldots,Y_n)$:
observationally strictly proper,
counterfactually proper.

## Parameter estimation

Let $F_\theta$ be a parametrised forecast, then in line with [5] we have

$$R(\theta_{t+1}, \theta_t) := \mathbb{E}_M[S(F_{\theta_{t+1}}, Y) \mid \mathrm{do}(F_{\theta_t})] \quad \textit{decoupled performative score}$$
$$R(\theta) := \mathbb{E}_M[S(F_\theta, Y) \mid \mathrm{do}(F_\theta)] = R(\theta,\theta) \quad \textit{performative score}$$
$$\theta_{PS} := \arg\min_\theta R(\theta, \theta_{PS}) \quad \textit{performatively stable}$$
$$\theta_{PO} := \arg\min_\theta R(\theta) \quad \textit{performatively optimal}$$

*Performative divergence* $R_D(\theta)$ and *decoupled performative divergence* $R_D(\theta_{t+1}, \theta_t)$ are defined similarly.

**Theorem**: If $P_M(A \mid \mathrm{do}(F))$ has full support for all $F$ and $M$, then for any given $\theta_t$, the re-trained parameter $\theta_{t+1} := \arg\min_\theta R_D(\theta, \theta_t)$:

- yields a correct forecast,
- is performatively stable and
- performatively optimal with respect to $R_D(\theta)$.


[1] Glenn W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3, January 1950.

[2] Yiling Chen, Ian Kash, Mike Ruberry, and Victor Shnayder. Decision Markets with Good Incentives. In Ning Chen, Edith Elkind, and Elias Koutsoupias, editors, *Internet and Network Economics*, pages 72–83, Berlin, Heidelberg, 2011. Springer.

[3] Caspar Oesterheld and Vincent Conitzer. Decision Scoring Rules. In *Web and Internet Economics: 16th International Conference, WINE 2020*, 2020.

[4] Abraham Othman and Tuomas Sandholm. Decision Rules and Decision Markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, 2010.

[5] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7599–7609. PMLR, November 2020.


p.a.boeken@uva.nl