Incentivizing honest performative predictions with proper scoring rules (Supplementary material)

Caspar Oesterheld[*1] **Johannes Treutlein**[*2] **Emery Cooper**[3] **Rubi Hudson**[4]

[1]Carnegie Mellon University
[2]University of California, Berkeley
[3]Center on Long-Term Risk
[4]University of Toronto

A PROOFS

A.1 PRELIMINARIES

We begin by proving a lemma characterizing the gradient $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p})))$, which we will use throughout.

Lemma 1. *Assume G, g, f are differentiable. Then*

$$\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p}))) = Dg(\boldsymbol{p})^\top (f(\boldsymbol{p}) - \boldsymbol{p}) + Df(\boldsymbol{p})^\top g(\boldsymbol{p}).$$

If S is strictly proper and $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ an optimal report, then

$$(\boldsymbol{p} - f(\boldsymbol{p}))^\top Dg(\boldsymbol{p}) = g(\boldsymbol{p})^\top Df(\boldsymbol{p}).$$

Proof. We have

$$\begin{aligned}
\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p}))) &= \nabla_{\boldsymbol{p}} \left(G(\boldsymbol{p}) + g(\boldsymbol{p})^\top (f(\boldsymbol{p}) - \boldsymbol{p}) \right) && (1)\\
&= g(\boldsymbol{p}) + Dg(\boldsymbol{p})^\top (f(\boldsymbol{p}) - \boldsymbol{p}) + Df(\boldsymbol{p})^\top g(\boldsymbol{p}) - Ig(\boldsymbol{p}) && (2)\\
&= Dg(p)^\top (f(\boldsymbol{p}) - \boldsymbol{p}) + Df(\boldsymbol{p})^\top g(\boldsymbol{p}). && (3)
\end{aligned}$$

Next, if $\boldsymbol{p}$ is an optimal report and an interior point, it must be $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p})))^\top \boldsymbol{v} = 0$ for any $\boldsymbol{v} \in \mathcal{T}$. Since $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p}))) \in \mathcal{T}$, it follows that $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p}))) = 0$. Hence, using the above, it follows that

$$\begin{aligned}
0 = \nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p}))) &= Dg(p)^\top (f(\boldsymbol{p}) - \boldsymbol{p}) + Df(\boldsymbol{p})^\top g(\boldsymbol{p}) && (4)\\
\Rightarrow Dg(\boldsymbol{p})^\top (\boldsymbol{p} - f(\boldsymbol{p})) &= Df(\boldsymbol{p})^\top g(\boldsymbol{p}). && (5)
\end{aligned}$$

□

A.2 PROOF OF PROPOSITION 1

Proposition 1. *Let S be any strictly proper scoring rule. For any interior fixed point $\boldsymbol{p}^* \in \operatorname{int}(\Delta(\mathcal{N}))$ there exists a function f with Lipschitz constant $L_f < 1$ and a unique fixed point at $\boldsymbol{p}^*$, such that there exists $\boldsymbol{p}' \neq \boldsymbol{p}^*$ with $S(\boldsymbol{p}', f(\boldsymbol{p}')) > S(\boldsymbol{p}^*, f(\boldsymbol{p}^*))$. That is, the unique fixed point of f is not performatively optimal.*

Proof. To begin, let $\boldsymbol{p}^* \in \operatorname{int}(\Delta(\mathcal{N}))$ arbitrary and define $f_\alpha(\boldsymbol{p}) := (1-\alpha)\boldsymbol{p} + \alpha \boldsymbol{p}^*$ for $\alpha \in [0, 1]$ and $\boldsymbol{p} \in \Delta(\mathcal{N})$. Note that since $\Delta(\mathcal{N})$ is convex, $f_\alpha(\boldsymbol{p}) \in \Delta(\mathcal{N})$. Let G be as in the Gneiting and Raftery characterization of S (Theorem 1).

[*]Equal contribution

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

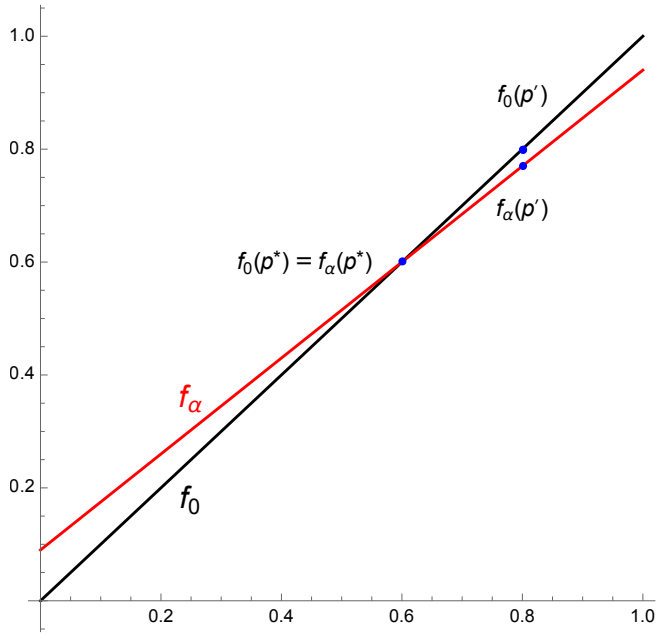


Figure 1: Illustration of the setup for our proof. We plot f_0 in black and f_α for $\alpha = 0.15$ in red, projected onto a single dimension.

To provide an intuition of how our proof will work, consider a binary prediction setting with f as given in Figure 1. For any $\alpha > 0$, f_α has a unique fixed point at $\boldsymbol{p}^*$, while f_0 is the identity function, so all points are fixed points of f_0. By strict convexity of G, there exists a point $\boldsymbol{p}'$ which receives a strictly higher score than $\boldsymbol{p}^*$ if it is a fixed point, so $S(\boldsymbol{p}', f_0(\boldsymbol{p}')) > S(\boldsymbol{p}^*, f_0(\boldsymbol{p}^*))$. $\boldsymbol{p}'$ is not a fixed point of f_α for $\alpha > 0$. However, we will show that $S(\boldsymbol{p}', f_\alpha(\boldsymbol{p}'))$ must be continuous in α, which means that we can choose a small enough $\alpha > 0$ to make sure that $\boldsymbol{p}'$ remains preferable over $\boldsymbol{p}^*$, i.e., $S(\boldsymbol{p}', f_\alpha(\boldsymbol{p}')) > S(\boldsymbol{p}^*, f_\alpha(\boldsymbol{p}^*))$, despite it not being a fixed point.

To formalize the proof, begin by noting that

$$\|f_\alpha(\boldsymbol{p}) - f_\alpha(\boldsymbol{p}')\| = \|(1-\alpha)(\boldsymbol{p} - \boldsymbol{p}')\| = (1-\alpha)\|\boldsymbol{p} - \boldsymbol{p}'\|$$

for any $\boldsymbol{p}, \boldsymbol{p}' \in \Delta(\mathcal{N})$, so f_α has Lipschitz constant $L := (1-\alpha) < 1$, and as mentioned, $\boldsymbol{p}^*$ is the unique fixed point of f_α.

Now consider the case $\alpha = 0$. As mentioned, every point is a fixed point of f_0. Then by strict convexity of G, since $\boldsymbol{p}^*$ is an interior point, there exists another interior point $\boldsymbol{p}' \in \text{int}(\Delta(\mathcal{N}))$ and $\epsilon > 0$ such that $G(\boldsymbol{p}') \geq G(\boldsymbol{p}^*) + \epsilon$. It follows that

$$S(\boldsymbol{p}', f_0(\boldsymbol{p}')) = S(\boldsymbol{p}', \boldsymbol{p}') \geq S(\boldsymbol{p}^*, \boldsymbol{p}^*) + \epsilon = S(\boldsymbol{p}^*, f_0(\boldsymbol{p}^*)) + \epsilon. \tag{6}$$

So for $\alpha = 0$, the model prefers to predict $\boldsymbol{p}'$ over $\boldsymbol{p}^*$ and gets at least ϵ additional expected score. Lastly, note that since $\boldsymbol{p}'$ is an interior point as well, it follows that $G(\boldsymbol{p}') < \infty$.

Now we show that the model still prefers to predict $\boldsymbol{p}'$, even for some small $\alpha > 0$. To that end, note that

$$S(\boldsymbol{p}', f_\alpha(\boldsymbol{p}')) = \mathbb{E}_{y \sim f_\alpha(\boldsymbol{p}')}[S(\boldsymbol{p}', y)]$$

is linear in $f_\alpha(\boldsymbol{p}')$, and $f_\alpha(\boldsymbol{p}')$ is affine-linear in α by construction. This means that $S(\boldsymbol{p}, f_\alpha(\boldsymbol{p}))$ is continuous in α. So there must exist some small $\alpha > 0$ such that

$$S(\boldsymbol{p}', f_\alpha(\boldsymbol{p}')) \geq S(\boldsymbol{p}', f_0(\boldsymbol{p}')) - \frac{\epsilon}{2} = S(\boldsymbol{p}', \boldsymbol{p}') - \frac{\epsilon}{2} \tag{7}$$

$$\underset{(6)}{\geq} S(\boldsymbol{p}^*, \boldsymbol{p}^*) + \frac{\epsilon}{2} > S(\boldsymbol{p}^*, \boldsymbol{p}^*) \tag{8}$$

$$= S(\boldsymbol{p}^*, f(\boldsymbol{p}^*)). \tag{9}$$

Choosing α in this way, we can define $f := f_\alpha$, and have thus provided a function that satisfies the statement that we wanted to prove. □

A.3 PROOF OF THEOREM 2

We begin with two lemmas. In the following, we always assume a strictly proper scoring rule S and accompanying functions G, g as in the Gneiting and Raftery characterization (Theorem 1). Moreover, we let $\Pi_{n-1}\colon \mathbb{R}^n \to \mathbb{R}^{n-1}$ be the projection onto $\mathbb{R}^n$, defined via $\Pi_{n-1}\boldsymbol{x} = (x_i)_{1\leq i\leq n-1}$ for $\boldsymbol{x} \in \mathbb{R}^n$. We will not go into issues of measurability in our proofs.

First, we show that if $\boldsymbol{p}^* \in \operatorname{int}(\Delta(\mathcal{N}))$ is a fixed point of f, then either $g(\boldsymbol{p}^*) = 0$ or $Df(\boldsymbol{p})|_{\mathcal{T}}$, i.e., the map

$$Df(\boldsymbol{p})\colon \mathcal{T} \to \mathcal{T}, \boldsymbol{v} \mapsto Df(\boldsymbol{p})\boldsymbol{v},$$

is singular.

Lemma 2. *Let G, g, and f be differentiable. Let $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ be a fixed point of f and a performatively optimal prediction. Then $Df(\boldsymbol{p})|_{\mathcal{T}}$ is singular or $g(\boldsymbol{p}) = 0$.*

Proof. Note that $f(\boldsymbol{p}) \in \Delta(\mathcal{N})$ for all $\boldsymbol{p} \in \Delta(\mathcal{N})$, so $\partial_{\boldsymbol{v}} f(\boldsymbol{p}) = Df(\boldsymbol{p})\boldsymbol{v} \in \mathcal{T}$ for all $\boldsymbol{v} \in \mathcal{T}$. Hence, $Df(\boldsymbol{p})$ defines an automorphism $Df(\boldsymbol{p})|_{\mathcal{T}}$.

It follows from Lemma 1 that $Dg(\boldsymbol{p})^\top(\boldsymbol{p} - f(\boldsymbol{p})) = Df(\boldsymbol{p})^\top g(\boldsymbol{p})$. Since $f(\boldsymbol{p}) - \boldsymbol{p} = 0$, it must be $Df(\boldsymbol{p})^\top g(\boldsymbol{p}) = 0$, so either $g(\boldsymbol{p}) = 0$, or $Df(\boldsymbol{p})^\top$ (and thus also $Df(\boldsymbol{p})$) is singular when restricted to $\mathcal{T}$. □

Next, we show that the fixed points of f are almost surely not at points $\boldsymbol{p}$ such that $g(\boldsymbol{p}) = 0$, under our assumptions on the distribution over f.

Lemma 3. *Let $\mathcal{F} := \{F(\boldsymbol{p})\}_{\boldsymbol{p}\in\operatorname{int}(\Delta(\mathcal{N}))}$ be a stochastic process with values in $\Delta(\mathcal{N})$ and assume that for each $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$, the random vector $\Pi_{n-1}F(\boldsymbol{p})$ has a density $h_{\Pi_{n-1}F(\boldsymbol{p})}$. Then almost surely if $F(\boldsymbol{p}) = \boldsymbol{p}$ for some $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ then $g(\boldsymbol{p}) \neq 0$. That is,*

$$\mathbb{P}(\exists \boldsymbol{p}\colon F(\boldsymbol{p}) = \boldsymbol{p} \wedge g(\boldsymbol{p}) = 0) = 0.$$

Proof. First note that if S is strictly proper, then G is strictly convex and so there exists at most one $\boldsymbol{p} \in \Delta(\mathcal{N})$ with $g(\boldsymbol{p}) = 0$. If there is no such point, then we are done. Otherwise, let that point be $\boldsymbol{p}^*$. Since we assume that $\Pi_{n-1}F(\boldsymbol{p}^*)$ has a density function $h_{\Pi_{n-1}F(\boldsymbol{p}^*)}$, it follows that

$$\mathbb{P}(F(\boldsymbol{p}^*) = \boldsymbol{p}^*) = \mathbb{P}(\Pi_{n-1}F(\boldsymbol{p}^*) = \Pi_{n-1}\boldsymbol{p}^*) = \int_{\{\Pi_{n-1}\boldsymbol{p}^*\}} h_{\Pi_{n-1}F(\boldsymbol{p}^*)}(\boldsymbol{x})d\boldsymbol{x} = 0.$$

□

Lastly, we require a result about random fields. The following is adapted from Proposition 6.11 in Azaïs and Wschebor [2009].

Proposition 2 (Azaïs and Wschebor [2009], Proposition 6.11)**.** *Let $\mathcal{Y} = \{Y(\boldsymbol{x})\}_{\boldsymbol{x}\in W}$ be a random field with values in $\mathbb{R}^d$ and W an open subset of $\mathbb{R}^{d'}$. Let $\boldsymbol{u} \in \mathbb{R}^d$ and $I \subseteq W$. Assume that*

- *the sample paths $\boldsymbol{x} \rightsquigarrow Y(\boldsymbol{x})$ are continuously differentiable*
- *for each $\boldsymbol{x} \in W$, $Y(\boldsymbol{x})$ has a density $h_{Y(\boldsymbol{x})}$ and there exists a constant C such that $h_{Y(\boldsymbol{x})}(\boldsymbol{y}) \leq C$ for all $\boldsymbol{x} \in I$ and $\boldsymbol{y} \in \mathbb{R}^d$.*
- *The Hausdorff dimension of I is strictly smaller than d.*

Then, almost surely, there is no point $\boldsymbol{x} \in I$ such that $Y(\boldsymbol{x}) = \boldsymbol{u}$.

Now we can turn to the proof of the main result.

Theorem 2. *Let S be a twice differentiable strictly proper scoring rule. Let $\mathcal{F} := \{F(\boldsymbol{p})\}_{\boldsymbol{p}\in\operatorname{int}(\Delta(\mathcal{N}))}$ be a stochastic field with values in $\Delta(\mathcal{N})$ and let $Y(\boldsymbol{p}, \boldsymbol{v}) := (\Pi_{n-1}F(\boldsymbol{p}), \Pi_{n-1}\partial_{\boldsymbol{v}}F(\boldsymbol{p}))$ for $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ and $\boldsymbol{v} \in \mathcal{T} \cap S^{n-1}$. Assume that*

- *the sample paths $\boldsymbol{p} \rightsquigarrow F(\boldsymbol{p})$ are twice continuously differentiable*
- *for each $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ and $\boldsymbol{v} \in \mathcal{T} \cap S^{n-1}$, the random vector $Y(\boldsymbol{p}, \boldsymbol{v})$ has a joint density $h_{Y(\boldsymbol{p},\boldsymbol{v})}$ and there exists a constant C such that $h_{Y(\boldsymbol{p},\boldsymbol{v})} \leq C$ for all $\boldsymbol{p} \in \Delta(\mathcal{N}), \boldsymbol{v} \in S^{n-1} \cap \mathcal{T}$.*

Then, almost surely, there is no point $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ *such that* $\boldsymbol{p} \in \arg\max_{\boldsymbol{p}'} S(\boldsymbol{p}', F(\boldsymbol{p}'))$ *and* $F(\boldsymbol{p}) = \boldsymbol{p}$.

Proof. We want to show that almost surely there does not exist $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ such that $F(\boldsymbol{p}) = \boldsymbol{p}$ and $\boldsymbol{p}$ is performatively optimal. I.e., we want to show that

$$\mathbb{P}(\exists \boldsymbol{p} \colon F(\boldsymbol{p}) = \boldsymbol{p} \wedge \boldsymbol{p} \in \arg\max S(\boldsymbol{p}, F(\boldsymbol{p}))) = 0.$$

First, let $\boldsymbol{p}^* \in \operatorname{int}(\Delta(\mathcal{N}))$ be a performatively optimal report. By Lemma 2, either $g(\boldsymbol{p}^*) = 0$ or $DF(\boldsymbol{p}^*)|_{\mathcal{T}}$ is singular. Moreover, by assumption, $(\Pi_{n-1}F(\boldsymbol{p}), \Pi_{n-1}DF(\boldsymbol{p})\boldsymbol{v})$ has a density function for any $\boldsymbol{v} \in \mathcal{T} \cap S^{n-1}$, and thus also $\Pi_{n-1}F(\boldsymbol{p})$ has one. Hence, by Lemma 3, it follows that if $F(\boldsymbol{p}) = \boldsymbol{p}$ for some $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$, then almost surely $g(\boldsymbol{p}) = 0$.

Second, we need to show that also almost surely $DF(\boldsymbol{p})|_{\mathcal{T}}$ is invertible at any fixed point of F. To that end, define the random field $\mathcal{Y} := \{Y(\boldsymbol{p}, \boldsymbol{v})\}_{(\boldsymbol{p},\boldsymbol{v}) \in W}$ where $W := \operatorname{int}(\Delta(\mathcal{N})) \times \mathcal{T}$ and

$$Y(\boldsymbol{p}, \boldsymbol{v}) := (\Pi_{n-1}F(\boldsymbol{p}) - \Pi_{n-1}\boldsymbol{p}, \Pi_{n-1}DF(\boldsymbol{p})\boldsymbol{v}),$$

with values in $\mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$.

Note that since F is in $\mathcal{C}^2$, DF is continuously differentiable, and thus also Y. Moreover

$$h_{Y(\boldsymbol{p},\boldsymbol{v})}(\boldsymbol{x}, \boldsymbol{y}) = h_{\Pi_{n-1}F(\boldsymbol{p}),\Pi_{n-1}DF(\boldsymbol{p})\boldsymbol{v}}(\boldsymbol{x} + \Pi_{n-1}\boldsymbol{p}, \boldsymbol{y}) \leq C$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n-1}$ by assumption. Finally, define $\boldsymbol{u} := (0, 0) \in \mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$ and $I := \Delta(\mathcal{N}) \times (\mathcal{T} \cap S^{n-1})$, where $\mathcal{T} \cap S^{n-1} = \{\boldsymbol{v} \in \mathcal{T} \mid \|\boldsymbol{v}\| = 1\}$. Note that the Hausdorff dimension of I is $n - 1 + n - 2 = 2n - 3$, while Y's values are $2n - 2$-dimensional.

This shows all conditions of Proposition 2, so we can apply it to Y to conclude that almost surely there exists no $\boldsymbol{p}, \boldsymbol{v} \in I$ such that $Y(\boldsymbol{p}, \boldsymbol{v}) = (0, 0)$. This means that almost surely there exists no point $\boldsymbol{p} \in \Delta(\mathcal{N})$ such that $F(\boldsymbol{p}) = \boldsymbol{p}$ and such that $DF(\boldsymbol{p})|_{\mathcal{T}}$ is singular, since if such a point existed, then also there would be a vector $\boldsymbol{v} \in \mathcal{T} \cap S^{n-1}$ such that $DF(\boldsymbol{p})\boldsymbol{v} = 0$ and thus $\Pi_{n-1}DF(\boldsymbol{p})\boldsymbol{v} = 0$, implying that

$$Y(\boldsymbol{p}, \boldsymbol{v}) = (\Pi_{n-1}F(\boldsymbol{p}) - \Pi_{n-1}\boldsymbol{p}, \Pi_{n-1}DF(\boldsymbol{p})\boldsymbol{v}) = 0.$$

Summarizing our argument, it follows that

$$\mathbb{P}(\exists \boldsymbol{p} \colon F(\boldsymbol{p}) = \boldsymbol{p} \wedge \boldsymbol{p} \in \arg\max S(\boldsymbol{p}, F(\boldsymbol{p})))$$

$$\leq \mathbb{P}(\exists \boldsymbol{p} \colon F(\boldsymbol{p}) = \boldsymbol{p} \wedge g(\boldsymbol{p}) = 0) + \mathbb{P}(\exists \boldsymbol{p} \colon F(\boldsymbol{p}) = \boldsymbol{p} \wedge DF(\boldsymbol{p}) \text{ is singular}) = 0.$$

This concludes the proof. □

We conclude by providing an example of a stochastic process that satisfies our conditions, for the binary prediction case.

Example 5. Consider a Gaussian process $\{X(p)\}_{p \in (0,1)}$ with values in $\mathbb{R}$, with infinitely differentiable kernel and mean functions. We can make it into a process $F(\boldsymbol{p})$ by defining $F(\boldsymbol{p}) = (X(p_1), 1 - X(p_1))$ for $\boldsymbol{p} \in \Delta([2])$. Note that the paths of F are infinitely differentiable and the values of $\Pi_1 F(\boldsymbol{p}) = X(p_1)$ and its directional derivatives $\Pi_1 DF(\boldsymbol{p})\boldsymbol{v} = X'(p_1)v_1$ are jointly Gaussian and thus have a bounded density [see Rasmussen and Williams, 2006, Ch. 9.4]. To deal with the restriction that $X(p) \in [0, 1]$ for $p \in (0, 1)$, we could condition on the event $E := \{\forall p \colon F(p) \in [0, 1]\}$, for instance. Then paths are still twice differentiable, and we claim that $h_{X(p)|E}$, defined as the density of X at point p, conditional on E, is still bounded. To see that, note that if $\mathbb{P}(E) > 0$, then we are done, since then

$$h_{X(p)|E}(x) = \frac{\mathbb{1}_E(x)}{\mathbb{P}(E)} h_{X(p)}(x).$$

We leave it as an exercise to the reader to prove that $\mathbb{P}(E) > 0$.

A.4 PROOF OF THEOREM 3

Theorem 3. *Let S be a strictly proper scoring rule, and let G, g as in the Gneiting and Raftery characterization (Theorem 1). Let $\boldsymbol{p} \in \Delta(\mathcal{N})$ and assume f, G, g are differentiable at $\boldsymbol{p}$. Assume $Dg(\boldsymbol{p})|_{\mathcal{T}} \succeq \gamma_{\boldsymbol{p}}$ for some $\gamma_{\boldsymbol{p}} > 0$. Then whenever $\boldsymbol{p}$ is a performatively optimal report,*

$$\|\boldsymbol{p} - f(\boldsymbol{p})\| \leq \frac{\|Df(\boldsymbol{p})\|_{\mathrm{op}} \|g(\boldsymbol{p})\|}{\gamma_{\boldsymbol{p}}}.$$

In particular, if f has Lipschitz constant L_f, G has Lipschitz constant L_G, and G is γ-strongly convex, then we have $\|\boldsymbol{p} - f(\boldsymbol{p})\| \leq \frac{L_f L_G}{\gamma}$.

Recall that we assume that $g(\boldsymbol{p})$ is normalized to be orthogonal to $\mathbf{1}$. Note that this is also the choice that minimizes $\|g(\boldsymbol{p})\|$ and makes sure that $\|g(\boldsymbol{p})\| = \|g(\boldsymbol{p})^\top|_{\mathcal{T}}\|_{\mathrm{op}}$, where $g(\boldsymbol{p})^\top|_{\mathcal{T}}$ denotes the function $\mathcal{T} \to \mathbb{R} \colon \mathbf{v} \mapsto g(\boldsymbol{p})^\top \mathbf{v}$. This is due to the Cauchy–Schwarz inequality and the Pythagorean theorem, since $\|g(\boldsymbol{p}) + \alpha \mathbf{1}\|^2 = \|g(\boldsymbol{p})\|^2 + |\alpha|^2 \|\mathbf{1}\|^2$ for any $\alpha \in \mathbb{R}$ when $g(\boldsymbol{p}) \in \mathcal{T}$. Moreover, by Cauchy–Schwarz, we have $\|g(\boldsymbol{p})^\top \boldsymbol{v}\| \leq \|g(\boldsymbol{p})\| \|\boldsymbol{v}\|$ for any $\boldsymbol{v} \in \mathcal{T}$ and if $g(\boldsymbol{p}) \in \mathcal{T}$ then $\|g(\boldsymbol{p})^\top|_{\mathcal{T}}\|_{\mathrm{op}} \geq \|g(\boldsymbol{p})^\top g(\boldsymbol{p})\| / \|g(\boldsymbol{p})\| = \|g(\boldsymbol{p})\|$.

Proof. Assume $\boldsymbol{p}$ is a performatively optimal report and that $Dg(\boldsymbol{p})|_{\mathcal{T}} \succeq \gamma_p$. Note that this is equivalent to all eigenvalues of the function $Dg(\boldsymbol{p})|_{\mathcal{T}}$ being at least γ_p, assuming $Dg(\boldsymbol{p})|_{\mathcal{T}}$ is symmetric. Moreover, $Dg(\boldsymbol{p})$ must be symmetric if G is twice differentiable (note that continuous differentiability is not needed since we assume differentiability in general, not just existence of the coordinate partial derivatives). This can be used to calculate our bound in practice.

Consider $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p})))^\top (f(\boldsymbol{p}) - \boldsymbol{p})$, the directional derivative of $\varphi \colon \boldsymbol{p} \mapsto S(\boldsymbol{p}, f(\boldsymbol{p}))$ in the direction $(f(\boldsymbol{p}) - \boldsymbol{p})$. Note that this derivative must be at most zero: The line from $\boldsymbol{p}$ to $f(\boldsymbol{p})$ lies entirely within the probability simplex, and so if the derivative were positive, $S(\boldsymbol{p}, f(\boldsymbol{p}))$ could be increased by moving in the direction of $f(\boldsymbol{p})$ from $\boldsymbol{p}$. By Lemma 1, we know that

$$\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p}))) = Dg(\boldsymbol{p})^\top (f(\boldsymbol{p}) - \boldsymbol{p}) + Df(\boldsymbol{p})^\top g(\boldsymbol{p}).$$

It follows that

$$0 \geq \nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p})))^\top (f(\boldsymbol{p}) - \boldsymbol{p}) = (f(\boldsymbol{p}) - \boldsymbol{p})^\top Dg(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p}) + g(\boldsymbol{p})^\top Df(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p}) \tag{10}$$

$$\Rightarrow - g(\boldsymbol{p})^\top Df(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p}) \geq (f(\boldsymbol{p}) - \boldsymbol{p})^\top (Dg(\boldsymbol{p}))(f(\boldsymbol{p}) - \boldsymbol{p}). \tag{11}$$

Using that $Dg(\boldsymbol{p})|_{\mathcal{T}} \succ \gamma_{\boldsymbol{p}}$ and thus $(f(\boldsymbol{p}) - \boldsymbol{p})^\top (Dg(\boldsymbol{p}))(f(\boldsymbol{p}) - \boldsymbol{p}) \geq \gamma_{\boldsymbol{p}} \|f(\boldsymbol{p}) - \boldsymbol{p}\|^2$, it follows that

$$\begin{aligned}
& \gamma_{\boldsymbol{p}} \|f(\boldsymbol{p}) - \boldsymbol{p}\|^2 \\
\leq\ & (f(\boldsymbol{p}) - \boldsymbol{p})^\top Dg(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p}) \\
\leq\ & -g(\boldsymbol{p})^\top Df(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p}) \\
\leq\ & |g(\boldsymbol{p})^\top Df(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p})| \\
\underset{\text{Cauchy-Schwarz}}{\leq}\ & \|g(\boldsymbol{p})\| \|Df(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p})\| \\
\leq\ & \|g(\boldsymbol{p})\| \|Df(\boldsymbol{p})\|_{\mathrm{op}} \|f(\boldsymbol{p}) - \boldsymbol{p}\|
\end{aligned}$$

Dividing by $\gamma_{\boldsymbol{p}} \|f(\boldsymbol{p}) - \boldsymbol{p}\|$, we get that $\|f(\boldsymbol{p}) - \boldsymbol{p}\| \leq \|Df(\boldsymbol{p})\|_{\mathrm{op}} \|g(\boldsymbol{p})\| / \gamma_{\boldsymbol{p}}$.

For the "in particular" part, note that if f is Lipschitz continuous with constant L_f, then $\|Df(\boldsymbol{p})\|_{\mathrm{op}} \leq L_f$ for all $\boldsymbol{p}$. Moreover, if G is Lipschitz continuous with constant L_G, we have

$$L_G \geq \|DG(\boldsymbol{p})\|_{\mathrm{op}} = \|g(\boldsymbol{p})^\top\|_{\mathrm{op}} = \|g(\boldsymbol{p})\|$$

for all $\boldsymbol{p} \in \Delta(\mathcal{N})$. Here, in the last step, we have used that for the Euclidean norm

$$\|g(\boldsymbol{p})^\top\|_{\mathrm{op}} = \max_{\boldsymbol{v} \in \mathcal{T}} \frac{g(\boldsymbol{p})^\top \boldsymbol{v}}{\|\boldsymbol{v}\|} = \frac{g(\boldsymbol{p})^\top g(\boldsymbol{p})}{\|g(\boldsymbol{p})\|} = \|g(\boldsymbol{p})\|.$$

Lastly, G being γ-strongly convex implies that $D^2 G(\boldsymbol{p}) \succeq \gamma$ for all $\boldsymbol{p} \in \Delta(\mathcal{N})$, and thus also $Dg(\boldsymbol{p}) = D^2 G(\boldsymbol{p})^\top \succeq \gamma$.

Putting everything together, we get

$$\|f(\boldsymbol{p}) - \boldsymbol{p}\| \leq \frac{\|g(\boldsymbol{p})\| \|Df(\boldsymbol{p})\|_{\mathrm{op}}}{\gamma_p} \leq \frac{L_G L_f}{\gamma}$$

for all performatively optimal reports $\boldsymbol{p}$. □

A.5 PROOF OF THEOREM 4

Theorem 4. *Same assumptions as Theorem 3. Assume further that f has Lipschitz constant $L_f < 1$. Let $\boldsymbol{p}^*$ be the unique fixed point of f. Then for the performatively optimal report $\boldsymbol{p}$,*

$$\|\boldsymbol{p} - \boldsymbol{p}^*\| \leq \frac{\|g(\boldsymbol{p})\|\|Df(\boldsymbol{p})\|_{\text{op}}}{(1 - L_f)\gamma_{\boldsymbol{p}}} \leq \frac{L_f L_G}{(1 - L_f)\gamma_{\boldsymbol{p}}}.$$

Proof. For *any* $\boldsymbol{p} \in \Delta(\mathcal{N})$, we have

$$\begin{aligned} \|\boldsymbol{p} - \boldsymbol{p}^*\| &\underset{\text{triangle ineq.}}{\leq} \|\boldsymbol{p} - f(\boldsymbol{p})\| + \|f(\boldsymbol{p}) - \boldsymbol{p}^*\| \\ &\underset{p^* \text{ fixpoint}}{=} \|\boldsymbol{p} - f(\boldsymbol{p})\| + \|f(\boldsymbol{p}) - f(\boldsymbol{p}^*)\| \\ &\leq \|\boldsymbol{p} - f(\boldsymbol{p})\| + L_f\|\boldsymbol{p} - \boldsymbol{p}^*\| \end{aligned}$$

Solving for $\|\boldsymbol{p} - \boldsymbol{p}^*\|$ yields

$$\|\boldsymbol{p} - \boldsymbol{p}^*\| \leq \frac{\|\boldsymbol{p} - f(\boldsymbol{p})\|}{1 - L_f}.$$

Hence, if $\boldsymbol{p} \in \Delta(\mathcal{N})$ is an optimal prediction, it follows by Theorem 3 that

$$\begin{aligned} \|\boldsymbol{p} - \boldsymbol{p}^*\|_2 &\leq \frac{\|\boldsymbol{p} - f(\boldsymbol{p})\|}{1 - L_f} \\ &\leq \frac{\|Df(\boldsymbol{p})\|_{\text{op}}\|g(\boldsymbol{p})\|}{\gamma_{\boldsymbol{p}}(1 - L_f)} \\ &\leq \frac{L_f\|g(\boldsymbol{p})\|}{\gamma_{\boldsymbol{p}}(1 - L_f)}, \end{aligned}$$

which concludes the proof. □

A.6 THERE IS NO NON-TRIVIAL BOUND ON THE DISTANCE TO THE FIXED POINT AS $L_f \to 1$

We here show why Theorem 4 requires that we have some bound $L_f < 1$ on the function f. Specifically, we show that if f can have Lipschitz constants arbitrarily close to 1, then even in the two-outcome case, only the trivial bound on the difference to the fixed point holds. (The trivial bound is $\|\boldsymbol{p}^* - \boldsymbol{p}\| \leq \sqrt{2}$, because any two points in $\Delta(\{1, 2\})$ are at most $\|(0, 1) - (1, 0)\| = \sqrt{2}$ apart.) We prove that this holds even for the binary case.

Proposition 3. *Consider the case of two outcomes, i.e., let $\mathcal{N} = \{1, 2\}$. Let S be any strictly proper scoring rule. Then there exist functions f with Lipschitz constants smaller than 1 such that $\|\boldsymbol{p}^* - \boldsymbol{p}\|$ is arbitrarily close to $\sqrt{2}$, where $\boldsymbol{p}^*$ is the fixed point of f and $\boldsymbol{p}$ is the optimal prediction for S, f.*

We here give some intuition for why the result holds. Recall that, roughly speaking, scoring rules generally induce a preference for extreme honest predictions over non-extreme honest predictions (see Appendix B). In particular, in the binary case any scoring rule must either incentivize near-0 honest predictions or near-1 honest predictions (or both) over honest relatively close-to-uniform predictions. Consider the case where S incentivizes predictions close to 0 over more uniform predictions and take the function f in Figure 2. The unique fixed point is at 0.8. But if a prediction close to 0 is made, the prediction is *approximately* honest while more extreme than 0.8. It turns out that a slight dishonesty (discrepancy between the report $\boldsymbol{p}$ and the true distribution $f(\boldsymbol{p})$) can be outweighed by the fact that the prediction is more extreme. Predicting near 0 may therefore be a better report than a of prediction 0.8.

Note that in this example, the distance of the optimal report to fixed point ($\|\boldsymbol{p} - \boldsymbol{p}^*\|$) and the inaccuracy of the optimal report ($\|\boldsymbol{p} - f(\boldsymbol{p})\|$) come apart: The optimal report might be far from the fixed point but still very accurate.

Proof. For notational convenience, we consider functions $f : [0, 1] \to [0, 1]$ on a single probability and similarly scoring rules $S : [0, 1] \times [0, 1] \to \mathbb{R}$.

Let ζ be a small positive real number and let δ be s.t. $0 < \delta < \zeta/2$. By the strict convexity of the function $x \mapsto S(x, x)$, one of the following must be the case:

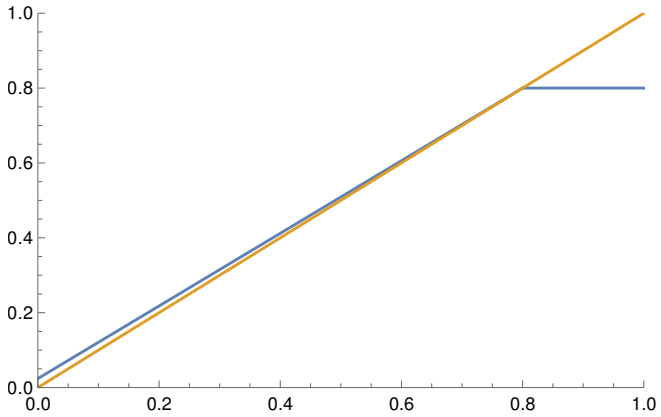


Figure 2: The blue line is the function used in the proof of Proposition 3. The orange line is the identity function.

1. $S(\delta, \delta) > \max_{x \in [2\delta, 1-2\delta]} S(x, x)$; or
2. $S(1-\delta, 1-\delta) > S(x, x)$ for all x between 2δ and $1-2\delta$.

Consider the first case: Now for small positive ϵ consider the function f_ϵ that starts at some small positive value and increases linearly at rate $1-\epsilon$ from 0 to $1-\zeta$ and is fixed at value $1-\zeta$ from $1-\zeta$ to 1. Figure 2 illustrates this function for $\zeta = 0.2, \epsilon = 0.03, \delta = 0.1$. Formally, $f_\epsilon(p) = 1-\zeta$ for $p \geq \zeta$ and otherwise $f_\epsilon(p) = 1-\zeta-(1-\zeta-p)(1-\epsilon)$. Note that f_ϵ's fixed point is $1-\zeta$ and f_ϵ's Lipschitz constant is $1-\epsilon$. We will show that for small enough ϵ the optimal report for f_ϵ and S is then close to 0 and thus almost 1 away from the fixed point, which means the distance is close to $\sqrt{2}$ in the simplex. Note that by continuity of f_ϵ and linearity (and thus continuity) of $S(p, q)$ in q, we have that $S(\delta, f_\epsilon(\delta)) \to S(\delta, \delta)$ as $\epsilon \to 0$. Thus, for small enough ϵ, we have for all x between 2δ and $1-2\delta$ that $S(\delta, f_\epsilon(\delta)) > S(x, x)$. It follows that the optimal report p cannot have $f_\epsilon(p) \in [2\delta, 1-2\delta] \supset [\zeta, 1-\zeta]$, because then we would have that $S(\delta, f_\epsilon(\delta)) > S(f_\epsilon(p), f_\epsilon(p)) > S(p, f_\epsilon(p))$, i.e., δ would be a better report. By construction of f_ϵ, this means that the optimal report cannot be in $[2\delta, 1]$. Thus, the distance of the optimal report to the fixed point is at least $1-\zeta-2\delta$. By choosing δ and ζ to be small, we can make this arbitrarily close to 1.

The second case can be considered analogously, by considering a function f that is constant at value ζ from 0 to ζ and then increases linearly at rate $1-\epsilon$. □

A.7 PROOF OF THEOREM 5

Theorem 5. *Consider the case of two outcomes, i.e., let $\mathcal{N} = \{1, 2\}$. Let $L_f \in \mathbb{R}$ and $\epsilon > 0$. Then there exists a scoring rule S s.t. under any f with Lipschitz constant L_f, any optimal report $\boldsymbol{p}$ satisfies $\|\boldsymbol{p} - f(\boldsymbol{p})\| \leq \epsilon$. If $L_f < 1$, then there also exists a scoring rule that additionally ensures that under any f with Lipschitz constant L_f, any optimal report satisfies $\|\boldsymbol{p} - \boldsymbol{p}^*\| \leq \epsilon$, where $\boldsymbol{p}^*$ is the (unique) fixed point of f.*

Proof. Consider the exponential scoring rule defined by $G(\boldsymbol{p}) = \frac{2}{K} e^{Kp_1}$ and $g(\boldsymbol{p}) = (e^{Kp_1}, -e^{Kp_1})^\top$ s.t. $Dg(\boldsymbol{p}) = \begin{pmatrix} Ke^{Kp_1} & -Ke^{Kp_1} \\ 0 & 0 \end{pmatrix}$ and $\|g(\mathbf{p})\| = \sqrt{2} e^{Kp_1}$. The only eigenvalue of $Dg(\boldsymbol{p})|_\mathcal{T}$ is Ke^{Kp_1}. Thus, $Dg(\boldsymbol{p}) \succeq Ke^{Kp_1}$. Therefore, by Theorem 3, the optimal report $\boldsymbol{p}$ satisfies $\|\boldsymbol{p} - f(\boldsymbol{p})\| \leq \sqrt{2} L_f / K$. Thus, by choosing $K = L_f/(\sqrt{2}\epsilon)$, we obtain the desired bound. If $L_f < 1$, then by Theorem 4 we further have that $\|\boldsymbol{p} - \boldsymbol{p}^*\| \leq \frac{L_f}{(1-L_f)} \frac{\sqrt{2}}{K}$, so that we can achieve the desired bound by setting $K = (1-L_f)/(\sqrt{2}\epsilon L_f)$. □

A.8 PROOF OF THEOREM 6

Throughout this section we use the following simplifying notation. Let S be a proper scoring rule for the two-outcome case with G, g as per Theorem 1. Then for a single probability $p \in [0, 1]$ we define $G(p) = G(p, 1-p)$ and $g(p) = (1, -1) g(p, 1-p)$. And $S(p, q) = S((p, 1-p), (q, 1-q))$. Then we have that $S(p, q) = g(p)(q-p) + G(p)$, where g as

a function on $[0,1]$ is a subgradient of G as a function on $[0,1]$. Conversely, note that any function of this form induces a proper scoring rule on $\Delta(\{1,2\})$.

First, we prove a result that reduces the claim about S to a claim about g.

Lemma 4. *Let $[p_1,p_2]$ be any interval and S be a proper scoring rule defined via g as usual. Then*

$$\frac{\sup_{p\in[p_1,p_2]} S(p,p) - S(p+x,p)}{\inf_{p\in[p_1,p_2]} S(p,p) - S(p+x,p)} \geq \frac{1}{4}\frac{\sup_{p\in[p_1,p_2]} g(p+x) - g(p)}{\inf_{p\in[p_1,p_2]} g(p+x) - g(p)}.$$

Proof. For our proof, we will use the following bounds:

$$\begin{aligned}
S(p,p) - S(p+x,p) &= xg(p+x) - \int_p^{p+x} g(t)dt \\
&\geq xg(p+x) - xg(p+x)/2 - xg(p+x/2)/2 \\
&= x(g(p+x) - g(p+x/2))/2 \\
S(p,p) - S(p+x,p) &= xg(p+x) - \int_p^{p+x} g(t)dt \\
&\leq xg(p+x) - xg(p) \\
&= x(g(p+x) - g(p)).
\end{aligned}$$

Using these bounds, we can prove the lemma as follows:

$$\begin{aligned}
\frac{\sup_p S(p,p) - S(p+x,p)}{\inf_p S(p,p) - S(p+x,p)} &\geq \frac{\sup_p x(g(p+x) - g(p+x/2))/2}{\inf_p x(g(p+x) - g(p))} \\
&= \frac{1}{2}\frac{\sup_p g(p+x) - g(p+x/2)}{\inf_p g(p+x) - g(p)} \\
&\geq \frac{1}{4}\frac{\sup_p g(p+x) - g(p)}{\inf_p g(p+x) - g(p)}.
\end{aligned}$$

□

Lemma 5. *Let $y \geq 0$, $h > 0$. Let $g \geq 0$ be strictly increasing on $[a,b]$ s.t. $g(x+h) - g(x) \geq yg(x)$ for all $x \in [a,b]$. Then*

$$\frac{\sup_{x\in[a,b]} g(x+h) - g(x)}{\inf_{x\in[a,b]} g(x+h) - g(x)} \geq y(1+y)^{\lfloor (b-a)/h \rfloor - 1}.$$

Proof. Let $N = \lfloor \frac{b-a}{h} \rfloor$.

Note that $g(x+h) \geq (1+y)g(x)$ for $x \in [a,b]$. Thus, iterating, we get that $g(a+Nh) \geq (1+y)^{N-1}g(a+h)$.

As a consequence, since $a + Nh \leq b$, we have:

$$\begin{aligned}
g(a+(N+1)h) - g(a+Nh) &\geq yg(a+Nh) \\
&\geq y(1+y)^{N-1}g(a+h) \\
&\geq y(1+y)^{N-1}(g(a+h) - g(a))
\end{aligned}$$

And so:

$$\frac{\sup_{x\in[a,b]} g(x+h) - g(x)}{\inf_{x\in[a,b]} g(x+h) - g(x)} \geq \frac{g(a+Nh+h) - g(a+Nh)}{g(a+h) - g(a)} \geq y(1+y)^{N-1}$$

□

Lemma 6. *Let S defined via g as usual be a proper scoring rule. Let $\epsilon > 0, L > 0$. Assume that S has the following property: For every f with Lipschitz constant L, we have $|p^* - f(p^*)| \leq \epsilon$ for the optimal report(s) p^*. Let $\delta = \frac{\epsilon}{L+1}$. Then for every 2δ interval contained in $[0, 1 - 3\epsilon + 2\delta]$, there is a p in that interval such that ($p + 2\delta \leq 1$ and)*

$$g(p + 2\delta) - g(p) \geq \frac{2L}{L+3} g(p).$$

(Note that this result doesn't assume $g > 0$. However, note that the consequent of the lemma is vacuous if $g(p) \leq 0$ (since g is monotone increasing.)

Proof. We shall show, equivalently, that for every interval of width 2δ in $[2\delta, 1 - 3\epsilon + 4\delta]$, there is some p (in $[0, 1]$) contained in the interval such that:

$$g(p) - g(p - 2\delta) \geq \frac{2L}{L+3} g(p - 2\delta) \tag{12}$$

Given an interval of width 2δ in $[2\delta, 1 - 3\epsilon + 4\delta]$, write the interval as $[p_0 - \delta, p_0 + \delta]$, where $p_0 \in [3\delta, 1 - 3L\delta] = [3\delta, 1 - 3\epsilon + 3\delta] \subseteq [0, 1]$.

Then, we construct f as follows.

Let

$$k_2 := p_0 \left(1 + \frac{1}{L}\right) \tag{13}$$

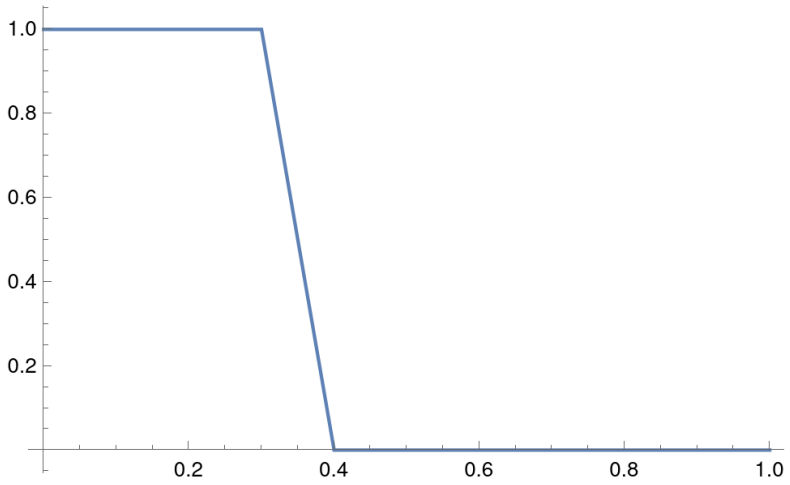
$$k_1 := k_2 - \frac{1}{L} = p_0 - \frac{1}{L}(1 - p_0). \tag{14}$$

Then consider

$$f(p) := \begin{cases} 1 & \text{if } p \leq k_1 \\ 0 & \text{if } p \geq k_2 \\ \frac{k_2 - p}{k_2 - k_1} = L(k_2 - p) & \text{if } k_1 \leq p \leq k_2 \end{cases}$$

for $p \in [0, 1]$.

For $k_1 = 0.3, k_2 = 0.4, L = 10$, this function looks as follows.



Note that f then has Lipschitz constant L.

Moreover, note that f has a unique fixed point, which occurs at $p_0 \in (k_1, k_2)$, since

$$f(p_0) = L(k_2 - p_0) = L\left(p_0 + \frac{1}{L} p_0 - p_0\right) = p_0.$$

We can see that for any proper scoring rule, the optimal report under f is in $[k_1, k_2] \cap [0, 1]$.

Next we will show that for $p \in [k_1, k_2] \cap [0, 1]$ to satisfy the bound $|f(p) - p| \leq \epsilon$, we must have that $p \in [p_0 - \delta, p_0 + \delta]$. To show this, observe that, if $p \in [k_1, k_2] \cap [0, 1]$:

$$\begin{aligned} f(p) - p &= f(p) - f(p_0) - (p - p_0) + f(p_0) - p_0 \\ &= -L(p - p_0) - (p - p_0) + f(p_0) - p_0 \\ &= -(L + 1)(p - p_0) + f(p_0) - p_0 \\ &\underset{p_0 \text{ fixed point}}{=} -(L + 1)(p - p_0), \end{aligned}$$

So that, for $k_1 \leq p \leq k_2$ we have $|f(p) - p| \leq \epsilon$ if and only if $|p - p_0| < \delta$. Thus, by assumption, we must have that the optimal report p^* satisfies $p^* \in [p_0 - \delta, p_0 + \delta]$. We will show the claim of the theorem by showing that that $p = p^*$ satisfies equation (12).

First, we will check that $p^* - 2\delta$ is in $[0, 1]$, and is still on the steep section of the graph, i.e., is in $[k_1, k_2]$.

We have

$$p^* - 2\delta \geq p_0 - 3\delta \underset{\text{Equation (14)}}{=} k_1 + \frac{1}{L}(1 - p_0) - 3\delta \underset{p_0 \leq 1 - 3L\delta}{\geq} k_1.$$

Also we chose p_0 to satisfy $p_0 - 3\delta \geq 0$, so that overall $p^* - 2\delta \geq \max(k_1, 0)$. Also, we must have that $\max(k_1, 0) \leq p^* \leq \min(k_2, 1)$.

Now we just use the optimality of p^* and the definition of f to get the result. $S(p^*, f(p^*)) \geq S(p^* - 2\delta, f(p^* - 2\delta))$, i.e.:

$$\begin{aligned} G(p^*) + g(p^*)(f(p^*) - p^*) &\geq G(p^* - 2\delta) + g(p^* - 2\delta)(f(p^* - 2\delta) - p^* + 2\delta) \\ &= G(p^* - 2\delta) + g(p^* - 2\delta)(f(p^*) + 2L\delta - p^* + 2\delta) \end{aligned}$$

Now, by the fact that g is a subgradient of G, we have that $G(p^*) - G(p^* - 2\delta) \leq 2\delta g(p^*)$. Thus, rearranging:

$$2\delta g(p^*) + g(p^*)(f(p^*) - p^*) \geq g(p^* - 2\delta)(f(p^*) + 2L\delta - p^* + 2\delta)$$

and so

$$(f(p^*) - p^* + 2\delta)(g(p^*) - g(p^* - 2\delta)) \geq 2\delta L g(p^* - 2\delta).$$

Thus, since $|f(p^*) - p^*| \leq \epsilon$

$$\begin{aligned} g(p^*) - g(p^* - 2\delta) &\geq g(p^* - 2\delta)\frac{2L\delta}{\epsilon + 2\delta} \\ &= \frac{2L}{L + 3} g(p^* - 2\delta). \end{aligned}$$

□

Lemma 7. *Let g be any monotonically increasing nonnegative function with the property that on each interval of length h fully contained in $[a, b]$ there is p in that interval such that $g(p + h) - g(p) \geq yg(p)$. Then for all $p \in [a, b - h]$, $g(p + 2h) - g(p) \geq yg(p)$, provided g is defined on $[a, b + h]$.*

Proof. For any $p \in [a, b - h]$, we have that the interval $[p, p + h]$ must contain some p^* with $g(p^* + h) - g(p^*) \geq yg(p^*)$. Then, since $p \leq p^*$, and $p^* + h \leq p + 2h$, we have, by monotonicity

$$g(p + 2h) - g(p) \geq g(p^* + h) - g(p^*) \geq yg(p^*) \geq yg(p).$$

□

Lemma 8. *Let $S(p, q) = g(p)(q - p) + G(p)$ be a (strictly) proper scoring rule, and $f : [0, 1] \to [0, 1]$ be L_f-Lipschitz. Let*

$$\begin{aligned} \tilde{g}(p) &:= -g(1 - p) \\ \tilde{G}(p) &:= G(1 - p) \\ \tilde{S}(p, q) &:= \tilde{g}(p)(q - p) + \tilde{G}(p) \\ \tilde{f}(p) &:= 1 - f(1 - p) \end{aligned}$$

Then $\tilde{S}$ is a (strictly) proper scoring rule, $\tilde{f}$ is L_f-Lipschitz and

$$\tilde{S}(1-p, \tilde{f}(1-p)) = S(p, f(p)).$$

Proof. First, we show that $\tilde{S}$ is a (strictly) proper scoring rule, by verifying that it conforms to the Gneiting and Raftery characterization. From the form of $\tilde{G}$, we can see that $\tilde{G}$ is (strictly) convex iff G is. It remains only to check that $\tilde{g}$ is a subderivative of g. Then, we have

$$\tilde{G}(p) - \tilde{G}(q) = G(1-p) - G(1-q) \geq g(1-q)(1-p-1+q) = \tilde{g}(q)(p-q).$$

as required. By inspection, $\tilde{f}$ is L_f-Lipschitz.

Finally,

$$\begin{aligned}
\tilde{S}(1-p, \tilde{f}(1-p)) &= \tilde{g}(1-p)(\tilde{f}(1-p) - (1-p)) + \tilde{G}(1-p) \\
&= (-g(p))(1 - f(p) - (1-p)) + G(p) \\
&= g(p)(f(p) - p) + G(p) \\
&= S(p, f(p)).
\end{aligned}$$

□

Lemma 9. *Suppose S is a proper scoring rule defined via g s.t. for some $\epsilon, L_f > 0$ we have that whenever f is L_f-Lipschitz, the optimal report p^* satisfies $|f(p^*) - p^*| < \epsilon$. Let $\delta = \frac{\epsilon}{L_f+1}$. Further consider p_l, p_h s.t. $3\epsilon - 3\delta \leq p_l \leq p_h \leq 1 - 3\epsilon - \delta$. Then we have:*

$$\frac{\sup_{x\in[p_l,p_h]} |g(x+4\delta) - g(x)|}{\inf_{x\in[p_l,p_h]} |g(x+4\delta) - g(x)|} \geq \frac{2L_f}{L_f+3} \left(3\frac{L_f+1}{L_f+3}\right)^{(L_f+1)(p_h-p_l)/(8\epsilon)-5/2}.$$

Proof. We will, show, equivalently, that for $3\epsilon - \delta \leq p_l \leq p_h \leq 1 - 3\epsilon + \delta$:

$$\frac{\sup_{x\in[p_l,p_h]} |g(x+2\delta) - g(x-2\delta)|}{\inf_{x\in[p_l,p_h]} |g(x+2\delta) - g(x-2\delta)|} \geq \frac{2L_f}{L_f+3} \left(3\frac{L_f+1}{L_f+3}\right)^{(L_f+1)(p_h-p_l)/(8\epsilon)-5/2}. \tag{15}$$

Consider first the case where $g((p_l+p_h)/2) < 0$. Then consider $\tilde{g}$ as specified by Lemma 8. Note that from Lemma 8 it follows that $\tilde{S}, \tilde{g}$ satisfy the claim of the Theorem in the form of equation (15) for $[p'_l, p'_h] := [1-p_h, 1-p_l]$ if and only if S, g satisfy the claim of the Theorem for $[p_l, p_h]$:

$$\begin{aligned}
\frac{\sup_{x\in[p_l,p_h]} |g(x+2\delta) - g(x-2\delta)|}{\inf_{x\in[p_l,p_h]} |g(x+2\delta) - g(x-2\delta)|} &= \frac{\sup_{x\in[p_l,p_h]} |\tilde{g}(1-x+2\delta) - \tilde{g}(1-x-2\delta)|}{\inf_{x\in[p_l,p_h]} |\tilde{g}(1-x+2\delta) - \tilde{g}(1-x-2\delta)|} \\
&= \frac{\sup_{y\in[1-p_h,1-p_l]} |\tilde{g}(y+2\delta) - \tilde{g}(y-2\delta)|}{\inf_{y\in[1-p_h,1-p_l]} |\tilde{g}(y+2\delta) - \tilde{g}(y-2\delta)|}.
\end{aligned}$$

Note further that $\tilde{g}((p'_l+p'_h)/2) > 0$. Thus, for our proof we can assume WLOG $g((p_l+p_h)/2) > 0$ and thus by monotonicity $g(x) > 0$ for $x > (p_l+p_h)/2$.

Note that $p_h \leq 1 - 3\epsilon + \delta \leq 1 - 3\epsilon + 2\delta$. So, by Lemma 6, we have that in every 2δ interval contained in $[(p_l+p_h)/2, p_h]$ there is a p such that

$$g(p+2\delta) - g(p) \geq \frac{2L_f}{L_f+3} g(p).$$

Hence, by Lemma 7, we have that for all $p \in [(p_l+p_h)/2, p_h - 2\delta]$,

$$g(p+4\delta) - g(p) \geq \frac{2L_f}{L_f+3} g(p)$$

since $p_h + 2\delta \leq 1 - 3\epsilon + \delta + 2\delta \leq 1$. Thus, by Lemma 5,

$$\begin{aligned}
&\frac{\sup_{x\in[p_l,p_h]} |g(x+2\delta) - g(x-2\delta)|}{\inf_{x\in[p_l,p_h]} |g(x+2\delta) - g(x-2\delta)|} \\
&= \frac{\sup_{x\in[p_l-2\delta,p_h-2\delta]} |g(x+4\delta) - g(x)|}{\inf_{x\in[p_l-2\delta,p_h-2\delta]} |g(x+4\delta) - g(x)|} \\
&\geq \frac{\sup_{x\in[(p_l+p_h)/2,p_h-2\delta]} g(x+4\delta) - g(x)}{\inf_{x\in[(p_l+p_h)/2,p_h-2\delta]} g(x+4\delta) - g(x)} \geq \frac{2L_f}{L_f+3}\left(1 + \frac{2L_f}{L_f+3}\right)^{\lfloor (p_h-p_l)/(8\delta) - 1/2 \rfloor - 1}.
\end{aligned}$$

□

Theorem 6. *Suppose S is a proper scoring rule s.t. for some $\epsilon, L_f > 0$ we have that whenever f is L_f-Lipschitz, the optimal report $\boldsymbol{p}$ satisfies $\|f(\boldsymbol{p}) - \boldsymbol{p}\| < \epsilon$. Let $3\epsilon \leq p_l \leq p_h \leq 1 - 4\epsilon$ and $\delta = \epsilon/(L_f + 1)$. Then the ratio of the supremum and infimum over $p_1 \in [p_l, p_h]$ of $S((p_1 + 4\delta, 1 - p_1 - 4\delta), (p_1, 1 - p_1)) - S((p_1, 1 - p_1), (p_1, 1 - p_1))$ is at least*

$$\frac{L_f}{2L_f + 6}\left(3\frac{L_f + 1}{L_f + 3}\right)^{(L_f+1)(p_h-p_l)/(8\epsilon) - 5/2}.$$

In particular, for fixed positive L_f, this term is exponential in $1/\epsilon$ and for fixed positive ϵ it is exponential in L_f.

Proof. Follows from Lemmas 4 and 9. □

A.9 PROOF OF THEOREM 7

We'll first need a lemma that we can find a section of an isoline of sufficient length that doesn't turn too much:

Lemma 10. *Let $\Delta(\mathcal{N})$ be the probability simplex in $\mathbb{R}^3$ (an equilateral triangle with side length $\sqrt{2}$ lying in a plane embedded in $\mathbb{R}^3$). Let $G : \Delta(\mathcal{N}) \to \mathbb{R}$ be strictly convex, with subgradient g (with entries summing to 0). Let $r = \frac{\sqrt{6}}{12}$, and $0 < l < r$. Then we can find a section of an isoline of G, γ, with the following properties:*

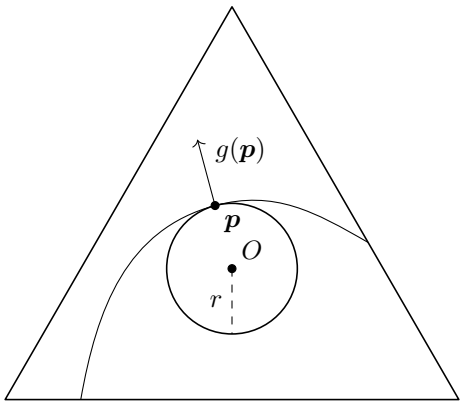
1. *γ has length l.*
2. *Let $\boldsymbol{p}$ be one endpoint of γ, and $\boldsymbol{n}$ be the unit vector parallel to $g(\boldsymbol{p})$, i.e. $\boldsymbol{n} = g(\boldsymbol{p})/\|g(\boldsymbol{p})\|$. Then, $\boldsymbol{p} - 2r\boldsymbol{n} \in \Delta(\mathcal{N})$, and $G(\boldsymbol{p} - 2r\boldsymbol{n}) \leq G(\boldsymbol{p})$.*
3. *For all $\boldsymbol{p}' \in \gamma$, the angle θ between $g(\boldsymbol{p})$ and $g(\boldsymbol{p}')$ satisfies $\theta \leq \frac{2l}{r}$.*
4. *Each point on γ has distance at least $r - l$ from the boundary of the simplex.*

Moreover, let γ be an isoline section with the above properties. Let $\boldsymbol{p}$ and $\boldsymbol{n}$ be defined as above, and $\boldsymbol{t}$ a unit vector perpendicular to $\boldsymbol{n}$ (and to $\mathbf{1}$). Let $\boldsymbol{q}$ be the other endpoint of γ. Then $|\boldsymbol{t}^\top(\boldsymbol{q} - \boldsymbol{p})| \geq l\left(1 - \frac{2l}{r}\right)$. I.e. the length of γ in the direction orthogonal to $\boldsymbol{n}$ is at least $l\left(1 - \frac{2l}{r}\right)$.

Proof. Note that each isoline $\{\boldsymbol{x} \in \Delta(\mathcal{N}) : G(\boldsymbol{x}) = y\}$ forms part of the boundary of the set $\{\boldsymbol{x} \in \Delta(\mathcal{N}) : G(\boldsymbol{x}) \leq y\}$, which by strict convexity of G is a convex set. Call this the *enclosed set* of the isoline. Then, at each point $\boldsymbol{p}$ on an isoline, there is at least one supporting line to the isoline, i.e., a straight line that touches the isoline but does not contain any of the interior points of the enclosed set. Moreover, $g(\boldsymbol{p})$ is always perpendicular to a supporting line to the isoline through $\boldsymbol{p}$, and points out of the enclosed set.

We will proceed by finding an isoline tangent to and enclosing a circle at the center of $\Delta(\mathcal{N})$, and then arguing that a section of this isoline has the desired properties.

First, consider the circle of radius r at the center of $\Delta(\mathcal{N})$. Let $\boldsymbol{p}$ be a point on the boundary of this circle at which G is maximal, and consider the isoline of G through $\boldsymbol{p}$.



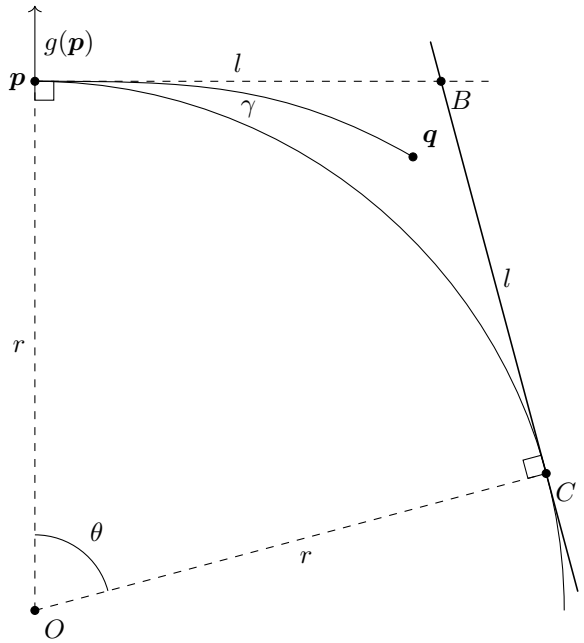
Observe that, by construction, the enclosed set of this isoline contains the circle. Moreover, the isoline is tangent to the circle at $\boldsymbol{p}$. Note further that the tangent line to the circle at $\boldsymbol{p}$ must be the unique supporting line to the isoline through $\boldsymbol{p}$, so that $g(\boldsymbol{p})$ is perpendicular to this tangent. As a consequence, we know that if $\boldsymbol{n} = \frac{g(\boldsymbol{p})}{\|g(\boldsymbol{p})\|}$, then $\boldsymbol{p} - 2r\boldsymbol{n}$ is the point on the opposite side of the circle. Thus, we must have $G(\boldsymbol{p} - 2r\boldsymbol{n}) \leq G(\boldsymbol{p})$, as required.

Now, we have two cases: either the isoline stays within the interior of the simplex, or it reaches the boundary of the simplex. In the former case, the total length of the isoline is at least the circumference of the circle, i.e., $2\pi r > l$. Thus, we may choose a section (with two distinct endpoints) of the isoline, γ, with length l and endpoint $\boldsymbol{p}$.

Now, note that the distance from the centre of the simplex to the (nearest point on the) boundary is $\frac{\sqrt{6}}{6} = 2r$. Hence, the minimum distance from the circle to the boundary of the simplex is at least the $2r - r = r > l$. Thus, in the latter case, the isoline must have a connected section starting at $\boldsymbol{p}$ of length l. Call this section γ.

In either case, γ has distance at least $r - l$ from the boundary of the simplex.

Now, we will bound the change in the angle of supporting lines, moving along γ. WLOG assume $\boldsymbol{p}$ is the anticlockwise-most point of γ.



Let the center of the circle be O. Consider the tangent to the circle at the point C, where C is such that OC is at an angle of θ from the line from O to $\boldsymbol{p}$. Let the intersection of the tangents through C and $\boldsymbol{p}$ be B. Let θ be such that the line from $\boldsymbol{p}$ to B has length l. Note that $\theta < \pi/2$, since $l < r$. Let $\boldsymbol{q}$ be the other (clockwise-most) endpoint of γ.

Note then that since the isoline must lie below the line $\boldsymbol{p}B$, and the length of γ is l, γ never crosses the line BC.

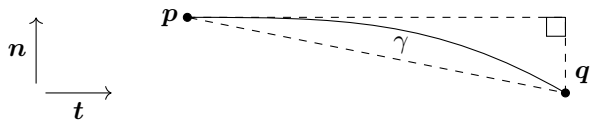
Meanwhile, consider a point $\boldsymbol{p}'$ on γ. Note that since the enclosed set of the isoline contains the circle, no supporting line to the isoline through $\boldsymbol{p}'$ contains interior points of the circle. The angle of such a line must then lie between the angles of the tangents at $\boldsymbol{p}$ and C (ie, is at most as steep as BC, on the diagram), and so makes angle at most θ with $\boldsymbol{p}B$. Since g points out of the enclosed set of the isoline, orthogonal to its supporting lines, the difference in angle between $g(\boldsymbol{p}')$ and $g(\boldsymbol{p})$ is then at most θ.

Therefore, we have that:

$$\frac{\theta}{2} \leq \tan\left(\frac{\theta}{2}\right) = \frac{l}{r}$$

since $\tan x \geq x$ for $x \in [0, \pi/2)$. Thus, $\theta \leq \frac{2l}{r}$, as required. We have now established that γ has the stated properties.

We need then only check the final part of the statement, i.e., that γ with these properties has sufficient length in the direction orthogonal to $\boldsymbol{n}$. Let $\boldsymbol{t}$ be as in the statement of the lemma, i.e., parallel to the supporting line through $\boldsymbol{p}$. We have by convexity that γ lies within the triangle defined by supporting lines to the isoline at $\boldsymbol{p}$ and $\boldsymbol{q}$ and the straight line from $\boldsymbol{p}$ to $\boldsymbol{q}$. Hence, since $\theta < \pi/2$, γ lies entirely within the triangle defined by the supporting line at $\boldsymbol{p}$ parallel to $\boldsymbol{t}$, the line through $\boldsymbol{q}$ parallel to $\boldsymbol{n}$, and the line from $\boldsymbol{p}$ to $\boldsymbol{q}$, as depicted in the diagram below.



Moreover, the maximum possible length of a convex path, within this triangle, from $\boldsymbol{p}$ to $\boldsymbol{q}$ is just the combined length of the two shorter sides, i.e. $|\boldsymbol{n}^\top(\boldsymbol{p}-\boldsymbol{q})| + |\boldsymbol{t}^\top(\boldsymbol{p}-\boldsymbol{q})|$. Thus, $\left|\boldsymbol{t}^\top(\boldsymbol{p}-\boldsymbol{q})\right| \geq l - \left|\boldsymbol{n}^\top(\boldsymbol{p}-\boldsymbol{q})\right|$. Then, note that the straight line from $\boldsymbol{p}$ to $\boldsymbol{q}$ makes angle at most θ with the line parallel to $\boldsymbol{t}$, since its angle must lie between the angle of supporting lines at $\boldsymbol{p}$ and $\boldsymbol{q}$.

Thus,

$$|\boldsymbol{n}^\top(\boldsymbol{p}-\boldsymbol{q})| \leq \sin\theta \, \|\boldsymbol{p}-\boldsymbol{q}\| \leq l \sin\theta \leq l\theta.$$

Hence, we have

$$|\boldsymbol{t}^\top(\boldsymbol{p}-\boldsymbol{q})| \geq l - l\theta = l\left(1 - \frac{2l}{r}\right)$$

and we are done. □

Now, our main result: We can't get arbitrarily good bounds for fixed Lipschitz constant, and the bound one can at best get scales linearly with L_f in the limit $L_f \to 0$.

Theorem 7. *For any Lipschitz constant L_f, for $\epsilon > 0$ sufficiently small, there is no proper scoring rule S for the three-outcome case that achieves the following property: Whenever f is L_f-Lipschitz, there is some performatively optimal report $\boldsymbol{p}$ with $\|f(\boldsymbol{p}) - \boldsymbol{p}\| \leq \epsilon$. In particular, there exists some function $\epsilon(L_f)$ with $\epsilon(L_f) \sim cL_f$ as $L_f \to 0$ for some fixed constant c, s.t. the above property cannot be achieved with $\epsilon = \epsilon(L_f)$. Thus, the best achievable bound is in $\Omega(L_f)$ as $L_f \to 0$, i.e. scales at least linearly with L_f in the limit.*

Proof. Let g and G be as in the Gneiting and Raftery characterization of S. Let $\lambda = \min(L_f, 2)$.

We will proceed as follows:

- Find an isoline of G on which the angle of $g(\boldsymbol{p})$ doesn't change much. On this isoline, we are then able to move along the isoline without $g(\boldsymbol{p})$ changing much in the direction of movement, and hence without $g(\boldsymbol{p})^\top \boldsymbol{p}$ changing much.
- Construct a λ-Lipschitz (and hence L_f-Lipschitz) function f with fixed point $\boldsymbol{p}_0$ such that as we move sideways along the isoline, $f(\boldsymbol{p})$ moves upwards, incentivising us to misrepresent in the direction of the isoline.
- We will then show that for a point $\boldsymbol{q}$, with $\|f(\boldsymbol{q}) - \boldsymbol{q}\| \geq \epsilon$, reporting $\boldsymbol{q}$ gives higher score than any point $\boldsymbol{p}$ for which $\|f(\boldsymbol{p}) - \boldsymbol{p}\| < \epsilon$ (for ϵ which we will choose).

Let $\theta = \arctan(\lambda/4) \leq \lambda/4 \leq 1/2$. Note that then $\theta \sim L_f/4$ as $L_f \to 0$.

Let $r = \frac{\sqrt{6}}{12}$. Then, let γ be an isoline satisfying the properties of Lemma 10 with $l = \frac{\theta r}{2}$. Let the end points of γ be $\boldsymbol{p}_1$ and $\boldsymbol{q}$ (chosen such that $\boldsymbol{q}$ is the same end as $\boldsymbol{q}$ in the statement of the Lemma), $\boldsymbol{n} := \frac{g(\boldsymbol{p}_1)}{\|g(\boldsymbol{p}_1)\|}$, and $\boldsymbol{t}$ a unit vector orthogonal to both $\boldsymbol{n}$ and $\mathbf{1}$. Then we have, in particular:

(P1) $\boldsymbol{p}_1 - 2r\boldsymbol{n} \in \Delta(\mathcal{N})$ and $G(\boldsymbol{p}_1 - 2r\boldsymbol{n}) \leq G(\boldsymbol{p}_1)$.

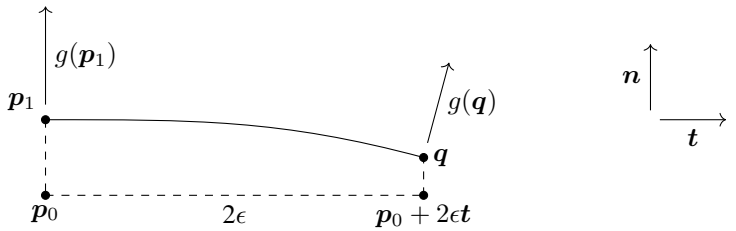
(P2) For all $\boldsymbol{p} \in \gamma$, the angle between $g(\boldsymbol{p})$ and $g(\boldsymbol{p}_1)$ (equivalently, $\boldsymbol{n}$) is at most $\frac{2l}{r} = \theta$.

(P3) Each point on γ has distance at least $r - l \geq 3l$ from the boundary of the simplex.

(P4) $|t^\top(\boldsymbol{q} - \boldsymbol{p}_1)| \geq l(1 - \frac{2l}{r}) = l(1 - \theta)$.

Let $\epsilon = \frac{1}{2}|t^\top(\boldsymbol{q} - \boldsymbol{p}_1)|$. We have, by (P4), $\epsilon \geq l(1-\theta)/2 \geq l/4 > 0$. Note that $l(1-\theta)/2 \leq \epsilon \leq l/2$, and $l\theta = o(L_f)$, and so $\epsilon \sim \frac{1}{2}l = \frac{1}{4}\theta r \sim \frac{1}{16}L_f r = \frac{\sqrt{6}}{192}L_f$ as $L_f \to 0$.

Let $\boldsymbol{p}_0 = \boldsymbol{p}_1 - \epsilon\lambda\boldsymbol{n}$. Note that since $\epsilon\lambda \leq 2\epsilon \leq l$, we have by (P3) that $\boldsymbol{p}_0$ is within the simplex.



By construction, there is a supporting line to γ, parallel to $\boldsymbol{t}$, through $\boldsymbol{p}_1$. Thus, $(\boldsymbol{q} - \boldsymbol{p}_1)^\top \boldsymbol{n} \leq 0$.

Now, let $f(\boldsymbol{p}) = \boldsymbol{p}_0 + \lambda \min(|\boldsymbol{t}^\top(\boldsymbol{p} - \boldsymbol{p}_0)|, 2\epsilon)\boldsymbol{n}$. Note that the image of f is the line segment $[\boldsymbol{p}_0, \boldsymbol{p}_0 + 2\epsilon\lambda\boldsymbol{n}]$, which has maximum distance $\lambda\epsilon \leq l$ from γ, and hence by (P3) is entirely within the probability simplex. Also, f has Lipschitz constant $\lambda \leq L_f$.

Then, if $\|f(\boldsymbol{p}) - \boldsymbol{p}\| \leq \epsilon$, we must have

$$\epsilon \geq |\boldsymbol{t}^\top(f(\boldsymbol{p}) - \boldsymbol{p})| = |\boldsymbol{t}^\top(\boldsymbol{p}_0 - \boldsymbol{p})|$$

and hence $f(\boldsymbol{p})$ must in fact lie in the line segment $[\boldsymbol{p}_0, \boldsymbol{p}_0 + \epsilon\lambda\boldsymbol{n}] = [\boldsymbol{p}_0, \boldsymbol{p}_1]$. Moreover, $\|f(\boldsymbol{q}) - q\| \geq 2\epsilon > \epsilon$.

Meanwhile, we have by convexity and (P1) that for $\boldsymbol{p} \in [\boldsymbol{p}_1 - 2r\boldsymbol{n}, \boldsymbol{p}_1]$, $G(\boldsymbol{p}) \leq \max(G(\boldsymbol{p}_1 - 2r\boldsymbol{n}), G(\boldsymbol{p}_1)) = G(\boldsymbol{p}_1)$. Hence, since $\lambda\epsilon < 2r$, the maximum of G on $[\boldsymbol{p}_0, \boldsymbol{p}_1]$ is $G(\boldsymbol{p}_1)$.

Therefore, whenever $\|f(\boldsymbol{p}) - \boldsymbol{p}\| \leq \epsilon$:

$$S(\boldsymbol{p}, f(\boldsymbol{p})) \leq S(f(\boldsymbol{p}), f(\boldsymbol{p})) = G(f(\boldsymbol{p})) \leq G(\boldsymbol{p}_1)$$

that is, the maximum achievable score is at most the score of honestly reporting $\boldsymbol{p}_1$.

We will now show that the score of reporting $\boldsymbol{q}$ is greater than $G(\boldsymbol{p}_1)$.

First, we have that

$$\begin{aligned} S(\boldsymbol{q}, f(\boldsymbol{q})) &= g(\boldsymbol{q})^\top(f(\boldsymbol{q}) - \boldsymbol{q}) + G(\boldsymbol{q}) && \text{(Gn\&Raf)} \\ &= g(\boldsymbol{q})^\top(\boldsymbol{p}_0 + 2\epsilon\lambda\boldsymbol{n} - \boldsymbol{q}) + G(\boldsymbol{p}_1) && \text{(Def. of } f, \boldsymbol{q}) \end{aligned}$$

It is left to show that the left summand is positive. We have that

$$
\begin{aligned}
& g(\boldsymbol{q})^\top (\boldsymbol{p}_0 + 2\epsilon\lambda \boldsymbol{n} - \boldsymbol{q}) \\
& = (2\epsilon\lambda + (\boldsymbol{p}_0 - \boldsymbol{q})^\top \boldsymbol{n}) g(\boldsymbol{q})^\top \boldsymbol{n} + ((\boldsymbol{p}_0 - \boldsymbol{q})^\top \boldsymbol{t}) g(\boldsymbol{q})^\top \boldsymbol{t} \\
& \geq \epsilon\lambda g(\boldsymbol{q})^\top \boldsymbol{n} - 2\epsilon |g(\boldsymbol{q})^\top \boldsymbol{t}| && \text{(Def. of } \boldsymbol{q}, \boldsymbol{p}_0) \\
& \geq \|g(\boldsymbol{q})\| \, \epsilon(\lambda \cos\theta - 2\sin\theta) && \text{(By (P2))} \\
& = 2\|g(\boldsymbol{q})\| \, \epsilon \cos(\theta)(\lambda/2 - \tan\theta) \\
& \geq \|g(\boldsymbol{q})\| \, \epsilon \cos(\theta)\lambda/2 > 0 && \text{(Choice of } \theta)
\end{aligned}
$$

□

B PREFERENCES BETWEEN DIFFERENT FIXED POINTS

Proposition 4. *Let $F = \{\boldsymbol{p} \colon f(\boldsymbol{p}) = \boldsymbol{p}\}$ be a set of fixed points of f. Let $\boldsymbol{p} \in F$ such that $\boldsymbol{p}$ is the convex combination of elements of $F - \{\boldsymbol{p}\}$. (In other words, $\boldsymbol{p}$ is in the interior of the convex hull of F). Then if S is strictly proper, there exists a $\boldsymbol{p}^* \in F$ s.t. $S(p^*, f(p^*)) > S(p, f(p))$. Thus, $\arg\max_{p \in F} S(p, f(p))$ is a subset of the extreme points of F.*

This follows directly from the convexity of the expected score under honest reporting as per Theorem 1, but for completeness we provide a detailed proof.

Proof. Let $\boldsymbol{p} = \sum_{i=1}^k c_i \boldsymbol{p}_i$ for $c_i \in [0, 1]$ with $\sum_{i=1}^k c_i = 1$ and $\boldsymbol{p}_i \in F - \boldsymbol{p}$. Then

$$
\begin{aligned}
S(\boldsymbol{p}, f(\boldsymbol{p})) &= g(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p}) + G(\boldsymbol{p}) \\
&\underset{\boldsymbol{p} \text{ fixed point}}{=} G(\boldsymbol{p}) \\
&= G\left(\sum_{i=1}^k c_i \boldsymbol{p}_i\right) \\
&\underset{\text{G strictly convex}}{<} \sum_{i=1}^k c_i G(\boldsymbol{p}_i) \\
&\underset{\boldsymbol{p}_i \text{ fixed point}}{=} \sum_{i=1}^k c_i (g(\boldsymbol{p}_i)(f(\boldsymbol{p}_i) - \boldsymbol{p}_i) + G(\boldsymbol{p}_i)) \\
&= \sum_{i=1}^k c_i S(\boldsymbol{p}_i, f(\boldsymbol{p}_i)).
\end{aligned}
$$

Now for the average of the $S(\boldsymbol{p}_i, f(\boldsymbol{p}_i))$ to be greater than $S(\boldsymbol{p}, f(\boldsymbol{p}))$, at least one of the $S(\boldsymbol{p}_i, f(\boldsymbol{p}_i))$ must be greater than $S(\boldsymbol{p}, f(\boldsymbol{p}))$. □

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 TWO OUTCOMES

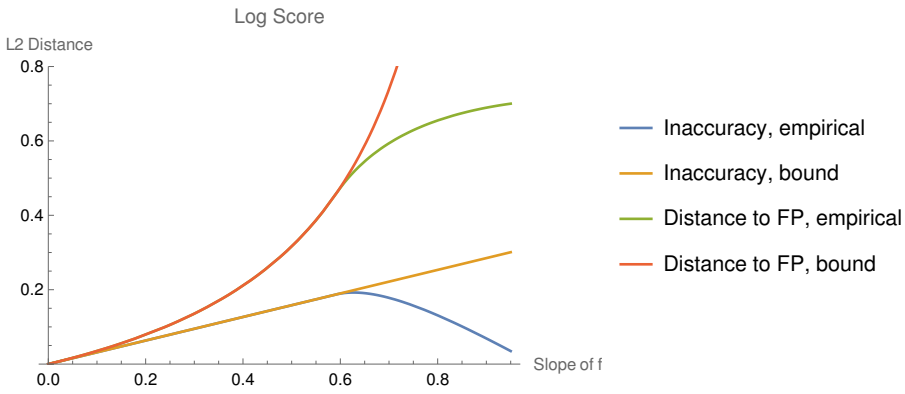


Figure 3: Maximal inaccuracy and maximal distance to fixed point (FP) of optimal predictions, depending on the slope of f, according to our simulation and our theoretical bound.

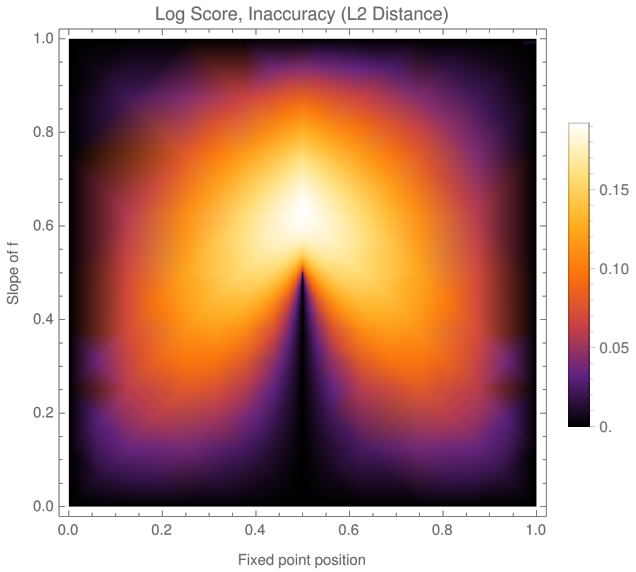


Figure 4: Heatmap of the L2 inaccuracy of optimal predictions, depending on fixed point position and slope of f, for the logarithmic scoring rule.

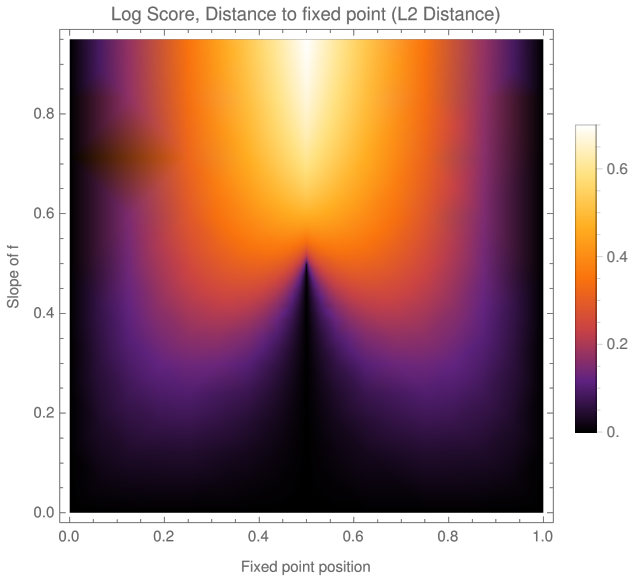


Figure 5: Heatmap of the L2 distance to the fixed point of optimal predictions, depending on fixed point position and slope of f, for the logarithmic scoring rule.

Figures 3 to 5 give the same graphs that we give for the Brier scoring rule in the main text.

Here the bounds for the log scoring rule are obtained as follows. First, note that for the log scoring rule we have that $g(p) = (\log p_i - 1/2(\log p_1 + \log p_2))_i$. So, $\|g(\boldsymbol{p})\| = \frac{|\log(p_1) - \log(p_2)|}{\sqrt{2}}$ and $Dg(p) = \begin{pmatrix} \frac{1}{2p_1} & -\frac{1}{2p_2} \\ -\frac{1}{2p_1} & \frac{1}{2p_2} \end{pmatrix}$. The eigenvalue of this on the tangent space is $1/(2p_1p_2)$. Thus, since Dg is symmetric, $DG(\boldsymbol{p}) \succeq 1/(2p_1p_2)$. By Theorem 3, $\|\boldsymbol{p} - f(\boldsymbol{p})\| \leq L_f \|g(p)\| 2p_1p_2 = \sqrt{2} L_f p_1 p_2 |\log(p_1) - \log(p_2)|$. Numerically this bound seems to be maximized at $p = 0.824$ so that we get a bound $\|\boldsymbol{p} - f(\boldsymbol{p})\| \leq 0.316 L_f$. Similarly, by Theorem 4, $\|\boldsymbol{p} - \boldsymbol{p}^*\| \leq 0.316 L_f/(1 - L_f)$.

For the logarithmic scoring rule, we also give the same plot for the absolute distance between the logits or log odds of the two probabilities (logit distance), see Figure 6. It is defined as $d(\boldsymbol{p}, \boldsymbol{p}') := |\sigma^{-1}(\boldsymbol{p}) - \sigma^{-1}(\boldsymbol{p}')|$, where $\sigma^{-1}(\boldsymbol{p}) := \log \frac{p_1}{p_2}$ is the logit of $\boldsymbol{p}$ (or the inverse sigmoid transform). If probabilities are close to 0 or 1, then L2 distance will always evaluate to very small distances. In contrast, the logit distance depends on order of magnitude differences between probabilities, which may be the more useful quantity.

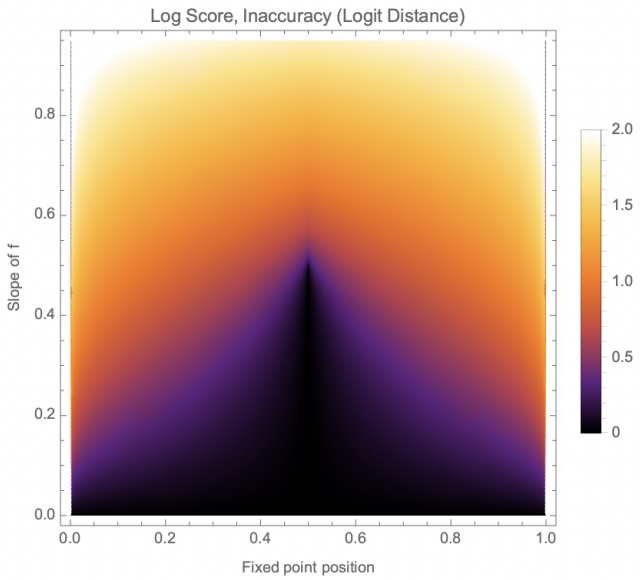


Figure 6: Heatmap of logit distance inaccuracy of optimal predictions for the log scoring rule.

We can see that inaccuracy remains high in logit space for fixed points close to 0 and 1. We don't plot logit distances for the quadratic score, since for that score, optimal predictions often take values close to or equal to $\{(0,1),(1,0)\}$ (even if neither $f(\boldsymbol{p})$ nor $\boldsymbol{p}^*$ lie in $\{(0,1),(1,0)\}$), so the corresponding distances become very large or infinite. The fact that logit distances are bounded for the log score is an advantage of that scoring rule.

C.2 MANY OUTCOMES

C.2.1 Inaccuracy and ditance to fixed point are strongly correlated

Throughout this paper we consider two measures of how wrong a prediction a prediction is, the inaccuracy, i.e., distance of the performatively optimal report $\boldsymbol{p}$ to $f(\boldsymbol{p})$, and the distance of the performatively optimal report to the fixed point. Our experiments show that these measures are closely but not perfectly correlated, see Figure 7. The correlation is 0.958.

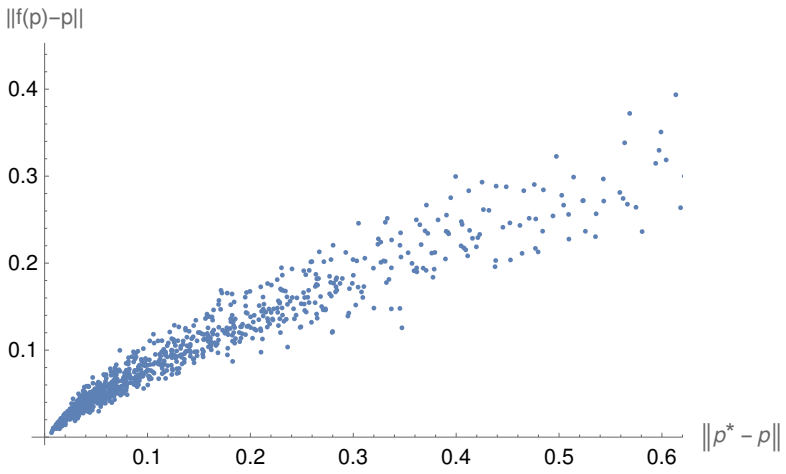


Figure 7: Scatter plot showing the L2 inaccuracy of the performatively optimal report against the L2 distance of the performatively optimal report to the fixed point report.

C.2.2 The effect of fixed point location

Figure 8 scatter-plots the distance to fixed points against the distance of the fixed points from the uniform distribution. The blue line is the best linear fit, which is $0.0274 + 0.751x$. Similarly Figure 9 scatter-plots the inaccuracy of the performatively optimal report against the distance of the fixed point report to the uniform distribution. The blue line is again given by the best linear fit, which is $0.0231 + 0.468x$.

The overall effect of the distance of $\boldsymbol{p}^*$ from uniform actually seems larger than the effect of the operator norm, as indicated by the correlation coefficients in Table 1.

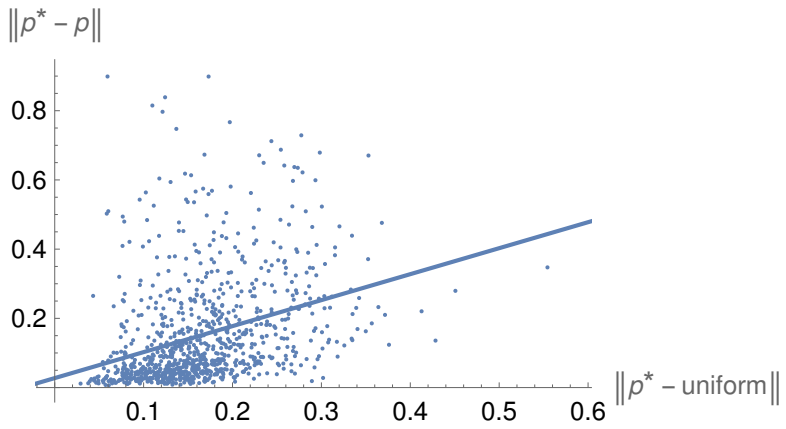


Figure 8: Scatter plot showing the L2 distance of the performatively optimal report to the fixed point report against distance of the fixed point to the uniform distribution in our experiments. The blue line is found by linear regression on the points.

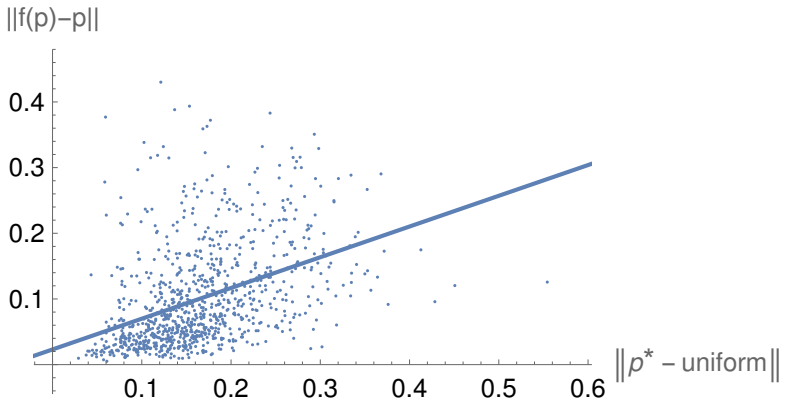


Figure 9: Scatter plot showing the L2 inaccuracy of the performatively optimal report against the distance of the fixed point of f to the uniform distribution in our experiments. The blue line is found by linear regression on the points.

	$\|\boldsymbol{p} - \boldsymbol{p}^*\|$	$\|\boldsymbol{p} - f(\boldsymbol{p})\|$
$\|f_A\|_{\text{op}}$	0.294	0.311
$\|p^* - \frac{1}{n}\mathbf{1}\|$	0.331	0.411

Table 1: Each entry shows the empirical correlation between the quantities determined by the row and column.

D FIXED POINTS VIA ALTERNATIVE NOTIONS OF OPTIMALITY

In this section, we will review alternatives to performance optimality under which fixed points are incentivized. We will elaborate on the settings introduced in Section 8 and provide formal statements and proofs.

To motivate the following, consider an expert AI that chooses its prediction to match its world model, but without explicitly considering the effect of its prediction. For instance, such cognition could arise in an AI trained via a purely supervised objective on historical data. This AI may not learn to take into account effects of its predictions on the outcome of the prediction. If it nevertheless has a world model that generalizes correctly to performative predictions, this could put the AI in a game in which it is trying to make a prediction to match its world model, while the world model updates its beliefs conditional on the AI's prediction. The only equilibria of this game would be fixed points.

Alternatively, fixed points could also result from different training schemes that explicitly optimize an AI's prediction to track empirical outcomes, without also incentivizing influencing the outcomes themselves, such as repeated risk minimization or repeated gradient descent [Perdomo et al., 2020].

Such expert AIs would likely be safer than ones optimizing for performative optimality. First, they report their true beliefs, which gives us better information to base decisions on. This also enables approaches in which we ensure that there is only one safe fixed point. Second, they do not explicitly optimize the choice of fixed point for a goal such as decreasing entropy. Instead, which fixed point is chosen will be contingent on initialization and specifics of the fixed point finding procedure.

D.1 PERFORMATIVE STABILITY AND GAME THEORY

We begin by defining performative stability and relating it to an equilibrium in a two-player game. This represents the core idea behind all of the following settings. A prediction $\boldsymbol{p}^*$ is called *performatively stable* [Perdomo et al., 2020] if

$$\boldsymbol{p}^* \in \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, f(\boldsymbol{p}^*)). \tag{16}$$

First, it is clear that in our case, this is equivalent to $\boldsymbol{p}^*$ being a fixed point.

Proposition 5. *Assume S is strictly proper. Then a prediction $\mathbf{p}^*$ is a fixed point if and only if it is performatively stable.*

Proof. "⇒". Assume $f(\boldsymbol{p}^*) = \boldsymbol{p}^*$. Then $S(\boldsymbol{p}^*, f(\boldsymbol{p}*)) = S(\boldsymbol{p}^*, \boldsymbol{p}^*) \geq S(\boldsymbol{p}, \boldsymbol{p}^*) = S(\boldsymbol{p}, f(\boldsymbol{p}^*))$ for any $\boldsymbol{p}$ since S is proper. Hence, $\boldsymbol{p}^* \in \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, f(\boldsymbol{p}^*))$.

"⇐". Assume $\boldsymbol{p}^* \in \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, f(\boldsymbol{p}^*))$. Then since S is strictly proper, it must be $\boldsymbol{p}^* = f(\boldsymbol{p}^*)$. □

Next, the above objective is equivalent to the definition of a Nash equilibrium in the following game.

Definition 1 (Oracle game)**.** Consider a two-player continuous game in which the first player controls $\boldsymbol{p} \in \Delta(\mathcal{N})$ and the second player controls $\boldsymbol{q} \in \Delta(\mathcal{N})$, with payoff functions $U_1(\boldsymbol{p}, \boldsymbol{q}) := S(\boldsymbol{p}, \boldsymbol{q})$ and $U_2(\boldsymbol{p}, \boldsymbol{q}) := S(\boldsymbol{q}, f(\boldsymbol{p}))$ for the two players, respectively.

If $\boldsymbol{p}^*, \boldsymbol{q}^*$ is a Nash equilibrium of the oracle game, we have $p^* = \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, \boldsymbol{q})$ and $\boldsymbol{q}^* = \arg\max_{\boldsymbol{q}} S(\boldsymbol{q}, f(\boldsymbol{p}^*))$. Substituting the optimal value $\boldsymbol{q}^* = f(\boldsymbol{p}^*)$ for the second player gives us exactly above definition of performative stability in Equation (16). Conversely, if a prediction $\boldsymbol{p}^*$ is performatively stable, then setting $\boldsymbol{q}^* := f(\boldsymbol{p}^*)$ yields a Nash equilibrium.

Proposition 6. *Assume S is a proper scoring rule. Then $\boldsymbol{p} \in \Delta(\mathcal{N})$, $\boldsymbol{q} := f(\boldsymbol{p})$ is a Nash equilibrium of the oracle game, if and only if $\boldsymbol{p}$ is performatively stable. By Proposition 5, this is equivalent to $\boldsymbol{p}$ being a fixed point.*

The oracle game could arise in an agent that uses a causal decision theory [Weirich, 2020] to maximize its score and that believes that S is influenced causally by $\boldsymbol{p}$, but only acausally by $f(\boldsymbol{p})$. In that case, the only *ratifiable* [Jeffrey, 1990, Bell et al., 2021, Ch. 1.7] decision is a Nash equilibrium of the above game. Similarly, the deliberational causal epistemic decision theory discussed by Greaves [2013] would output Nash equilibria of this game (whereas performative optimality would correspond to an agent using evidential epistemic decision theory in this case).

Note that it is important that both players act simultaneously. Perdomo et al. [2020] introduce a Stackelberg version of the oracle game that produces performatively optimal instead of performatively stable reports. Consider a game in which player 1 acts first and chooses $\boldsymbol{p}$, after which player 2 responds with a prediction $\boldsymbol{q}$. Then player 2 responds $\boldsymbol{q} = f(\boldsymbol{p})$ to player 1's action, and player 1's optimization problem becomes

$$p^* = \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, \arg\max_{\boldsymbol{q}} S(\boldsymbol{q}, f(\boldsymbol{p}))) = \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, f(\boldsymbol{p})).$$

D.2 REPEATED RISK MINIMIZATION AND REPEATED GRADIENT DESCENT

Above, we have defined performative stability and a related game which yield fixed points, but we have not defined methods for solving these problems. In the performative prediction context, Perdomo et al. [2020] introduce *repeated risk minimization* and *repeated gradient descent*, both methods that converge to performatively stable points. In this section, we review both schemes and show how repeated gradient descent can be seen as gradient descent on a *stop-gradient* [Foerster et al., 2018, Demski, 2019] objective.

We assume direct access to $\boldsymbol{q}$, instead of having only access to samples distributed according to $\boldsymbol{q}$. In the next section, we discuss online learning when we only have access to samples. One way to understand this distinction is that the former corresponds to the internal cognition of an agent with a belief $\boldsymbol{q} = f(\boldsymbol{p})$ optimizing a prediction $\boldsymbol{p}$. The latter instead corresponds to a machine learning training setup for an oracle AI, where $\boldsymbol{q}$ is the ground truth environment distribution instead of the oracle's belief. Of course, there is no strict divide between the two. Any optimization algorithm could be used either by the agent itself or to train the agent.

First, *repeated risk minimization* is a procedure by which we start with a prediction $\boldsymbol{p}_0$ and then iteratively update the prediction as $\boldsymbol{p}_{t+1} = \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, f(\boldsymbol{p}_t))$. This is also the same as alternating best response learning in the oracle game, where player 1 iteratively updates their prediction, responding to predictions $\boldsymbol{q}_t = f(\boldsymbol{p}_t)$ from player 2. If S is strictly proper, $\boldsymbol{p}_{t+1} = f(\boldsymbol{p}_t)$, and this results in *fixed point iteration* for f. Fixed point iteration converges globally to a fixed point if f has Lipschitz constant $L_f < 1$. It also converges locally to a fixed point $\boldsymbol{p}^*$ if f is continuously differentiable at $\boldsymbol{p}^*$ and $\rho(Df(\boldsymbol{p}^*)) < 1$, where $\rho(Df(\boldsymbol{p}^*))$ is the spectral radius of the Jacobian matrix $Df(\boldsymbol{p}^*)$.

Second, assume that S is differentiable. Then *repeated gradient ascent* updates points via

$$\boldsymbol{p}_{t+1} := \Pi_\Delta(\boldsymbol{p}_t + \alpha \mathbb{E}_{y\sim f(\boldsymbol{p}_t)}[\nabla_{\boldsymbol{p}} S(\boldsymbol{p}_t, y)]),$$

where Π_Δ is the Euclidean projection onto the probability simplex $\Delta(\mathcal{N})$, and $\alpha > 0$ is the learning rate.

Using the definition of $S(\boldsymbol{p}, \boldsymbol{q})$, we have

$$\mathbb{E}_{y\sim f(\boldsymbol{p}_t)}[\nabla_{\boldsymbol{p}} S(\boldsymbol{p}_t, y)] = \nabla_{\boldsymbol{p}}(\mathbb{E}_{y\sim \boldsymbol{q}}[S(\boldsymbol{p}_t, y)])|_{\boldsymbol{q}=f(\boldsymbol{p}_t)} = \nabla_{\boldsymbol{p}}(S(\boldsymbol{p}_t, \boldsymbol{q}))|_{\boldsymbol{q}=f(\boldsymbol{p})}$$

We can express this as

$$\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}_t, \perp f(\boldsymbol{p}_t))) := \nabla_{\boldsymbol{p}}(S(\boldsymbol{p}_t, \boldsymbol{q}))|_{\boldsymbol{q}=f(\boldsymbol{p})},$$

where $\perp$ is the *stop-gradient operator*, which evaluates to the identity function but sets gradients to zero, $\nabla_x(\perp x) = 0$ [Foerster et al., 2018, Demski, 2019].[1] In the following, we call $S(\boldsymbol{p}, \perp f(\boldsymbol{p}))$ the *stop-gradient objective*.

Importantly, it matters that the gradient in repeated gradient ascent lies inside instead of outside the expectation:

$$\mathbb{E}_{y\sim f(\boldsymbol{p}_t)}[\nabla_{\boldsymbol{p}} S(\boldsymbol{p}_t, y)] = \nabla_{\boldsymbol{p}}(S(\boldsymbol{p}_t, \perp f(\boldsymbol{p}_t))) \neq \nabla_p(S(\boldsymbol{p}_t, f(\boldsymbol{p}_t))) = \nabla_{\boldsymbol{p}}\mathbb{E}_{y\sim f(\boldsymbol{p}_t)}[S(\boldsymbol{p}_t, y)]).$$

Unlike repeated gradient ascent, the latter implements gradient ascent on $S(\boldsymbol{p}, f(\boldsymbol{p}))$ and thus leads to performatively optimal reports.

Perdomo et al. [2020] show that, given their assumptions, repeated gradient descent globally converges to stable fixed points. They also provide convergence rates. We will show an analogous result relating repeated gradient ascent to fixed points in our setting, though we won't analyze global convergence or rates of convergence.

To begin, we show that repeated gradient descent is equivalent to Naive Learning [Letcher et al., 2019] in the oracle game, assuming that player 2 always plays $\boldsymbol{q} = f(\boldsymbol{p})$.

Proposition 7. *Assume player* 1 *is performing gradient ascent on its objective with learning rate* α*, under the assumption that player* 2 *always plays* $\boldsymbol{q} = f(\boldsymbol{p})$*. Then player* 1*'s update is*

$$\boldsymbol{p}_{t+1} = \Pi_\Delta(\boldsymbol{p}_t + \alpha \nabla_{\boldsymbol{p}}(\mathbf{S}(\boldsymbol{p}_t, \perp f(\boldsymbol{p}_t)))).$$

[1] This is not a mathematical function (there is no function that is equal to the identity but has gradient zero everywhere), but rather a notational convention in reference to the `stop_gradient` or `detach` functions from the tensorflow or pytorch python libraries. Interestingly, one can perform valid derivations using the stop-gradient operator (e.g., using the chain rule). We leave it to future work to explore the mathematics behind stop-gradients further.

Proof. The proof follows immediately from the definitions. Player 1's update is, by assumption,

$$\boldsymbol{p}_{t+1} = \Pi_\Delta(\boldsymbol{p}_t + \alpha\nabla_{\boldsymbol{p}}(U_1(\boldsymbol{p}_t, \boldsymbol{q}))) = \Pi_\Delta(\boldsymbol{p}_t + \alpha\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}_t, \boldsymbol{q})))$$

where $\boldsymbol{q}$ is player 2's action. Assuming player 2 plays $\boldsymbol{q} = f(\boldsymbol{p}_t)$, we get

$$\boldsymbol{p}_{t+1} = \Pi_\Delta(\boldsymbol{p}_t + \alpha\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}_t, \boldsymbol{q}))) = \Pi_\Delta(\boldsymbol{p}_t + \alpha\nabla_p(S(\boldsymbol{p}_t, \perp f(\boldsymbol{p}_t))))$$

□

Next, we show that fixed points are critical points of the stop-gradient objective.

Proposition 8. *Assume S is proper and let G, g as in the Gneiting and Raftery characterization of S (Theorem 1) be differentiable. Then for any $\boldsymbol{p} \in \Delta(\mathcal{N})$, we have*

$$\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, \perp f(\boldsymbol{p}))) = Dg(\boldsymbol{p})^\top(f(\boldsymbol{p}) - \boldsymbol{p}).$$

In particular, if $\boldsymbol{p}$ is a fixed point, it follows that $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, \perp f(\boldsymbol{p}))) = 0$. The reverse is true if $Dg(\boldsymbol{p})|_{\mathcal{T}} \succ 0$.

Proof.

$$\begin{aligned}\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, \perp f(\boldsymbol{p}))) = \nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, \boldsymbol{q}))|_{\boldsymbol{q}=f(\boldsymbol{p})} &= \nabla_{\boldsymbol{p}}(G(\boldsymbol{p}) + g(\boldsymbol{p})^\top(\boldsymbol{q} - \boldsymbol{p}))|_{\boldsymbol{q}=f(\boldsymbol{p})} \\ &= (g(\boldsymbol{p}) + Dg(\boldsymbol{p})^\top(\boldsymbol{q} - \boldsymbol{p}) - g(\boldsymbol{p}))|_{\boldsymbol{q}=f(\boldsymbol{p})} = Dg(\boldsymbol{p})^\top(f(\boldsymbol{p}) - \boldsymbol{p}). \quad (17)\end{aligned}$$

If $\boldsymbol{p}$ is a fixed point, it follows that $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, \perp f(\boldsymbol{p}))) = 0$. Moreover, if $Dg(\boldsymbol{p})|_{\mathcal{T}} \succ 0$, then if $f(\boldsymbol{p}) - \boldsymbol{p} \neq 0$,

$$\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, \perp f(\boldsymbol{p})))^\top(f(\boldsymbol{p}) - \boldsymbol{p}) = (f(\boldsymbol{p}) - \boldsymbol{p})^\top Dg(\boldsymbol{p})(f(\boldsymbol{p}) - \boldsymbol{p}) > 0$$

and thus $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, \perp f(\boldsymbol{p}))) \neq 0$. □

Finally, we show that in our setting, repeated gradient ascent locally converges to fixed points $\boldsymbol{p}^*$, assuming that $\|Df(\boldsymbol{p}^*)\|_{\text{op}}$ is sufficiently small. This is a local version of convergence results from Perdomo et al. [2020], adapted to our setting.

Proposition 9. *Let S be a strictly proper scoring rule. Let $\boldsymbol{p}^* \in \text{int}(\Delta(\mathcal{N}))$ be a fixed point of f such that G is three times differentiable at $\boldsymbol{p}^*$, i.e. $D^2g(\boldsymbol{p}^*) = D^2\nabla g(\boldsymbol{p}^*)$ exists. Assume $\beta \succeq Dg(\boldsymbol{p}^*)|_{\mathcal{T}} \succeq \gamma > 0$, that f is differentiable at $\boldsymbol{p}^*$, and $\|Df(\boldsymbol{p}^*)\|_{\text{op}} < \frac{\gamma}{\beta}$. Then, for small enough $\alpha > 0$, an agent taking updates $\boldsymbol{p}_{t+1} = \Pi_\Delta(\boldsymbol{p}_t + \alpha\nabla_{\boldsymbol{p}}(\mathbf{S}(\boldsymbol{p}_t, \perp f(\boldsymbol{p}_t))))$ will locally converge to $\boldsymbol{p}^*$.*

For the proof, we use the following generalization of Ostrowski's theorem, adapted from Kitchen [1966].

Theorem 8 (Kitchen [1966])**.** *Let $\varphi\colon D \subseteq V \to W$ where V, W are Banach spaces. Assume*

- *φ has a fixed point $\boldsymbol{x}^* \in \text{int}(D)$*
- *φ is differentiable at $\boldsymbol{x}^*$*
- *$\rho(D\varphi(\boldsymbol{x}^*)) < 1$.*

Then there exists an open set $U \subseteq D$ with $x^* \in U$ such that, letting $\boldsymbol{x}_0 \in U$ and $\boldsymbol{x}_t := \varphi(\boldsymbol{x}_{t-1})$ for $k \in \mathbb{N}$, we have $\boldsymbol{x}_t \in U$ for all k and $\lim_t \boldsymbol{x}_t = \boldsymbol{x}^*$.

Proof of Proposition 9. The Banach space we consider will be $\mathcal{T}$. Note that, since $\boldsymbol{p}^* \in \text{int}(\Delta(\mathcal{N}))$, there exists an open set $\mathcal{D} \subseteq \mathcal{T}$ (with respect to the standard topology on $\mathcal{T}$) with $0 \in \mathcal{D}$ such that $\boldsymbol{v} + \boldsymbol{p}^* \subseteq \Delta(\mathcal{N})$ for all $\boldsymbol{v} \in \mathcal{D}$. Our iteration function then is

$$\varphi\colon \mathcal{D} \subseteq \mathcal{T} \to \mathcal{T}, \boldsymbol{v} \mapsto \boldsymbol{v} + \alpha\nabla_{\boldsymbol{v}}(S(\boldsymbol{v} + \boldsymbol{p}^*, \perp f(\boldsymbol{v} + \boldsymbol{p}^*))).$$

Note that φ has a fixed point at 0. Our goal is now to show that there exists $\alpha > 0$ and an open set $U \subseteq \mathcal{D}$ such that iterates of φ starting in U stay in U and converge to 0.

To that end, note that, using Proposition 8, we have $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, \perp f(\boldsymbol{p}))) = Dg(\boldsymbol{p})^\top (f(\boldsymbol{p}) - \boldsymbol{p})$ and thus

$$
\begin{aligned}
D\varphi(0) &= D(\boldsymbol{v} \mapsto \boldsymbol{v} + \alpha \nabla_{\boldsymbol{v}} \mathbf{S}(\boldsymbol{v} + \boldsymbol{p}^*, \perp f(\boldsymbol{v} + \boldsymbol{p}^*)))(0) && (18)\\
&= I + \alpha D(\boldsymbol{v} \mapsto Dg(\boldsymbol{v} + \boldsymbol{p}^*)^\top (f(\boldsymbol{v} + \boldsymbol{p}^*) - \boldsymbol{v} - \boldsymbol{p}^*))(0) && (19)\\
&= I + \alpha D^2 g(\boldsymbol{p}^*)[f(\boldsymbol{p}^*) - \boldsymbol{p}^*] + \alpha Dg(\boldsymbol{p}^*)^\top (Df(\boldsymbol{p}^*) - I). && (20)
\end{aligned}
$$

Here, $D^2 g(\boldsymbol{v} + \boldsymbol{p}^*)$ is a third-degree tensor, and $D^2 g(\boldsymbol{v} + \boldsymbol{p}^*)[f(\boldsymbol{p}^*) - \boldsymbol{p}^*]$ is a linear map. Since $f(\boldsymbol{p}^*) = \boldsymbol{p}^*$, it follows $D\varphi(0) = I + \alpha Dg(\boldsymbol{p}^*)^\top (Df(\boldsymbol{p}^*) - I)$. In particular, φ is differentiable at 0.

Now let $\boldsymbol{v}$ be an arbitrary eigenvector of $D\varphi(0)$, with eigenvalue λ and w.l.o.g. assume $\|\boldsymbol{v}\| = 1$. Note that $\boldsymbol{v}^\top Dg(\boldsymbol{p}^*)\boldsymbol{v} \geq \gamma \|\boldsymbol{v}\| = \gamma$ and $\boldsymbol{v}^\top Dg(\boldsymbol{p}^*)\boldsymbol{v} \leq \beta \|\boldsymbol{v}\| \leq \beta$ by assumption. Letting $\alpha := \frac{1}{\beta}$, it follows that $\alpha \boldsymbol{v}^\top Dg(\boldsymbol{p}^*)\boldsymbol{v} \leq 1$ and thus

$$
|1 - \alpha \boldsymbol{v}^\top Dg(\boldsymbol{p}^*)^\top \boldsymbol{v}| = |1 - \alpha \boldsymbol{v}^\top Dg(\boldsymbol{p}^*)\boldsymbol{v}| = 1 - \alpha \boldsymbol{v}^\top Dg(\boldsymbol{p}^*)\boldsymbol{v} \leq 1 - \alpha\gamma.
$$

Moreover, since $Dg(\boldsymbol{p}^*)$ is the Hessian of G and thus symmetric since G is twice differentiable, we have $\|Dg(\boldsymbol{p}^*)\|_{\mathrm{op}} \leq \beta$. Using this, as well as our assumption $\|Df(\boldsymbol{p}^*)\|_{\mathrm{op}} < \frac{\gamma}{\beta}$, we get

$$
\begin{aligned}
|\lambda| = |\lambda \boldsymbol{v}^\top \boldsymbol{v}| = |\boldsymbol{v}^\top D\varphi(0)\boldsymbol{v}| &= |\boldsymbol{v}^\top (I + \alpha Dg(\boldsymbol{p}^*)^\top (Df(\boldsymbol{p}^*) - I))\boldsymbol{v}|\\
&= |\boldsymbol{v}^\top \boldsymbol{v} - \alpha \boldsymbol{v}^\top Dg(\boldsymbol{p}^*)^\top \boldsymbol{v} + \alpha \boldsymbol{v}^\top Dg(\boldsymbol{p}^*)^\top Df(\boldsymbol{p}^*))\boldsymbol{v}|\\
&\leq |1 - \alpha \boldsymbol{v}^\top Dg(\boldsymbol{p}^*)^\top \boldsymbol{v}| + \alpha |\boldsymbol{v}^\top Dg(\boldsymbol{p}^*)^\top Df(\boldsymbol{p}^*)\boldsymbol{v}|\\
&\underset{\text{Cauchy-Schwarz}}{\leq} 1 - \alpha\gamma + \alpha \|Dg(\boldsymbol{p}^*)\boldsymbol{v}\| \|Df(\boldsymbol{p}^*)\boldsymbol{v}\|\\
&\leq 1 - \alpha\gamma + \alpha \|Dg(\boldsymbol{p}^*)\|_{\mathrm{op}} \|\boldsymbol{v}\| \|Df(\boldsymbol{p}^*)\|_{\mathrm{op}} \|\boldsymbol{v}\|\\
&= 1 - \alpha\gamma + \alpha \|Dg(\boldsymbol{p}^*)\|_{\mathrm{op}} \|Df(\boldsymbol{p}^*)\|_{\mathrm{op}} < 1 - \alpha\gamma + \alpha\gamma = 1. \qquad (21)
\end{aligned}
$$

This shows that $\rho(D\varphi(0)) < 1$. Hence, by Theorem 8, we can conclude that there exists an open set $U \subseteq \mathcal{D}$ such that for arbitrary $\boldsymbol{v}_0 \in U$, $\boldsymbol{v}_t := \varphi(\boldsymbol{v}_{t-1}) \in U$ for all $t \geq 1$, and $\lim_{t\to\infty} \boldsymbol{v}_t = 0$. In particular, note that since $\boldsymbol{v}_t \in U$ for all t, $\boldsymbol{v}_t + \boldsymbol{p}^* \in \Delta(\mathcal{N})$ and

$$
\boldsymbol{p}^* + \boldsymbol{v}_{t+1} = \boldsymbol{p}^* + \boldsymbol{v}_t + \alpha \nabla_{\boldsymbol{v}}(S(\boldsymbol{v}_t + \boldsymbol{p}^*, \perp f(\boldsymbol{v}_t + \boldsymbol{p}^*))) = \Pi_\Delta(\boldsymbol{p}^* + \boldsymbol{v}_t + \alpha \nabla_{\boldsymbol{v}}(S(\boldsymbol{v}_t + \boldsymbol{p}^*, \perp f(\boldsymbol{v}_t + \boldsymbol{p}^*))))
$$

for all t. Hence, setting $\boldsymbol{p}_t := \boldsymbol{p}^* + \boldsymbol{v}_t$, it follows $\boldsymbol{p}_{t+1} = \Pi_\Delta(\boldsymbol{p}_t + \alpha \nabla_{\boldsymbol{p}}(\mathbf{S}(\boldsymbol{p}_t, \perp f(\boldsymbol{p}_t))))$ for all t and

$$
\lim_{t\to\infty} \boldsymbol{p}_t = \boldsymbol{p}^* + \lim_{t\to\infty} \boldsymbol{v}_t = \boldsymbol{p}^*.
$$

This concludes the proof. □

D.3 ONLINE LEARNING

Now consider a machine learning setup in which we train an oracle with stochastic gradient ascent on environment samples. We assume that at time t, a model makes a prediction $\boldsymbol{P}_t$ and receives a score $S(\boldsymbol{P}_t, Y_t)$, where $Y_t \sim f(\boldsymbol{P}_t)$. The model is then updated using gradient ascent on $S(\boldsymbol{P}_t, Y_t)$. That is, for some learning rate schedule $(\alpha_t)_t$, we have

$$
\boldsymbol{P}_{t+1} = \Pi_\Delta(\boldsymbol{P}_t + \alpha_t \nabla_{\boldsymbol{p}} S(\boldsymbol{P}_t, Y_t)),
$$

where Π_Δ is the Euclidean projection onto $\Delta(\mathcal{N})$ as before.

We discuss this as a theoretical model for oracles trained using machine learning, to show how training setups may incentivize predicting fixed points. There are many issues with the setting beyond giving accurate predictions; for instance, learning may fail to converge at all, and even if the training process sets the right incentives on training examples, the learned model may be optimizing a different objective when generalizing to new predictions [Hubinger et al., 2019] .

To see that this setting leads to fixed points, note that we have

$$
\mathbb{E}_{Y_t \sim f(\boldsymbol{P}_t)}[\nabla_{\boldsymbol{p}} S(\boldsymbol{P}_t, Y_t)] = \nabla_{\boldsymbol{p}} \mathbb{E}_{Y_t \sim \perp f(\boldsymbol{P}_t)}[S(\boldsymbol{P}_t, Y_t)] = \nabla_{\boldsymbol{p}}(S(\boldsymbol{P}_t, \perp f(\boldsymbol{P}_t))).
$$

That is, the expectation of this gradient, conditional on $\boldsymbol{P}_t$, is exactly the repeated gradient from the previous section. Hence, given the right assumptions, this converges to fixed points instead of performative optima. We do not show this here, but an analogous result in performative prediction was proved by Mendler-Dünner et al. [2020].

There are several variations of this setup that essentially set the same incentives. For instance, one could also draw entire batches of outcomes $Y_{t,1:B}$ and then perform updates based on the batch gradient $\nabla_{\boldsymbol{p}} \sum_{b=1}^{B} S(\boldsymbol{P}_t, Y_{t,b})$. This is a Monte Carlo estimate of the repeated gradient and hence also converges to performatively stable points and thus fixed points [Perdomo et al., 2020]. One could also mix the two algorithms and, e.g., perform gradient ascent on an average of past losses, yielding a version of the backwards-facing oracle discussed in Armstrong [2018].

Note that finding fixed points depends on the fact that we differentiate $S(\boldsymbol{P}_t, Y_t)$ instead of the expectation $\mathbb{E}_{Y_t \sim f(\boldsymbol{P}_t)}[S(\boldsymbol{P}_t, Y_t)] = S(\boldsymbol{P}_t, f(\boldsymbol{P}_t))$. If we used policy gradients to differentiate $S(\boldsymbol{P}_t, f(\boldsymbol{P}_t))$, for instance, we would again optimize for performative optimality. Similarly, we could learn a Q-function representing scores for each prediction, and update the function based on randomly sampled predictions $\boldsymbol{p}$. Then the Q-function would converge to estimates of $S(\boldsymbol{p}, f(\boldsymbol{p}))$, and the highest Q-value prediction would be a performative optimum. There are also some more recent results in performative prediction that explicitly try to estimate the gradient $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p})))$ and thus find performatively optimal instead of stable points [Izzo et al., 2021].

Stop-gradients could also be circumvented in a hidden way [Krueger et al., 2020]. For instance, consider a hyperparameter search to meta-learn a learning algorithm, where the evaluation criterion is the accumulated score during an episode. Then this search would prefer algorithms that optimize $S(\boldsymbol{p}, f(\boldsymbol{p}))$ directly, without a stop-gradient.

Lastly, repeated gradient descent is related to *decoupled approval* in RL Uesato et al. [2020]. The decoupled approval policy gradient samples actions and approval queries independently and can thus differentiate with a stop-gradient in front of the approval signal. In our setting, we can differentiate through $S(\boldsymbol{P}_t, Y_t)$ directly, so it is not necessary to calculate this gradient with a decoupled policy gradient. Decoupled gradients could be used to implement the stop-gradient objective if scores were discrete or otherwise not differentiable.

D.4 NO-REGRET LEARNING

In this section, we consider no-regret learning and show that algorithms have sublinear regret if and only if their prediction error is sublinear. Regret takes environment outcomes as given and asks which predictions would have been optimal in hindsight. It thus corresponds to an alternative notion of optimality with a "stop-gradient" in front of environment probabilities.

As in the previous section, we assume that at time $t \in \mathbb{N}$, the agent (i.e., the oracle AI) makes a prediction $\boldsymbol{P}_t$ and receives a score $S(\boldsymbol{P}_t, Y_t)$, where $Y_t \sim f(\boldsymbol{P}_t)$. The agent's cumulative score at step T is defined as $\sum_{t=1}^{T} S(\boldsymbol{P}_t, Y_t)$. In no-regret learning, we compare performance against *experts*, which choose sequences of probabilities $(\boldsymbol{P}'_t)_t$, $\boldsymbol{P}'_t \in \Delta(\mathcal{N})$. We assume that an expert's prediction $\boldsymbol{P}'_t$ is independent of Y_t conditional on $\boldsymbol{P}_t$. I.e., an expert knows the predictions $\boldsymbol{P}_t$ and thus probabilities $f(\boldsymbol{P}_t)$, but it does not know the outcome of Y_t. Let $\mathcal{P}$ be the set of all such experts.

The regret of the agent is the difference between the cumulative score received by the best expert in expectation and the cumulative score received by the agent. To define it formally, let

$$\boldsymbol{P}^*_t \in \arg\max_{\boldsymbol{P}'_t \in \mathcal{P}} \mathbb{E}[S(\boldsymbol{P}'_t, Y_t) \mid \boldsymbol{P}_t]$$

for $t \in \mathbb{N}$. $\boldsymbol{P}^*_t$ is a random variable that maximizes the expectation of $S(\boldsymbol{P}^*_t, Y_t)$ before Y_t is drawn, but conditional on $\boldsymbol{P}_t$.

Definition 2 (Regret)**.** The regret of agent $(\boldsymbol{P}_t)_t$ at time T is

$$\mathrm{Regret}(T) \coloneqq \sum_{t=1}^{T} S(\boldsymbol{P}^*_t, Y_t) - S(\boldsymbol{P}_t, Y_t).$$

The agent is said to have *sublinear regret* or *no-regret* if

$$\limsup_{T \to \infty} \frac{1}{T} \mathrm{Regret}(T) \leq 0.$$

First, note that we define regret relative to the best expert in expectation instead of the best expert in hindsight. The latter would always be the one that made confident predictions and accidentally got all predictions exactly right. We are interested

in algorithms with sublinear regret, and for that purpose it would be too much to ask the agent to perform well compared to the best expert in hindsight. Moreover, for scoring rules that are symmetric between the outcomes, this expert would have a constant score C. This would imply that $\text{Regret}(T) = \sum_{t=1}^{T} C - S(\boldsymbol{P}_t, Y_t)$ and reduce the problem to minimizing the negative score, which would lead to performatively optimal predictions.

Second, we evaluate the performance of the expert with respect to the environment outcomes Y_t generated by the agent $(\boldsymbol{P}_t)_t$, instead of evaluating the expert according to outcomes $\tilde{Y}_t \sim f(\boldsymbol{P}_t^*)$ generated using the expert's own predictions. This means that, to receive sublinear regret, the agent only has to make accurate predictions—it does not have to find a performatively optimal prediction. This is different from the no-regret learning setup discussed in Jagadeesan et al. [2022], where regret is defined with respect to $S(\boldsymbol{P}_t^*, f(\boldsymbol{P}_t^*))$. In that setting, only agents converging to performatively optimal predictions have sublinear regret.

We begin by showing that the best expert in expectation actually exists, and that $\boldsymbol{P}_t^* = f(\boldsymbol{P}_t)$.

Proposition 10. *Let S be a proper scoring rule and $(\boldsymbol{P}'_t)_t \in \mathcal{P}$ an expert. Then for any $t \in \mathbb{N}$, we have*

$$\mathbb{E}[S(\boldsymbol{P}'_t, Y_t)] = \mathbb{E}[S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))].$$

Moreover, we have $(\boldsymbol{P}_t^*)_t = (f(\boldsymbol{P}_t))_t$ and thus

$$\text{Regret}(T) = \sum_{t=1}^{T} S(f(\boldsymbol{P}_t), Y_t) - S(\boldsymbol{P}_t, Y_t).$$

Proof. Let $t \in \mathbb{N}$ and let $(\boldsymbol{P}'_t)_t \in \mathcal{P}$ be any expert. Conditional on $\boldsymbol{P}_t$, $Y_t \sim f(\boldsymbol{P}_t)$ and Y_t is independent of $\boldsymbol{P}'_t$ by assumption. Hence,

$$\mathbb{E}\left[S(\boldsymbol{P}'_t, Y_t)\right] = \mathbb{E}\left[\mathbb{E}[S(\boldsymbol{P}'_t, Y_t) \mid \boldsymbol{P}_t, \boldsymbol{P}'_t]\right] = \mathbb{E}\left[S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))\right].$$

Next, since S is proper,

$$\mathbb{E}\left[S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))\right] \leq \mathbb{E}\left[S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t))\right].$$

It follows that

$$\max_{(\boldsymbol{P}'_t)_t \in \mathcal{P}} \mathbb{E}\left[S(\boldsymbol{P}'_t, Y_t)\right] = \max_{(\boldsymbol{P}'_t)_t \in \mathcal{P}} \mathbb{E}\left[S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))\right] \leq \mathbb{E}\left[S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t))\right] = \mathbb{E}\left[S(f(\boldsymbol{P}_t), Y_t)\right].$$

Moreover, $(f(\boldsymbol{P}_t))_t \in \mathcal{P}$, as $f(\boldsymbol{P}_t)$ is constant given $\boldsymbol{P}_t$ and thus independent of Y_t.

It follows that, for any $t \in \mathbb{N}$, $\boldsymbol{P}_t^* \in \arg\max_{(\boldsymbol{P}'_t)_t \in \mathcal{P}} \mathbb{E}[S(\boldsymbol{P}'_t, Y_t)]$, and thus

$$\text{Regret}(T) = \sum_{t=1}^{T} S(f(\boldsymbol{P}_t), Y_t) - S(\boldsymbol{P}_t, Y_t).$$

□

D.4.1 Characterization of regret in the limit

If S is unbounded (such as the log scoring rule), then the agent's scores can become arbitrarily low, and the limit of $\frac{1}{T}\text{Regret}(T)$ may be undefined. To simplify our analysis, we will thus assume that there is a bound on the variance of the received score $S(\boldsymbol{P}'_t, Y_t)$ and on the expected score $S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))$ of both the agent, $\boldsymbol{P}'_t = \boldsymbol{P}_t$, and the best expert, $\boldsymbol{P}'_t = \boldsymbol{P}_t^*$. In the case of the log scoring rule, this would be satisfied, for instance, if the agent's predictions are bounded away from the boundary of the probability simplex.

Our next proposition shows that, given these assumptions, $\lim_{T\to\infty} \frac{1}{T}\text{Regret}(T)$ exists and is nonnegative, and having sublinear regret is equivalent to $\lim_{t\to\infty} \frac{1}{T}\text{Regret}(T) = 0$.

Proposition 11. *Let S be a proper scoring rule. Assume that $\sup_t |S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))| < \infty$ and that $\sup_t \text{Var}(S(\boldsymbol{P}'_t, Y_t)) < \infty$ for $\boldsymbol{P}'_t \in \{\boldsymbol{P}_t, f(\boldsymbol{P}_t)\}$. Then almost surely*

$$\lim_{T\to\infty} \frac{1}{T}\text{Regret}(T) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t)) \geq 0.$$

In particular, almost surely both limits exist and are finite, and the agent has sublinear regret if and only if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t)) = 0.$$

Proof. We will use a version of the strong law of large numbers for uncorrelated random variables with bounded variance, adapted from Neely [2021, Theorem 2].

Theorem 9 (Neely [2021], Theorem 2)**.** *Let $\{X_t\}_{t\in\mathbb{N}_0}$ be a sequence of pairwise uncorrelated random variables with mean 0 and bounded variances. I.e., assume that*

1. $\mathbb{E}[X_t] = 0$ *for all* $t \in \mathbb{N}_0$
2. *There exists* $c > 0$ *such that* $\operatorname{Var}(X_t) \leq c$ *for all* $t \in \mathbb{N}_0$
3. $\operatorname{Cov}(X_t, X_{t'}) = 0$ *for all* $t \neq t' \in \mathbb{N}_0$.

Then almost surely

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} X_t = 0.$$

We will apply this law to random variables $X_t := S(\boldsymbol{P}'_t, Y_t) - S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))$, where $\boldsymbol{P}'_t$ is either $\boldsymbol{P}_t$ or $f(\boldsymbol{P}_t)$.

First, by Proposition 10, $\mathbb{E}[X_t] = \mathbb{E}[S(\boldsymbol{P}'_t, Y_t) - S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))] = 0$. Second, by assumption,

$$\sup_t \operatorname{Var}(S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))) < \infty.$$

Hence, also

$$\sup_t \operatorname{Var}(S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))) = \sup_t \operatorname{Var}(\mathbb{E}[S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t)) \mid \boldsymbol{P}_t]) \leq \sup_t \operatorname{Var}(S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))) < \infty.$$

It follows that also $\sup_t \operatorname{Var}(X_t) < \infty$.

Third, we know that Y_t is independent of $\boldsymbol{P}_{t'}$ and $Y_{t'}$ for $t > t'$, conditional on $\boldsymbol{P}_t$. Moreover, $\boldsymbol{P}'_t$ is constant given $\boldsymbol{P}_t$. Hence, given $\boldsymbol{P}_t$, also $X_t = S(\boldsymbol{P}'_t, Y_t) - S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))$ is independent of $X_{t'}$. Moreover,

$$\mathbb{E}[X_t \mid \boldsymbol{P}_t] = \mathbb{E}[S(\boldsymbol{P}'_t, Y_t) - S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t)) \mid \boldsymbol{P}_t] = S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t)) - S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t)) = 0.$$

It follows for $t > t'$ that

$$\operatorname{Cov}(X_t, X_{t'}) = \mathbb{E}[X_t X_{t'}] = \mathbb{E}[\mathbb{E}[X_t X_{t'} \mid \boldsymbol{P}_t]] = \mathbb{E}[\mathbb{E}[X_t \mid \boldsymbol{P}_t]\mathbb{E}[X_{t'} \mid \boldsymbol{P}_t]] = 0.$$

This shows all conditions of the theorem and thus

$$\lim_{t\to\infty} \frac{1}{T} \sum_{t=1}^{T} X_t = 0$$

almost surely.

Now we turn to the limit of $\frac{1}{T}\sum_{t=1}^{T} S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))$. By assumption, $\sup_t |S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))| < \infty$, so this limit exists and is finite. Thus, almost surely

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t)) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} S(\boldsymbol{P}'_t, Y_t) - X_t$$

$$= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} S(\boldsymbol{P}'_t, Y_t) - \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} X_t = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} S(\boldsymbol{P}'_t, Y_t).$$

Using Proposition 10, it follows that almost surely

$$\lim_{T\to\infty}\frac{1}{T}\mathrm{Regret}(T) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} S(f(\boldsymbol{P}_t), Y_t) - S(\boldsymbol{P}_t, Y_t) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} S(f(\boldsymbol{P}_t), Y_t) - \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} S(\boldsymbol{P}_t, Y_t)$$

$$= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} S(\boldsymbol{P}_t, f(\boldsymbol{P}_t)) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t)).$$

Turning to the "in particular" part, note that this limit is finite by the above, and it is nonnegative since S is assumed to be proper. Moreover, it follows that almost surely

$$\limsup_{T\to\infty}\frac{1}{T}\mathrm{Regret}(T) = \lim_{T\to\infty}\frac{1}{T}\mathrm{Regret}(T) \geq 0.$$

Thus, almost surely $\limsup_{T\to\infty}\frac{1}{T}\mathrm{Regret}(T) \leq 0$ if and only if $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t)) = 0$. This concludes the proof. □

D.4.2 Sublinear regret ⇔ sublinear prediction error

Now we turn to the main result of this section. We show that given our assumptions, agents have sublinear regret if and only if their prediction error is sublinear. Note that here, we do *not* require the $\boldsymbol{P}_t$ to converge; they could also oscillate between different fixed points.

Theorem 10. *Let $(\boldsymbol{P}_t)_t$ be the sequence of the agent's predictions and S a strictly proper scoring rule. Assume that $\sup_t \mathrm{Var}(S(\boldsymbol{P}'_t, Y_t)) < \infty$ for $\boldsymbol{P}'_t \in \{\boldsymbol{P}_t, f(\boldsymbol{P}_t)\}$, and assume that there exists a closed set $\mathcal{C} \subseteq \Delta(\mathcal{N})$ such that $\boldsymbol{P}_t \in \mathcal{C}$ for all t and $S(\boldsymbol{p}, f(\boldsymbol{p}))$, $S(f(\boldsymbol{p}), f(\boldsymbol{p}))$, and $f(\boldsymbol{p})$ are continuous in $\boldsymbol{p}$ at any $\boldsymbol{p} \in \mathcal{C}$. Then almost surely the agent has sublinear regret if and only if $\sum_{t=1}^{T} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\|$ is sublinear, i.e., if $\lim_{t\to\infty}\frac{1}{T}\sum_{t=1}^{T} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\| = 0$.*

To show the result, we begin by proving an analytic lemma.

Lemma 11. *Let $\varphi, \psi\colon \mathbb{N} \to [0, \infty)$ and assume there exists a constant $C > 0$ such that for all $t \in \mathbb{N}$, we have $\psi(t) \leq C$. Assume that for any $\epsilon > 0$, there exists $\delta > 0$ such that if $\psi(t) > \epsilon$ for any $t \in \mathbb{N}$, then $\varphi(t) > \delta$. Then*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\varphi(t) = 0 \Rightarrow \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\psi(t) = 0.$$

Proof. We prove the contrapositive. That is, we assume that there exists some constant $c > 0$ such that there are infinitely many $T \in \mathbb{N}$ such that $\frac{1}{T}\sum_{t=1}^{T}\psi(t) > c$. Let $\mathcal{T}$ be the set of such T. We show that then there exists a constant $c' > 0$ such that for infinitely many T, $\frac{1}{T}\sum_{t=1}^{T}\varphi(t) > c'$.

Let $T \in \mathcal{T}$. Since by assumption $\frac{1}{T}\sum_{t=1}^{T}\psi(t) > c$, it follows that $\sum_{t=1}^{T}\frac{\psi(t)}{c} > T$. Let $C' := \max\{C, 1/c\} + 1$. Since $\psi(t) < C'$ it must be $\psi(t) > \frac{c}{2C}$ for more than $\frac{c}{2C}$ fraction of the times $t \leq T$. Otherwise, it would be

$$\sum_{t=1}^{T}\frac{\psi(t)}{c} \leq T\left(\frac{C}{c}\frac{c}{2C} + \left(1 - \frac{c}{2C}\right)\frac{\epsilon}{2C}\right) \leq T\left(\frac{1}{2} + \frac{\epsilon}{2C}\right) < T.$$

By assumption, this gives us a $\delta > 0$ such that whenever $\psi(t) > \epsilon := \frac{c}{2C}$, also $\varphi(t) > \delta$. In particular, this applies to at least ϵ fraction of $t \leq T$. Hence, it follows that for any $T \in \mathcal{T}$,

$$\sum_{t=1}^{T}\varphi(t) \geq \delta\epsilon T.$$

This shows that there are infinitely many T such that $\frac{1}{T}\sum_{t=1}^{T}\varphi(t) > \delta\epsilon$ and thus concludes the proof. □

Proof of Theorem 10. To begin, note that since $\mathcal{C} \subseteq \Delta(\mathcal{N})$ is closed and $\Delta(\mathcal{N})$ compact, also $\mathcal{C}$ is compact. Hence, continuity of $S(\boldsymbol{p}, f(\boldsymbol{p}))$ and $S(f(\boldsymbol{p}), f(\boldsymbol{p}))$ implies that both are also bounded on $\mathcal{C}$ and thus $\sup_t |S(\boldsymbol{P}'_t, f(\boldsymbol{P}_t))| < \infty$ for $\boldsymbol{P}'_t \in \{\boldsymbol{P}_t, f(\boldsymbol{P}_t)\}$. Hence, by our assumptions, the conditions for Proposition 11 are satisfied.

"$\Rightarrow$". Assume $\mathrm{Regret}(T)$ is sublinear. We want to show that then $\sum_{t=1}^{T} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\|$ is sublinear. To do this, we will apply Lemma 11.

To begin, define $\varphi(t) := S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t))$ and note that $\varphi(t) \geq 0$ since S is proper. By Proposition 11, it follows that if $\mathrm{Regret}(T)$ is sublinear, also $\sum_{t=1}^{T} \varphi(t)$ is sublinear almost surely. For brevity, we omit the "almost surely" qualification in the following.

Next, define $\psi(t) := \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\|$, and note that $0 \leq \psi(t) \leq n$. Next, let $\epsilon > 0$ arbitrary. To apply Lemma 11 to φ and ψ, it remains to show that there exists $\delta > 0$ such that whenever $\psi(t) \geq \epsilon$, then $\varphi(t) \geq \delta$.

To that end, let

$$\delta := \min_{\{\boldsymbol{p} \in \mathcal{C} \mid \|\boldsymbol{p} - f(\boldsymbol{p})\| \geq \epsilon\}} S(f(\boldsymbol{p}), f(\boldsymbol{p})) - S(\boldsymbol{p}, f(\boldsymbol{p})).$$

Since $\|\cdot\|$ is continuous and f is continuous at any $\boldsymbol{p} \in \mathcal{C}$, the set $\{\boldsymbol{p} \in \mathcal{C} \mid \|\boldsymbol{p} - f(\boldsymbol{p})\| \geq \epsilon\}$ is closed and thus compact. Moreover, $S(f(\boldsymbol{p}), f(\boldsymbol{p}))$ and $S(\boldsymbol{p}, f(\boldsymbol{p}))$ are continuous by assumption, and thus the minimum is attained at some point $\hat{\boldsymbol{p}} \in \mathcal{C}$. But since S is strictly proper, it follows $\delta = S(f(\hat{\boldsymbol{p}}), f(\hat{\boldsymbol{p}})) - S(\hat{\boldsymbol{p}}, f(\hat{\boldsymbol{p}})) > 0$. Hence, since $\boldsymbol{P}_t \in \mathcal{C}$ for any $t \in \mathbb{N}$, it follows that whenever $\varphi(t) \geq \epsilon$, it follows

$$\varphi(t) = S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t)) \geq \delta.$$

This shows all conditions for Lemma 11. Hence, we conclude that $\lim_{t\to\infty} \frac{1}{T} \sum_{t=1}^{T} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\| = 0$.

"$\Leftarrow$". Let $\varphi(t) := \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\|$ and $\psi := S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t))$. We assume that $\sum_{t=1}^{T} \varphi(t)$ is sublinear in T and want to show that then $\mathrm{Regret}(T)$ is sublinear as well. To do so, we will show that $\sum_{t=1}^{T} \psi(t)$ is sublinear using our lemma, and then the required statement follows again from Proposition 11.

Now we have to show the conditions of the lemma. First, as before, $\varphi(t) \geq 0$ and $\psi(t) \geq 0$. Second, as noted in the beginning, we have $\sup_t \psi(t) < \infty$ by our assumption that $S(f(\boldsymbol{p}), f(\boldsymbol{p}))$ and $S(\boldsymbol{p}, f(\boldsymbol{p}))$ are continuous on $\mathcal{C}$. Now let $\epsilon > 0$ arbitrary. Assume that $S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t)) > \epsilon$ for some $\epsilon > 0$ and $t \in \mathbb{N}$.

Consider the set $\mathcal{C}' := \{\boldsymbol{p} \in \mathcal{C} \mid S(f(\boldsymbol{p}), f(\boldsymbol{p})) - S(\boldsymbol{p}, f(\boldsymbol{p})) \geq \epsilon\}$. Since $S(f(\boldsymbol{p}), f(\boldsymbol{p}))$ and $S(\boldsymbol{p}, f(\boldsymbol{p}))$ are continuous on $\mathcal{C}$ by assumption, this set is compact. Moreover, the function $\boldsymbol{p} \in \mathcal{C} \mapsto \|\boldsymbol{p} - f(\boldsymbol{p})\|$ is continuous since f is continuous on $\mathcal{C}$ by assumption. Hence, the minimum $\delta := \min_{\boldsymbol{p} \in \mathcal{C}'} \|\boldsymbol{p} - f(\boldsymbol{p})\|$ is attained at some point $\hat{\boldsymbol{p}} \in \mathcal{C}'$.

Now, if $\delta = 0$, we would have $\hat{\boldsymbol{p}} = f(\hat{\boldsymbol{p}})$ and thus

$$S(f(\hat{\boldsymbol{p}}), f(\hat{\boldsymbol{p}})) - S(\hat{\boldsymbol{p}}, f(\hat{\boldsymbol{p}})) = S(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}) - S(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}) = 0 < \epsilon,$$

which is a contradiction. Hence, $\delta > 0$. Since $\boldsymbol{P}_t \in \mathcal{C}$, it follows from $S(f(\boldsymbol{P}_t), f(\boldsymbol{P}_t)) - S(\boldsymbol{P}_t, f(\boldsymbol{P}_t)) \geq \epsilon$, for $t \in \mathbb{N}$ that $\boldsymbol{P}_t \in \mathcal{C}'$ and thus $\|\boldsymbol{P}_t - f(\boldsymbol{P}_t)\| \geq \delta$. This shows the third condition for the lemma. We can thus conclude that $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \psi(t) = 0$. Using Proposition 11, this concludes the proof. □

D.4.3 Convergence to fixed points

The next result shows that if the agent's predictions converge to some distribution $\boldsymbol{p}$, then $\boldsymbol{p}$ must be a fixed point.

Corollary 1. *In addition to the assumptions from Theorem 10, assume that $\boldsymbol{P}_t$ converges almost surely to a limit $\lim_{t\to\infty} \boldsymbol{P}_t = \boldsymbol{p}^*$. Then almost surely $\boldsymbol{p}^*$ is a fixed point if and only if the agent has sublinear regret.*

Proof. By Theorem 10, almost surely the agent has sublinear regret if and only if

$$\lim_{t\to\infty} \frac{1}{T} \sum_{t=1}^{T} \|f(\mathbf{P}_t) - \mathbf{P}_t\| = 0.$$

It remains to show that, given that the $\boldsymbol{P}_t$ converge, the latter is equivalent to convergence to a fixed point.

Since $\mathcal{C}$ is compact and $\boldsymbol{P}_t \in \mathcal{C}$ for all $t \in \mathbb{N}$, also $\boldsymbol{p}^* \in \mathcal{C}$. Hence, f is continuous at $\boldsymbol{p}^*$, so

$$\|f(\boldsymbol{p}^*) - \boldsymbol{p}^*\| = \left\| f\left(\lim_{t\to\infty} \boldsymbol{P}_t\right) - \lim_{t\to\infty} \boldsymbol{P}_t \right\| = \lim_{t\to\infty} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\|.$$

Since this sequence converges, it is equal to its Cesàro mean,

$$\lim_{t\to\infty} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\| = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\|.$$

Hence,

$$\|f(\boldsymbol{p}^*) - \boldsymbol{p}^*\| = \lim_{t\to\infty} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\| = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\|.$$

It follows that, if $\lim_{t\to\infty} \boldsymbol{P}_t = \boldsymbol{p}^*$, then

$$\|f(\boldsymbol{p}^*) - \boldsymbol{p}^*\| = 0 \Leftrightarrow \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\| = 0.$$

This shows that, almost surely, $\boldsymbol{p}^*$ is a fixed point, if and only if $\sum_{t=1}^{T} \|f(\boldsymbol{P}_t) - \boldsymbol{P}_t\|$ is sublinear. □

D.5 PREDICTION MARKETS

Lastly, we consider prediction markets. We assume a simplified model of a prediction market, in which traders submit a single prediction and get scored using a proper scoring rule. The prediction that is output by the market and that influences the outcome is just a weighted average of the individual traders' predictions. In this situation, if a trader has a small weight and can thus barely influence the market prediction, the trader's score will mostly be determined by the accuracy of the report, rather than the influence of the report on the market. Thus, if all traders are small relative to the market, the equilibrium prediction will be close to a fixed point.

A similar result was shown by Hardt et al. [2022] in the performative prediction context. They define a firm's performative power as the degree to which the firm can influence the overall outcome with their prediction. Hardt et al. show that in an equilibrium, the distance between a player's (performatively optimal) equilibrium strategy and their strategy when optimizing loss against the fixed equilibrium distribution (here, this means predicting the market probability) is bounded by the power of the trader. We give an analogous result for our formal setting and assumptions.

To formalize the setting, assume that there are N players. We associate with each player $n \in [N]$ a number $w_n \in [0, 1]$ s.t. $\sum_n w_n = 1$, representing, intuitively, what fraction of the overall capital in the market is provided by player n. In the game, all players simultaneously submit a probability distribution $\boldsymbol{p}_n$. Then the event Y is sampled according to the distribution $\boldsymbol{q} = f(\sum_n w_n \boldsymbol{p}_n)$. Finally, each player is scored in proportion to $S(\boldsymbol{p}_n, Y)$ for some strictly proper scoring rule S. Typical market scoring rules would consider terms like $S(\boldsymbol{p}_n, Y) - S(\boldsymbol{p}_n, Y)$, but subtracting $S(\boldsymbol{p}_n, Y)$ (or multiplying by constants) does not matter for the game. We assume that players maximize their expected score, $\mathbb{E}[S(\boldsymbol{p}_n, Y)] = S(\boldsymbol{p}_n, f(\sum_m w_m \boldsymbol{p}_m))$.

For discussions of market scoring rules, see Hanson [2003] and Pennock and Sami [2007]. Prior work has connected these market scoring rules to more realistic prediction markets that trade Arrow–Debreu securities markets such as PredictIt [e.g., Hanson, 2003; Pennock and Sami, 2007, Section 4; Chen and Pennock, 2007; Agrawal et al., 2009].

We assume that f is common knowledge. Moreover, in the following we only consider pure strategy equilibria, and we do not investigate the existence of equilibria.

Theorem 11. *Let S be a proper scoring rule and let G, g as in the Gneiting and Raftery characterization of S. Let $(\boldsymbol{p}_n)_n$ be a pure strategy Nash equilibrium of the aforedefined game and let $\hat{\boldsymbol{p}} := \sum_n w_n \boldsymbol{p}_n$ be the market prediction. Assume f is differentiable at $\hat{\boldsymbol{p}}$. For any player n, if G, g are differentiable at $\boldsymbol{p}_n$ and $Dg(\boldsymbol{p}_n) \succ \gamma_{\boldsymbol{p}_n}$, it follows that*

$$\|f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n\| \leq \frac{w_n \|Df(\hat{p})\|_{\mathrm{op}} \|g(\boldsymbol{p}_n)\|}{\gamma_{\boldsymbol{p}_n}}.$$

In particular, this theorem shows that players n with very low w_n (little capital/influence on $\boldsymbol{q}$) will accurately predict $\boldsymbol{q} = f(\hat{\boldsymbol{p}})$. Note, however, that $\hat{\boldsymbol{p}}$ is not necessarily a fixed point or close to a fixed point. If there are are also players n with very high w_n, then their prediction and the overall market prediction may be wrong. (So interestingly the overall market probability $\hat{\boldsymbol{p}} = \sum_n w_n \boldsymbol{p}_n$ is worse than the prediction of individuals. One might take this to suggest that anyone interested in $\boldsymbol{q}$ should look at the latter type of predictions. Of course, if this is what everyone does, it is not so clear anymore that the model $\boldsymbol{q} = f(\sum_n w_n \boldsymbol{p}_n)$ is accurate.)

Proof. The proof is analogous to that of Theorem 3. Let $(\boldsymbol{p}_n)_n$ be a pure strategy Nash equilibrium and $\hat{\boldsymbol{p}} := \sum_m w_m \boldsymbol{p}_m$. Each player must play a best response to the other player's strategies, so $\boldsymbol{p}_n$ must be a global maximum of the function $\varphi\colon \boldsymbol{p}_n \mapsto S(\boldsymbol{p}_n, \sum_m w_m \boldsymbol{p}_m)$ Hence, it must be $\nabla\varphi(\boldsymbol{p}_n)^\top (f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) \leq 0$, i.e., the directional derivative of φ in the direction $f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n)$ must be at most zero. Otherwise, player n could improve their loss by changing their prediction marginally towards $f(\hat{\boldsymbol{p}})$.

Computing the gradient, we have

$$\begin{aligned}
\nabla_{\boldsymbol{p}_n}\left(S\left(\boldsymbol{p}_n, f\left(\sum_m w_m \boldsymbol{p}_m\right)\right)\right) &= \nabla_{\boldsymbol{p}_n}\left(G(\boldsymbol{p}_n) + g(\boldsymbol{p}_n)^\top\left(f\left(\sum_m w_m \boldsymbol{p}_m\right) - \boldsymbol{p}_n\right)\right) \\
&= g(\boldsymbol{p}_n) + Dg(\boldsymbol{p}_n)^\top\left(f\left(\sum_m w_m \boldsymbol{p}_m\right) - \boldsymbol{p}_n\right) + w_n Df(\hat{\boldsymbol{p}})^\top g(\boldsymbol{p}_n) - Ig(\boldsymbol{p}_n) \\
&= Dg(\boldsymbol{p}_n)^\top (f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) + w_n Df(\hat{\boldsymbol{p}})^\top g(\boldsymbol{p}_n).
\end{aligned}$$

It follows

$$0 \geq \nabla\varphi(\boldsymbol{p}_n)^\top (f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) = (f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n)^\top Dg(\boldsymbol{p}_n)(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) + w_n g(\boldsymbol{p}_n)^\top Df(\hat{\boldsymbol{p}})(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) \tag{22}$$

$$\Rightarrow - w_n g(\boldsymbol{p}_n)^\top Df(\hat{\boldsymbol{p}})(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) \geq (f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n)^T (Dg(\boldsymbol{p}_n))(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n). \tag{23}$$

Using that $Dg(\boldsymbol{p}_n)|_{\mathcal{T}} \succ \gamma_{\boldsymbol{p}_n}$ and thus $(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n)^\top (Dg(\boldsymbol{p}_n))(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) \geq \gamma_{\boldsymbol{p}} \|f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n\|^2$, it follows that

$$\begin{aligned}
& \gamma_{\boldsymbol{p}_n} \|f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n\|^2 \\
\leq\ & (f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n)^\top Dg(\boldsymbol{p}_n)(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) \\
\leq\ & -w_n g(\boldsymbol{p}_n)^\top Df(\hat{\boldsymbol{p}})(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n) \\
\leq\ & w_n |g(\boldsymbol{p}_n)^\top Df(\hat{\boldsymbol{p}})(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n)| \\
\underset{\text{Cauchy-Schwarz}}{\leq}\ & w_n \|g(\boldsymbol{p}_n)\| \|Df(\hat{\boldsymbol{p}})(f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n)\| \\
\leq\ & w_n \|g(\boldsymbol{p}_n)\| \|Df(\hat{\boldsymbol{p}})\|_{\mathrm{op}} \|f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n\|
\end{aligned}$$

The result follows by dividing by $\gamma_{\boldsymbol{p}_n} \|f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n\|$.

□

Corollary 2. *In addition to the assumptions from Theorem 11, assume that f is Lipschitz-continuous and $C := \sup_{p \in \Delta(\mathcal{N})} \frac{\|g(\boldsymbol{p})\|}{\gamma_{\boldsymbol{p}}} < \infty$. Let $(\boldsymbol{p}_n)_n$ be a Nash equilibrium and let $\epsilon > 0$ arbitrary. Then there exists a $\delta > 0$ such that if for all n, $w_n < \delta$, all of $\boldsymbol{p}_n$ and $f(\boldsymbol{p}_n)$, for all i, as well as $\sum_m w_m \boldsymbol{p}_m$ and $f(\sum_m w_m \boldsymbol{p}_m)$ are within ϵ of each other.*

Proof. Let $\epsilon > 0$ arbitrary. Let L_f be the Lipschitz constant of f and note that then $\|Df(\boldsymbol{p})\|_{\mathrm{op}} \leq L_f$ for all $\boldsymbol{p} \in \Delta(\mathcal{N})$. By Theorem 11, it follows for $\hat{\boldsymbol{p}} := \sum_m w_m \boldsymbol{p}_m$ and any player n that

$$\|f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n\| \leq w_n L_f C.$$

Now let $\lambda := \min(\{1, \frac{1}{L_f}\})$ and $\delta := \frac{\epsilon\lambda}{4CL_f}$, and assume $w_n < \delta$ for all $n \in [N]$. Then it follows

$$\|f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n\| \leq \delta L_f C \leq \frac{\lambda}{4}\epsilon.$$

Moreover, since $\hat{\boldsymbol{p}}$ is a convex combination of probabilities $\boldsymbol{p}_n$, it follows that

$$\|f(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\| \leq \max_n \|f(\hat{\boldsymbol{p}}) - \boldsymbol{p}_n\| \leq \frac{\lambda}{4}\epsilon.$$

Thus, by the triangle equality, we have $\|\boldsymbol{p}_n - \hat{\boldsymbol{p}}\| \leq \frac{2\lambda}{4}\epsilon$, and since f is Lipschitz-continuous,

$$\|f(\hat{p}) - f(\boldsymbol{p}_n)\| \leq L_f \|\hat{\boldsymbol{p}} - \boldsymbol{p}_n\| \leq L_f \frac{2\lambda}{4}\epsilon \leq \frac{1}{2}\epsilon$$

for any $n \in [N]$.

This shows that all of $\boldsymbol{p}_n, \hat{\boldsymbol{p}}, f(\boldsymbol{p}_n)$ are within $\epsilon/2$ of $f(\hat{\boldsymbol{p}})$ and thus by the triangle inequality within ϵ of each other. □

It would be interesting to extend these results. For example, it is unclear what happens when players make predictions *repeatedly*. (To keep things simple, one should probably still imagine that all players know f and that the environment probability is determined by f applied to the majority forecast. If the traders have private information, prediction markets become harder to analyze. For some discussions, see Ostrovsky [2009], Chen and Waggoner [2016].)

References

Shipra Agrawal, Erick Delage, Mark Peters, Zizhuo Wang, and Yinyu Ye. A unified framework for dynamic pari-mutuel information market design. In *EC '09 Proceedings of the 10th ACM conference on Electronic commerce*, pages 255–264. 2009.

Stuart Armstrong. Standard ML Oracles vs counterfactual ones. AI Alignment Forum, 2018. URL `https://www.alignmentforum.org/posts/hJaJw6LK39zpyCKW6/standard-ml-oracles-vs-counterfactual-ones`.

Jean-Marc Azaïs and Mario Wschebor. *Level sets and extrema of random processes and fields*. John Wiley & Sons, 2009.

James Bell, Linda Linsefors, Caspar Oesterheld, and Joar Skalse. Reinforcement learning in newcomblike environments. *NeurIPS*, 34:22146–22157, 2021.

Yiling Chen and David M. Pennock. A utility framework for bounded-loss market makers. In *UAI'07 Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 49–56. 2007.

Yiling Chen and Bo Waggoner. Informational substitutes. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 239–247, 2016. doi: 10.1109/FOCS.2016.33.

Abram Demski. Partial agency. AI Alignment Forum, 2019. `https://www.alignmentforum.org/posts/4hdHto3uHejhY2F3Q/partial-agency`.

Jakob Foerster, Gregory Farquhar, Maruan Al-Shedivat, Tim Rocktäschel, Eric Xing, and Shimon Whiteson. Dice: The infinitely differentiable monte carlo estimator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1529–1538. PMLR, 2018.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Hilary Greaves. Epistemic decision theory. *Mind*, 122(488):915–952, 2013.

Robin Hanson. Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119, 2003.

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. In *NeurIPS*, 2022.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: performative gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4641–4650. PMLR, 2021.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9760–9785. PMLR, 2022.

Richard C Jeffrey. *The logic of decision*. University of Chicago press, 1990.

J. Kitchen. Concerning the convergence of iterates to fixed points. *Studia Mathematica*, 27(3):247–249, 1966.

David Krueger, Tegan Maharaj, and Jan Leike. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*, 2020.

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. In *International Conference on Learning Representations*, 2019.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *NeurIPS*, 33:4929–4939, 2020.

Michael J. Neely. Infinitely often, probability 1, Borel-Cantelli, and the law of large numbers, 2021. URL `https://viterbi-web.usc.edu/~mjneely/Borel-Cantelli-LLN.pdf`.

Michael Ostrovsky. Information aggregation in dynamic markets with strategic traders. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 253–254, 2009.

David M. Pennock and Rahul Sami. Computational aspects of prediction markets. In Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, editors, *Algorithmic Game Theory*, chapter 26, pages 651–675. Cambridge University Press, 2007.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7599–7609. PMLR, 2020.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Jonathan Uesato, Ramana Kumar, Victoria Krakovna, Tom Everitt, Richard Ngo, and Shane Legg. Avoiding tampering incentives in deep RL via decoupled approval. *arXiv preprint arXiv:2011.08827*, 2020.

Paul Weirich. Causal Decision Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.