Binary Scoring Rules that Incentivize Precision

ERIC NEYMAN, Columbia University, United States
GEORGY NOAROV, University of Pennsylvania, United States
S. MATTHEW WEINBERG, Princeton University, United States

All proper scoring rules incentivize an expert to predict *accurately* (report their true estimate), but not all proper scoring rules equally incentivize *precision*. Rather than treating the expert's belief as exogenously given, we consider a model where a rational expert can endogenously refine their belief by repeatedly paying a fixed cost, and is incentivized to do so by a proper scoring rule.

Specifically, our expert aims to predict the probability that a biased coin flipped tomorrow will land heads, and can flip the coin any number of times today at a cost of c per flip. Our first main result defines an *incentivization index* for proper scoring rules, and proves that this index measures the expected error of the expert's estimate (where the number of flips today is chosen adaptively to maximize the predictor's expected payoff). Our second main result finds the unique scoring rule which optimizes the incentivization index over all proper scoring rules.

We also consider extensions to minimizing the ℓ^{th} moment of error, and again provide an incentivization index and optimal proper scoring rule. In some cases, the resulting scoring rule is differentiable, but not infinitely differentiable. In these cases, we further prove that the optimum can be uniformly approximated by polynomial scoring rules.

Finally, we compare common scoring rules via our measure, and include simulations confirming the relevance of our measure even in domains outside where it provably applies.

CCS Concepts: • **Theory of computation** → *Algorithmic mechanism design.*

Additional Key Words and Phrases: proper scoring rules, information elicitation

ACM Reference Format:
Eric Neyman, Georgy Noarov, and S. Matthew Weinberg. 2021. Binary Scoring Rules that Incentivize Precision. In *Proceedings of the 22nd ACM Conference on Economics and Computation (EC '21), July 18–23, 2021, Budapest, Hungary.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3465456.3467639

EC '21, July 18–23, 2021, Budapest, Hungary

ACM ISBN 978-1-4503-8554-1/21/07…$15.00
https://doi.org/10.1145/3465456.3467639

1 Introduction

In the context of decision theory, a *scoring rule* rewards predictors for the accuracy of their predictions [2, 9, 15]. In the context of a binary choice (e.g. "Will it rain tomorrow?"), a scoring rule can be thought of as a function $f : (0, 1) \to \mathbb{R}$, where if a predictor reports a probability p of rain, then the predictor's reward is $f(p)$ if it rains and $f(1 - p)$ if it does not rain.[1] We consider settings in which there are two possible outcomes that are treated symmetrically (as this definition assumes), and henceforth refer to scoring rules in terms of this function f. Traditionally, scoring rules are concerned with incentivizing *accurate* reports. For example, a scoring rule is called *proper* if a predictor is always incentivized to tell the truth, in the sense that reporting the predictor's true belief strictly maximizes the predictor's expected reward.

Of course, there is an extraordinary amount of flexibility in selecting a proper scoring rule. For example, if a continuously differentiable scoring rule $f : (0, 1) \to \mathbb{R}$ satisfies $xf'(x) = (1 - x)f'(1 - x)$ and $f'(x) > 0$ for all x, then f is proper. Any increasing C^1 function on $[\frac{1}{2}, 1)$ can therefore be extended to a C^1 proper scoring rule on $(0, 1)$ (see Corollary 2.6). Much prior work exists comparing proper scoring rules by various measures, e.g. [7, 8, 18], but there is little which formally analyzes the extent to which proper scoring rules incentivize *precision* (see Section 1.2 for a discussion of prior work).

As a motivating example, consider the problem of guessing the probability that one of two competing advertisements will be clicked. With zero effort, a predictor could blindly guess that each is equally likely. But the predictor is not exogenously endowed with this belief, they can also endogenously exert costly effort to refine their prediction. For example, the predictor could sample which ad they would click themselves, or poll members of their household for additional samples. A more ambitious predictor could run a crowdsourcing experiment, paying users to see which link they would click. Any proper scoring rule will equally incentivize the predictor to accurately report their resulting belief, but not all scoring rules equally incentivize the costly gathering of information.

We propose a simple model to formally measure the extent to which a scoring rule incentivizes costly refinement of the predictor's beliefs. Specifically, we consider a two-sided coin that comes up heads with probability p, and p is drawn uniformly from $(0, 1)$ (we refer to p as the bias of the coin). Tomorrow the coin will be flipped, and we ask the predictor to guess the probability that it lands heads. Today, the predictor can flip the coin (with bias p) any number of times, at cost c per flip. While we choose this model for its mathematical simplicity, it captures examples like the previous paragraph surprisingly well: tomorrow, a user will be shown the two advertisements (clicking one). Today, the predictor can run a crowdsourcing experiment and pay any number of workers c to choose between the two ads. This simple model also captures weather forecasting using ensemble methods surprisingly well, and we expand on this connection in Appendix A in the supplement.

With this model in mind, consider the following two extreme predictions: on one hand, the predictor could never flip the coin, and always output a guess of $1/2$. On the other, the predictor could flip the coin infinitely many times to learn p exactly, and output a guess of p. Note that both predictions are accurate: the predictor is truthfully reporting their belief, and that belief is correct given the observed flips. However, the latter prediction is more precise. All proper scoring rules incentivize the predictor to accurately report their true prediction in both cases, but different scoring rules incentivize the predictor to flip the coin a different number of times. More specifically, every scoring rule induces a different optimization problem for the predictor, thereby leading

[1] To be clear: if $f(x) = \ln(x)$ and a predictor predicts a probability of 0.7 to it raining, then the predictor receives reward $\ln(0.7)$ if it rains and $\ln(0.3)$ if it does not rain.

them to produce predictions of different quality. In this model, the key question we answer is the following: *which scoring rules best incentivize the predictor to produce a precise prediction?*

1.1 Our Results

Our first main result is the existence of an *incentivization index*. Specifically, if $\mathsf{Error}_c(f)$ denotes the expected error that a rational predictor makes when incentivized by scoring rule f with cost c per flip, we give a closed-form index $\mathsf{Ind}(f)$ with the following remarkable property: for all respectful (see Definition 3.1) proper scoring rules f and g, the inequality $\mathsf{Ind}(f) < \mathsf{Ind}(g)$ implies the existence of a sufficiently small $c_0 > 0$ such that $\mathsf{Error}_c(f) < \mathsf{Error}_c(g)$ for all $c \le c_0$ (Theorem 3.3). We formally introduce this index in Definition 3.2, but remark here that it is not a priori clear that such an index should exist at all, let alone that it should have a closed form.[2]

With an index in hand, we can now pose a well-defined optimization problem: which proper scoring rule minimizes the incentivization index? Our second main result nails down this scoring rule precisely; we call it $g_{1,\mathrm{Opt}}$ (see Theorem 4.1).

We also extend our results to the ℓ^{th} moment for $\ell \ge 1$, where now $\mathsf{Error}_c^{\ell}(f)$ denotes the expected ℓ^{th} power of the error that a rational predictor makes when incentivized by f with cost c per flip, and again derive an incentivization index $\mathsf{Ind}^{\ell}(f)$ and an optimal scoring rule $g_{\ell,\mathrm{Opt}}$.

Some optimal rules $g_{\ell,\mathrm{Opt}}$ have a particularly nice closed form (for example, as $\ell \to \infty$, the optimal rule pointwise converges to a polynomial), but many do not. We also prove, using techniques similar to the Weierstrass approximation theorem [17], that each of these rules can be approximated by polynomial scoring rules whose incentivization indices approach the optimum.

Finally, beyond characterizing the optimal rules, the incentivization indices themselves allow for comparison among popular scoring rules, such as logarithmic ($f_{\log}(x) := \ln(x)$), quadratic ($f_{\mathrm{quad}}(x) := 2x-(x^2+(1-x)^2)$), and spherical ($f_{\mathrm{sph}}(x) := x/\sqrt{x^2+(1-x)^2}$). We plot the predictions made by our incentivization index (which provably binds only as $c \to 0$) for various values of c, and also confirm via simulation that the index has bite for reasonable choices of c.

1.2 Related Work

To the best of our knowledge, [13] was the first to consider scoring rules as motivating the predictor to seek additional information about the distribution before reporting their belief. This direction is revisited in [6], and has gained more attention recently [10, 14, 16]. While these works (and ours) each study the same phenomenon, there is little technical overlap and the models are distinct: each explores a different aspect of this broad agenda. For example, [14] considers the predictor's incentive to outperform competing predictors (but there is no costly effort — the predictors' beliefs are still exogenous). [10] (which is contemporaneous and independent of our work) is the most similar in motivation, but still has significant technical differences (beyond the two subsequent examples). On one hand, their model is more general than ours in that they consider multi-dimensional state spaces (rather than binary ones, in our model). On another hand, it is more restrictive in that they consider only two levels of effort (versus infinitely many, in our model).

Our work also fits into the broad category of principal-agent problems. For example, works such as [3–5, 11] consider a learning principal who incentivizes agents to make costly effort and produce an accurate data point. Again, the models are fairly distinct, as these works focus on more sophisticated learning problems (e.g. regression), whereas we perform a more comprehensive dive into the problem of simply eliciting the (incentivized-to-be-precise) belief.

[2]Indeed, a priori it is possible that $\mathsf{Error}_{0.1}(f) < \mathsf{Error}_{0.1}(g)$, but $\mathsf{Error}_{0.01}(f) > \mathsf{Error}_{0.01}(g)$, and $\mathsf{Error}_{0.001}(f) < \mathsf{Error}_{0.001}(g)$, but $\mathsf{Error}_{0.0001}(f) > \mathsf{Error}_{0.0001}(g)$, and so on. The existence of an incentivization index rules out this possibility.

In summary, there is a sparse, but growing, body of work addressing the study of incentivizing effort in forming predictions, rather than just accuracy in reporting them. The above-referenced works pose various models to tackle different aspects of this agenda. In comparison, our model is arguably the simplest, and we develop a deep understanding of optimal scoring rules in this setting.

1.3 Summary and Roadmap

Section 2 lays out our model, and contains some basic facts to help build intuition for reasoning about the incentivization properties of scoring rules. Our main results are detailed in Sections 3 through 6, along with intuition for our techniques.

- Section 3 defines the incentivization index, and provides a sufficient condition (Definition 3.1) for the incentivization index to nail down the expected error of a rational predictor, up to $o(1)$. This is our first main result, which gives a framework to formally reason about scoring rules that incentivize precision.
- Section 4 finds the unique proper scoring rule which optimizes the incentivization index. This is our second main result, which finds novel scoring rules, and also sets a benchmark with which to evaluate commonly-studied scoring rules.
- Section 5 studies the optimal scoring rules from Section 4, and compares their incentivization indices to those of some well-known scoring rules. Appendix H in the supplement provides a few simulations confirming that Ind seems to have predictive value for $c \gg 0$.
- Section 6 proves that there exist polynomial scoring rules with incentivization indices arbitrarily close to the optimum.
- All sections additionally consider the expected ℓ^{th} power of the error for any $\ell \geq 1$.
- Section 7 concludes.

2 Model and Preliminaries

2.1 Scoring Rules and their Rewards

This paper considers predicting a binary outcome for tomorrow: heads or tails. The *expert* or *predictor* is asked to output a probability p with which they believe the coin will land heads. Tomorrow, should the coin land heads, their reward is $f(p)$; should it not, their reward is $f(1-p)$ (note that the reward is symmetric: it is invariant under swapping the labels 'heads' and 'tails'). Throughout this paper, we consider a scoring rule to be defined by this function $f(\cdot)$. Observe that if the expert believes the true probability of heads to be p, and chooses to guess x, then the expected reward is $p \cdot f(x) + (1-p) \cdot f(1-x)$.

Definition 2.1 (Expected Reward). *For scoring rule $f : (0,1) \to \mathbb{R}$, denote by $r_p^f(x) := p \cdot f(x) + (1-p) \cdot f(1-x)$ the expected reward of an expert who predicts x when their true belief is p.*

Let also $R^f(p) := r_p^f(p)$ be the expected reward of an expert who reports their true belief p. We may drop the superscript when the scoring rule f is clear from context.

A scoring rule is (weakly) proper if it (weakly) incentivizes accurate reporting. In our notation:

Definition 2.2. *A scoring rule $f : (0,1) \to \mathbb{R}$ is* proper *(resp.* weakly proper*) if for all $p \in (0,1)$, the expected reward function $r_p^f(x)$ is strictly (resp. weakly) maximal at $x = p$ on $(0,1)$.*

Note that the optimal scoring rules *designed* in this paper are all (strictly) proper. However, we will show them to be optimal even with respect to the larger class of weakly proper scoring rules.

2.2 Modeling the Expert's Behavior

We model the expert as Bayesian. Specifically, the expert initially believes the coin bias is uniformly distributed in $(0, 1)$. Today, the expert may flip the coin any number of times in order to gauge its true bias, and pays c per flip. After having flipped the coin n times, and seen k heads, the expert believes the true bias is $\frac{k+1}{n+2}$ (Fact C.2 in the supplemental appendix).[3] Once done flipping, the expert reports the coin bias. Tomorrow, the coin is flipped once, and the expert receives reward for the prediction based on the outcome via scoring rule f (known to the expert in advance), as described in Section 2.1.

It remains to define when the expert should stop flipping. Below, an *adaptive strategy* simply refers to a (possibly randomized) stopping rule for the expert, i.e. a rule that, given any number of past flips and the scoring rule f, tells the expert whether to stop or flip once again. The payoff of an adaptive strategy is simply the expert's expected reward for following that strategy, minus the expected number of coin flips.

Definition 2.3. *A* globally-adaptive expert *uses the payoff-maximizing adaptive strategy.*

Nailing down the expert's optimal behavior as a function of c is quite unwieldy. Thus, we derive our characterizations up to $o(1)$ terms (as $c \to 0$). When c is large, one may reasonably worry that these $o(1)$ terms render our theoretical results irrelevant. In Appendix H in the supplement we simulate the expert's optimal behavior for large c, and confirm that our results hold qualitatively in this regime.

Finally, we define a natural measure of precision for the expert's prediction.

Definition 2.4. *The* expected error *associated with a scoring rule f and cost c is* $\mathsf{Error}_c(f) := \mathbb{E}\left[|p - q|\right]$. *The expectation is taken over p, drawn uniformly from $(0, 1)$, and q, the prediction of a globally-adaptive expert after flipping the coin (q is a random variable which depends on f, p, c).*

We will also consider generalizations to other moments, and define $\mathsf{Error}_c^\ell(f) := \mathbb{E}[|p - q|^\ell]$.

2.3 Scoring Rule Preliminaries

Our proofs will make use of fairly heavy single-variable analysis, and therefore will require making some assumptions on $f(\cdot)$: continuity, differentiability, but also more technical ones. We will clearly state them when necessary, and confirm that all scoring rules of interest satisfy them. For these preliminaries, we need only assume that f is continuously differentiable so that everything which follows is well-defined. First, Lemma 2.5 provides an alternative characterization of proper (and weakly proper) scoring rules. The proof is in Appendix C in the supplement.

Lemma 2.5. *A continuously differentiable scoring rule f is weakly proper if and only if for all $p \in (0, 1)$, $pf'(p) = (1 - p)f'(1 - p)$ and $f'(p) \geq 0$. It is (strictly) proper if and only if additionally $f'(p) > 0$ almost everywhere[4] in $(0, 1)$.*

Corollary 2.6. *Let f be strictly increasing almost everywhere (resp., nondecreasing everywhere) and continuously differentiable on $(0, \frac{1}{2}]$. Then f can be extended to a continuously differentiable proper (resp., weakly proper) scoring rule on $(0, 1)$ by defining $f'(p) = \frac{1-p}{p} f'(1 - p)$ for $p \in (\frac{1}{2}, 1)$.*

Put another way: every continuously differentiable proper scoring rule can be defined by first providing a strictly increasing function on $(0, \frac{1}{2}]$, and then extending it as in Corollary 2.6. Remark C.3 in the supplemental appendix provides a short example to help parse this extension.

[3] By this, we mean the expert believes the coin would land heads with probability $\frac{k+1}{n+2}$, if it were flipped again.
[4] Almost everywhere on $(0, 1)$ refers to the interval $(0, 1)$ except a set of measure zero.

2.4 First Steps towards Understanding Incentivization

In this section, we state a few basic facts about the expert's expected reward, and how it changes with additional flips. We defer all proofs to Appendix C in the supplement. Reading these proofs may help a reader gain technical intuition for the model. Our analysis will focus mostly on the reward function $R^f(\cdot)$ rather than $f(\cdot)$, so the following fact will be useful:

Fact 2.7. *For a weakly proper scoring rule f, we have $(R^f)'(x) = f(x) - f(1-x)$ and $(R^f)''(x) = f'(x) + f'(1-x) = \frac{f'(x)}{1-x} \geq 0$ on $(0,1)$.*

Lemma 2.8 observes how this expected reward evolves with an additional flip.

Lemma 2.8. *If the expert has already flipped the coin n times, seeing k heads, then their expected increase in reward for exactly one additional flip is $\frac{k+1}{n+2}R^f(\frac{k+2}{n+3}) + \frac{n-k+1}{n+2}R^f(\frac{k+1}{n+3}) - R^f(\frac{k+1}{n+2})$.*

Lemma 2.8 suggests that the function $R^f(x)$ should be convex: if it were not, that would leave open the possibility of the expert potentially *losing* expected reward as a result of performing *more* flips (meaning that the expert might get a smaller reward for a better estimate of the coin bias).

Lemma 2.9 ([12]). *Let f be any proper (resp., weakly proper) scoring rule. Then $R^f(x)$ is strictly convex (resp., weakly convex) almost everywhere on $(0,1)$.*

Corollary 2.10. *Let f be a proper (resp., weakly proper) scoring rule. Then the expert's increased expected reward from an additional flip is strictly positive (resp., weakly positive).*

Because we are interested in incentivizing the expert to take costly actions, the scale of a proper scoring rule will also be relevant. For example, if f is proper, then so is $2f$, and $2f$ clearly does a better job of incentivizing the expert (Lemma 2.8). As such, we will want to first *normalize* any scoring rule under consideration to be on the same scale. A natural normalization is to consider two scoring rules to be on the same scale if expected payoff they provide to the expert is the same (where the expectation is taken over both the bias and the flips of the coin).

Definition 2.11. *We define $\mathrm{Cost}_c(f)$ to be the expected payoff to a globally-adaptive expert via scoring rule f (when the bias is drawn uniformly from $(0,1)$, and the expert may pay c per flip).*

Recall that the (expected) payoff of a perfect expert is $\int_0^1 R(x)dx$, since a perfect expert has expected payoff $R(x)$ if the coin has bias x, and the coin's bias is chosen uniformly from $[0,1]$. For proper (but not necessarily weakly proper) scoring rules, we show that as $c \to 0$ the expected payoff of a globally-adaptive expert approaches the payoff of a perfect expert. (This is true no matter the coin's bias, though we only need this result in expectation over the bias.) Intuitively, this is because the number of flips approaches ∞ as $c \to 0$, so the expert is rewarded as if they are perfect.

Proposition 2.12. *Let f be a proper scoring rule. Then $\lim_{c\to 0} \mathrm{Cost}_c(f) = \int_0^1 R(x)dx$. That is, $\mathrm{Cost}_c(f) = \int_0^1 R(x)dx \pm o(1)$.*

Assuming that two scoring rules f, g have $\mathrm{Cost}_c(f) = \mathrm{Cost}_c(g)$ addresses one potential scaling issue. But there is another issue as well: whenever f is proper, the scoring rule $2f - 1$ is also proper, and again clearly does a better job incentivizing the expert (again directly by Lemma 2.8). As such, we will also normalize so that $R^f(x) \geq 0$ for all x: the expert's expected reward is always non-negative if they are perfect. We conclude this section with a formal statement of this normalization. Appendix C in the supplement confirms the implications of the definition, and also contains a few lemmas stating equivalent conditions.

Definition 2.13. *A scoring rule $f(\cdot)$ is* normalized *if $\int_0^1 R^f(x)dx = 1$, and $f(1/2) = 0$. This implies that $\text{Cost}_c(f) = 1 \pm o(1)$, and that a perfectly calibrated expert gets non-negative expected reward. It also implies that an expert who flips zero coins gets zero expected reward.*

3 An Incentivization Index

This section presents our first main contribution: an incentivization index which characterizes the expert's expected error. The main result of this section, Theorem 3.3, requires scoring rules to be analytically nice in a specific way. We term such scoring rules *respectful*.

Definition 3.1. *A proper scoring rule f with reward function $R := R^f$ is* respectful *if:*
(1) R is strongly convex on $(0,1)$. That is, $R''(x) \geq a$ on $(0,1)$ for some $a > 0$.
(2) R''' is Riemann integrable on any closed subinterval of $(0,1)$.[5]
(3) $\exists t > \frac{1}{4}$, and $c_0 > 0$ s.th. for all $c \in (0, c_0)$: $|R'''(x)| \leq \frac{1}{c^{0.16}\sqrt{x(1-x)}} R''(x)$ on $[c^t, 1-c^t]$.[6]

Recall that R is strictly convex for any strictly proper scoring rule, so strong convexity is a minor condition. Likewise, the second condition is a minor "niceness" assumption. We elaborate on the third condition in detail in Appendix D in the supplement, and confirm that frequently used proper scoring rules are indeed respectful. We briefly note here that intuitively, the third condition asserts that R'' does not change too quickly (except possibly near zero and one) for small enough coin-flipping costs c. The particular choice of 0.16 is not special, and could be replaced with any constant $< 1/6$.

Definition 3.2 (Incentivization Index). *We define the* incentivization index *of a scoring rule f:*

$$\text{Ind}(f) := \int_0^1 \left(\frac{x(1-x)}{(R^f)''(x)}\right)^{1/4} dx. \quad \textit{More generally, for } \ell \geq 1\textit{:} \ \text{Ind}^\ell(f) := \int_0^1 \left(\frac{x(1-x)}{(R^f)''(x)}\right)^{\ell/4} dx.$$

Theorem 3.3. *If f is a respectful, continuously differentiable proper scoring rule, then:*

$$\lim_{c\to 0} c^{-1/4} \cdot \text{Error}_c(f) = \sqrt{2/\pi} \cdot 2^{1/4} \cdot \text{Ind}(f).$$

More generally, if $\mu_\ell := \frac{2^{\ell/2}\Gamma\left(\frac{\ell+1}{2}\right)}{\sqrt{\pi}}$ is the ℓ^{th} moment of the standard normal distribution, then:

$$\lim_{c\to 0} c^{-\ell/4} \cdot \text{Error}_c^\ell(f) = \mu_\ell \cdot 2^{\ell/4} \cdot \text{Ind}^\ell(f).$$

Intuitively, the incentivization index captures the expert's error as $c \to 0$.[7] More formally, for any two respectful proper scoring rules f, g, $\text{Ind}(f) < \text{Ind}(g)$ implies that there exists a sufficiently small $c_0 > 0$ such that $\text{Error}_c(f) < \text{Error}_c(g)$ for all $c \leq c_0$. As previously referenced, Theorem 3.3 says nothing about how big or small this c_0 might be, although simulations in Appendix H in the supplement confirm that it does not appear to be too small for typical scoring rules.

The rest of this section is organized as follows. Sections 3.1 through 3.6 outline our proof of Theorem 3.3. The key steps are given as precisely-stated technical lemmas with mathematical intuition alongside them, to illustrate where precision is needed for the proof to carry through. Complete proofs of these lemmas can be found in Appendix E in the supplement. In Appendix D, we confirm that natural scoring rules are respectful (which is mostly a matter of validating the third condition in Definition 3.1).

[5] Note this does not necessarily require R''' be defined on the entire $(0,1)$, just that it is defined almost everywhere.
[6] Except in places where R''' is undefined.
[7] Proposition 3.9 in Section 3.4 gives intuition for why $\text{Error}_c(f)$ is proportional to $\sqrt[4]{c}$.

3.1 Proof Outline of Theorem 3.3

We provide below an executive overview of our approach. The concrete steps are separated out as formally-stated technical lemmas in the following sections, with proofs deferred to Appendix E in the supplement. Before beginning, we highlight the main challenge: to prove Theorem 3.3, we need to capture the *precise asymptotics* of the expert's expected error. Upper bounds can be easily shown via concentration inequalities; however, traditional lower bounds via anti-concentration results would simply state that the expected error tends to 0 as $c \to 0$ (which holds for every proper scoring rule, and doesn't distinguish among them). So not only are we looking for two-sided bounds on the error, but we need to gauge the precise *rate* at which it approaches zero. Moreover, even obtaining the order of magnitude of the error as $c \to 0$, which turns out to be $c^{-\ell/4}$, still does not suffice: we need to compute the exact coefficient of $c^{-\ell/4}$. This difficulty motivates the need for the technical lemmas stated in this section to be very precise. Our outline is as follows:

- All of our analysis first considers a locally-adaptive expert, who flips the coin one additional time if and only if the expected increase in reward *from that single flip* exceeds c.
- Our first key step, Section 3.2, provides an asymptotic *lower bound* on the number of times an expert flips the coin, for all respectful f.
- Our second key step, Section 3.3, provides a coupling of the expert's flips across all possible true biases p. This helps prove uniform convergence bounds over all p for the expert's error: we can now define an unlikely "bad" event of overly-slow convergence without reference to p.
- Our third key step, Section 3.4, provides tight bounds on the number of flips by a locally-adaptive expert, up to $(1 \pm o(1))$ factors. Note that the first three steps have not referenced an error measure at all, and only discuss the expert's behavior.
- Our fourth key step, Section 3.5, shows how to translate the bounds in Section 3.4 to tight bounds on the error of a locally-adaptive expert, again up to $(1 \pm o(1))$ factors.
- Finally our last step, Section 3.6, shows that the globally-adaptive expert behaves nearly-identically to the locally-adaptive expert, up to an additional $o(1)$ factor of flips.

We now proceed to formally state the main steps along this outline, recalling that the first several steps consider a locally-adaptive expert, whose definition is restated formally below:

Definition 3.4 (Locally-Adaptive Expert). *The* locally-adaptive expert *flips one more time if and only if making a* single *additional coin flip (and then stopping) increases their expected payoff.*

3.2 Step One: Lower Bounding Expert's Number of Flips

We begin by tying the expert's expected marginal reward from one additional flip to R''. Below, $Q(n)$ denotes the random variable which is the expert's belief after n flips. The important takeaway from Claim 3.5 is that for fixed n, the expert's expected belief as a function of $Q(n)$ changes (roughly) as $Q(n) \cdot (1 - Q(n)) \cdot R''(Q(n))$ — this takeaway will appear in later sections.

Claim 3.5. *Let $\Delta_{n+1}(q) := \mathbb{E}[R(Q(n+1))|Q(n) = q] - R(q)$ be the expected increase in the expert's reward (not counting the paid cost c) from the $(n+1)^{th}$ flip of the coin, given current belief $Q(n) = q$. Then there exist $c_1, c_2 \in [q - 1/n, q + 1/n]$ such that:*

$$\Delta_{n+1} = \frac{q \cdot (1-q)}{2(n+3)^2}(q \cdot R''(c_1) + (1-q) \cdot R''(c_2))$$

Recalling that the locally-adaptive expert decides to flip the coin for the $(n+1)^{th}$ time if and only if $\Delta_{n+1} \geq c$, and assuming that R'' is bounded away from zero (Condition 1 in Definition 3.1), we arrive at a simple lower bound on the number of coin flips.

Claim 3.6. *For all f such that $(R^f)''$ is bounded away from zero, there exists α, c_0 such that the expert is guaranteed to flip the coin at least $\frac{1}{\alpha c^{1/3}}$ times for all $c \le c_0$ (no matter the true bias).*

Using basic concentration inequalities, Claim 3.6 immediately implies an asymptotic *upper bound* on the expert's error. Recall, however, that we need a two-sided bound, and moreover that we need precise asymptotics of the error. Still, Claim 3.6 is the first step towards this.

3.3 Step Two: Ruling Out Irregular Coin-Flipping Trajectories

The expert's coin-flipping behavior depends on $Q(n)$, which depends on the fraction of realized coin flips which are heads, which itself depend on the coin's true bias p. Note, of course, that $Q(n) \to p$ as $n \to \infty$. If instead we had that $Q(n) = p$ *exactly*, we could leverage Claim 3.5 to better understand the number of flips as a function of p. Unfortunately, $Q(n)$ will not equal p exactly, and it is even possible to have $Q(n)$ far from p, albeit with low probability.

The challenge, then, is how to handle these low-probability events, and importantly how to do so *uniformly over p*. To this end, we consider the following coupling of coin-flipping processes over all possible biases. Specifically, rather than first drawing bias p and then flipping coins with bias p, we use the following identically distributed procedure:

(1) Generate an infinite sequence $r_1, r_2, \ldots$ of uniformly random numbers in $[0, 1]$.
(2) Choose p uniformly at random from $[0, 1]$.
(3) For each n, coin n comes up heads if and only if $r_n \le p$.

Under this sampling procedure, $Q_p(n) := \frac{h_p(n)+1}{n+2}$ is the expert's estimate after flipping n coins, where $h_p(n)$ is the number of heads in the first n flips, if p is the value chosen in step (2).

With this procedure, we can now define a single bad event *uniformly over all p*. Intuitively, Ω_N holds when, no matter what p is chosen in step (2), the expert's Bayesian estimate of p never strays too far from p after N flips. More formally, the complement of Ω_N is our single bad event:

$$\overline{\Omega_N} := \bigcup_{n=N}^{\infty} \bigcup_{j=1}^{n-1} \left\{ \left| Q_{j/n}(n) - \frac{j}{n} \right| > \frac{\sqrt{j(n-j)}}{2n^{1.49}} \right\}.$$

The expression on the right-hand side of the inequality can be rewritten as $\sqrt{\frac{\frac{j}{n}\left(1-\frac{j}{n}\right)}{n}} \cdot \frac{n^{.01}}{2}$, where the radical term gives the order of the expected difference between $Q_{j/n}(n)$ and $\frac{j}{n}$. So intuitively, Ω_N holds unless the actual difference between $Q_{j/n}(n)$ and $\frac{j}{n}$ far exceeds its expected value.

We have defined Ω_N so that, on the one hand, our subsequent analysis becomes tractable when Ω_N holds, and on the other hand, Ω_N fails to hold with probability small enough that our asymptotic results are not affected. Below, Claim 3.7 gives the property we desire from Ω_N, and Claim 3.8 shows that $\overline{\Omega_N}$ is unlikely. The key takeaway from Claim 3.7 is that when Ω_N holds, the expert's prediction is close to p *for all $n \ge N$ and $p \in (0, 1)$* and this closeness *shrinks with n*.

Claim 3.7. *The exists a sufficiently large N_0 such that for all $N \ge N_0$: if Ω_N holds, then*

$$\left|Q_p(n) - p\right| \le \frac{\sqrt{p(1-p)}}{n^{.49}} \quad \text{for all } n \ge N \text{ and } p \in [1/n, 1 - 1/n].$$

Claim 3.8.

$$Pr\left[\overline{\Omega_N}\right] = O\left(e^{-N^{.01}}\right).$$

While it is trivial to see that $Q_p(n)$ approaches p as $n \to \infty$, we reiterate that Claims 3.7 and 3.8 guarantee quantitatively that: (a) when Ω_N holds, $|Q_p(n) - p|$ *shrinks with n*, (b) the probability that Ω_N fails *shrinks exponentially fast in N*, and (c) both previous bounds are *uniform over p*.

3.4 Step Three: Tightly Bounding Expert's Number of Flips

We now nail down the precise asymptotics of the number of the expert's flips as a function of the true bias p. This becomes significantly more tractable after assuming Ω_N holds. Below, the random variable n_{stop} denotes the number of flips that a locally-adaptive expert chooses to make.

Proposition 3.9. *Assume that Ω_N holds for some N, and let t be as in Definition 3.1. There exists a constant γ and cost $c_0 > 0$ such that for all $c \le c_0$ and all $p \in [2c^t, 1-2c^t]$, we have*

$$\sqrt{\frac{p(1-p)R''(p)}{2c}(1-\gamma c^{1/300})} \le n_{\text{stop}} \le \sqrt{\frac{p(1-p)R''(p)}{2c}(1+\gamma c^{1/300})}.$$

Proposition 3.9 has two key aspects. First, the upper and lower bounds on n_{stop} match up to a $1 \pm o(1)$ factor. Second, the $o(1)$ term is independent of p. To get intuition for why $n_{\text{stop}} \approx \sqrt{\frac{p(1-p)R''(p)}{2c}}$, recall that Claim 3.5 shows after n flips, the expected marginal gain is $\Delta_{n+1} \approx \frac{p(1-p)}{2n^2}R''(p)$. This quantity first falls below c, the cost per flip, after $n = \sqrt{\frac{p(1-p)R''(p)}{2c}}$ flips.

3.5 Step Four: Translating Number-of-Flips Bounds to Error Bounds

Having pinned down n_{stop} quite precisely, we will now obtain a tight bound on the error of the locally-adaptive expert's reported prediction. By contrast, the previous three steps performed an analysis of the locally-adaptive expert's coin-flipping behavior, which does not depend on the choice of error metric. Lemma 3.10 below is a formal statement of the main step of this process, which nails down the asymptotics of the error conditioned on Ω_N. Below, $\mathsf{Err}_c(p)$ denotes a random variable equal to the locally-adaptive expert's error when the cost is c and the true bias is p (and the scoring rule f is implicit).

Lemma 3.10. *Let $\ell \ge 1$ and $\mu_\ell := \frac{2^{\ell/2}\Gamma\left(\frac{\ell+1}{2}\right)}{\sqrt{\pi}}$ be the ℓ^{th} moment of a standard Gaussian. Let $N = \frac{1}{\alpha c^{1/3}}$ (so N is implicitly a function of c). For all $p \in [2c^t, 1-2c^t]$ we have*

$$(1-o(1))\cdot\mu_\ell\cdot\left(\frac{2p(1-p)}{R''(p)}\right)^{\ell/4} \le c^{-\ell/4}\cdot\mathbb{E}\left[(\mathsf{Err}_c(p))^\ell \mid \Omega_N\right] \le (1+o(1))\cdot\mu_\ell\cdot\left(\frac{2p(1-p)}{R''(p)}\right)^{\ell/4}$$

where the $o(1)$ term is a function of c (but not p) that approaches zero as c approaches zero.

Lemma 3.10 is the key, but far from only, step in translating Proposition 3.9 to tight bounds on the locally-adaptive expert's error. Intuitively, it states that the value of the expert's error will be, up to a $1 \pm o(1)$ factor, consistent with what one would expect from using a quantitative central limit theorem in conjunction with the bound on n_{stop} from Proposition 3.9.

3.6 Step Five: From Locally-Adaptive to Globally-Adaptive Behavior

Finally, we extend our previous analysis from locally-adaptive to globally-adaptive experts. In particular, for a scoring rule that gives finite expected reward to a perfect expert, we prove that the globally-adaptive expert does not flip significantly more than a locally-adaptive expert would, and therefore their achieved errors are equal up to a $1 \pm o(1)$ factor. Below, the random variable n_g denotes the number of flips by the globally-adaptive expert.

Lemma 3.11. *Assume f is respectful and normalizable (i.e. $\int_0^1 R(x)dx < \infty$). Let γ be as in Proposition 3.9. There exists a $c_0 > 0$, such that for all $c \le c_0$: If $\Omega_{n_{\text{stop}}}$ holds and $4c^t \le Q(n_{\text{stop}}) \le 1-4c^t$, then*

$$n_{\text{stop}} \le n_g \le (1+6\gamma c^{1/300})n_{\text{stop}}.$$

Lemma 3.11 is the key step in this portion of the analysis. The remaining work is to bound the impact of negligible events (such as $\Omega_{n_{\text{stop}}}$ failing, or $Q(n_{\text{stop}})$ being extremely close to 0 or 1) on our analysis. This completes our outline of the proof of Theorem 3.3 (and we refer the reader back to Section 3.1 for a reminder of this outline).

4 Finding Optimal Scoring Rules

Now that we have shown that the incentivization index characterizes how well any respectful scoring rule incentivizes a globally-adaptive expert to minimize error, we have a well-defined optimization problem: *which normalized proper scoring rule has the lowest incentivization index* (and therefore minimizes the expert's expected error)? Recall the following necessary and sufficient set of conditions for a continuously differentiable and normalized scoring rule $g(\cdot)$ to be weakly proper:[8]

- (Lemma 2.5) For all $x \in (0,1)$, $xg'(x) = (1-x)g'(1-x)$ and $g'(x) \geq 0$.
- (Definition 2.13, Corollary C.7 in supplemental appendix) $g\left(\frac{1}{2}\right) = 0$.
- (Definition 2.13, Corollary C.7 in supplemental appendix) $\int_{\frac{1}{2}}^{1}(1-x)g'(x)dx = 1$.

So our goal is just to find the scoring rule which satisfies these constraints and minimizes the incentivization index:

$$\mathsf{Ind}^{\ell}(g) = \int_0^1 \left(\frac{x(1-x)}{R''(x)}\right)^{\ell/4} dx = \int_0^1 \left(\frac{x(1-x)^2}{g'(x)}\right)^{\ell/4} dx.$$

The main result of this section is the following theorem:

Theorem 4.1. *The unique continuously differentiable normalized proper scoring rule which minimizes* $\mathsf{Ind}^{\ell}(g)$ *is:*

$$g_{\ell,\text{Opt}}(x) = \begin{cases} \kappa_\ell \int_{\frac{1}{2}}^{x} (t^{\ell-8}(1-t)^{2\ell+4})^{1/(\ell+4)} dt & x \leq \frac{1}{2} \\ \kappa_\ell \int_{\frac{1}{2}}^{x} (t^{\ell}(1-t)^{2\ell-4})^{1/(\ell+4)} dt & x \geq \frac{1}{2}. \end{cases},$$

where κ_ℓ *is the appropriate normalization constant.*

While $g_{\ell,\text{Opt}}$ is certainly challenging to parse, importantly it is a closed form, and can thus be numerically evaluated (and, it is provably optimal). A complete proof of Theorem 4.1 appears in Appendix F in the supplement. Appendix B contains several plots of these scoring rules, alongside traditional ones. Section 5 immediately below also gives further discussion of these rules.

5 Comparing Scoring Rules

In this section we compare various scoring rules by their incentivization indices, for various values of ℓ. Of particular interest are the values $\ell = 1$ (expected absolute error), $\ell = 2$ (expected squared error), and the limit as $\ell \to \infty$ (which penalizes bigger errors "infinitely more" than smaller ones, so this regime corresponds to minimizing the probability of being very far off).

5.1 Optimal Scoring Rules for Particular Values of ℓ

We begin by noting some values of ℓ for which the function $g_{\ell,\text{Opt}}$ takes a nice closed form. $\ell = 1$ happens to not be one such value. For $\ell = 2, 4, 8$, the functions $g_{\ell,\text{Opt}}$ can be written in terms of elementary functions on the entire interval $(0,1)$. For $\ell = 2$, the closed form on $(1/2, 1)$ is a polynomial, although its extension via Corollary 2.6 to $(0, 1/2)$ is not. For $\ell = 8$, the closed form on

[8] Including weakly proper scoring rules in our optimization domain makes the analysis simpler. The optimal scoring rules are in fact strictly proper.

both $(0, 1/2)$ and $(1/2, 1)$ is a polynomial, although they are different. Interestingly, as $\ell \to \infty$, the closed form converges pointwise to a single polynomial. Specifically, for these values of ℓ:

For $\ell = 2$: On $[\frac{1}{2}, 1)$, we have

$$g_{2,\mathrm{Opt}}(x) = \kappa_2 \int_{\frac{1}{2}}^{x} t^{2/3} dt = \frac{3}{5}\kappa_2 \left(x^{5/3} - \left(\frac{1}{2}\right)^{5/3} \right).$$

For $\ell = 8$: On $(0, \frac{1}{2}]$, we have

$$g_{8,\mathrm{Opt}}(x) = \kappa_8 \int_{\frac{1}{2}}^{x} (1-t)^{5/3} dt = \frac{3}{8}\kappa_8 \left(\left(\frac{1}{2}\right)^{8/3} - (1-x)^{8/3} \right)$$

and on $[\frac{1}{2}, 1)$, we have

$$g_{8,\mathrm{Opt}}(x) = \kappa_8 \int_{\frac{1}{2}}^{x} (t^{2/3} - t^{5/3}) dt = \kappa_8 \left(\frac{3}{5}\left(x^{5/3} - \left(\frac{1}{2}\right)^{5/3} \right) - \frac{3}{8}\left(x^{8/3} - \left(\frac{1}{2}\right)^{8/3} \right) \right).$$

Finally, **as $\ell \to \infty$:** on the entire interval $(0, 1)$, $g_{\ell,\mathrm{Opt}}$ pointwise converges to

$$\lim_{\ell \to \infty} \kappa_\ell \cdot \int_{\frac{1}{2}}^{x} t(1-t)^2 dt = \frac{320}{3}\left(\frac{1}{4}x^4 - \frac{2}{3}x^3 + \frac{1}{2}x^2 - \frac{11}{192} \right) = \frac{5}{9}(48x^4 - 128x^3 + 96x^2 - 11).$$

We refer to this last rule as $g_{\infty,\mathrm{Opt}}$. Intuitively, minimizing the expected value of error raised to a power that approaches infinity punishes any error infinitely more than an even slightly smaller error. Put otherwise, this metric judges a scoring rule by the maximum (over $p \in (0, 1)$) of the spread of the distribution of expert error. The scoring rule $g_{\infty,\mathrm{Opt}}$ has a very special property, which is that the quantity $\frac{x(1-x)}{R''(x)} = \frac{x(1-x)^2}{g'_{\infty,\mathrm{Opt}}(x)}$, which appears in the incentivization index, is a constant regardless of x. This means that, in the limit as $c \to \infty$, the distribution of the expert's error is the same regardless of p. It makes intuitive sense that making the spread of the distribution of expert error uniform over all p also minimizes the maximum of these spreads, which explains why $g_{\infty,\mathrm{Opt}}$ has this interesting property.

As some of these rules are not infinitely differentiable, a natural question to ask is: what infinitely differentiable normalized function minimizes Ind^ℓ? While (as we have shown by virtue of $g_{\ell,\mathrm{Opt}}$ being the unique minimizer) achieving an incentivization index equal to $\mathsf{Ind}^\ell(g_{\ell,\mathrm{Opt}})$ with an infinitely differentiable scoring rule is impossible, it turns out that it is possible to get arbitrarily close — and in fact it is possible to get arbitrarily close with *polynomial* scoring rules. The main idea of the proof is to use the Weierstrass approximation theorem to approximate $g_{\ell,\mathrm{Opt}}$ with polynomials. See Section 6 for a full proof.

5.2 Comparison of Incentivization Indices of Scoring Rules

We compare commonly studied scoring rules such as quadratic, logarithmic, and spherical, and refer to their normalizations as $g_{\mathrm{quad}}, g_{\log}, g_{\mathrm{sph}}$, respectively. Additionally we include for comparison the normalization g_{hs} of the *hs* scoring rule, defined as $hs(x) = -\sqrt{\frac{1-x}{x}}$. This scoring rule was prominently used in [1] to prove their minimax theorem for randomized algorithms.

Figure 1 states $\mathsf{Ind}^\ell(g)$ for various scoring rules g (the lower the better). Figure 1 lets us compare the performance of various scoring rules by our metric for any particular value of ℓ. However, as one can see, Ind^ℓ decreases as ℓ increases. This makes sense, since Ind^ℓ measures the expected ℓ-th power of error. For this reason, if we wish to describe how a given scoring rule performs over a range of values of ℓ, we need to normalize these values. We do so by taking the ℓ-th root and

$\mathrm{Ind}^\ell(\cdot)$	$\ell = 1$	$\ell = 2$	$\ell = 4$
$g_{\log}$	0.260	0.0732	0.00644
g_{quad}	0.279	0.0802	0.00694
g_{sph}	0.296	0.0889	0.00819
g_{hs}	0.255	0.0723	0.00658
$g_{1,\mathrm{Opt}}$	0.253	0.0728	0.00719
$g_{2,\mathrm{Opt}}$	0.255	0.0718	0.00661
$g_{4,\mathrm{Opt}}$	0.261	0.0732	0.00639
$g_{\infty,\mathrm{Opt}}$	0.311	0.0968	0.00974

Fig. 1. Rows correspond to scoring rules, and columns correspond to error measures.

Scoring Rule	1	2	4	8	16	32	64	128	256	512
hs	0.990	0.997	0.992	0.979	0.962	0.947	0.935	0.927	0.922	0.919
Logarithmic	0.970	0.990	0.998	0.993	0.982	0.969	0.959	0.951	0.946	0.943
Quadratic	0.905	0.946	0.979	0.996	0.999	0.995	0.989	0.984	0.980	0.978
Spherical	0.853	0.899	0.940	0.968	0.984	0.992	0.995	0.995	0.995	0.994
OPT (l = 1)	1.000	0.993	0.971	0.938	0.905	0.877	0.856	0.842	0.833	0.827
OPT (l = 2)	0.992	1.000	0.991	0.969	0.941	0.915	0.896	0.882	0.873	0.868
OPT (l = 4)	0.966	0.991	1.000	0.992	0.973	0.953	0.936	0.924	0.916	0.910
OPT (l = 8)	0.925	0.964	0.991	1.000	0.994	0.981	0.969	0.958	0.951	0.946
OPT (l = 16)	0.885	0.931	0.971	0.994	1.000	0.996	0.989	0.981	0.976	0.972
OPT (l = 32)	0.854	0.903	0.949	0.980	0.996	1.000	0.998	0.994	0.990	0.987
OPT (l = 64)	0.835	0.885	0.932	0.967	0.988	0.998	1.000	0.999	0.997	0.995
OPT (l = 128)	0.824	0.874	0.921	0.958	0.981	0.994	0.999	1.000	0.999	0.998
OPT (l = 256)	0.818	0.868	0.915	0.952	0.976	0.990	0.997	0.999	1.000	1.000
OPT (l = 512)	0.815	0.864	0.912	0.949	0.973	0.987	0.995	0.998	1.000	1.000
OPT (l -> Infinity)	0.812	0.861	0.908	0.945	0.970	0.984	0.992	0.996	0.998	0.999

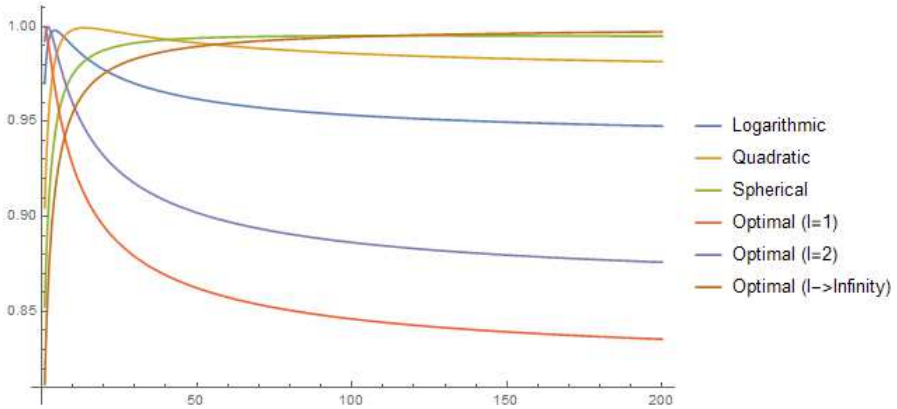
Fig. 2. Rows correspond to different scoring rules g, and columns correspond to different measures of error ℓ. The corresponding entry is $\left(\frac{\mathrm{Ind}^\ell(g_{\ell,\mathrm{Opt}})}{\mathrm{Ind}^\ell(g)}\right)^{1/\ell}$.

dividing these values by the ℓ-th root of the optimal (smallest) index (and take the inverse so that larger numbers are better). This gives us the following measure of scoring rule precision, which makes sense across different values of ℓ:

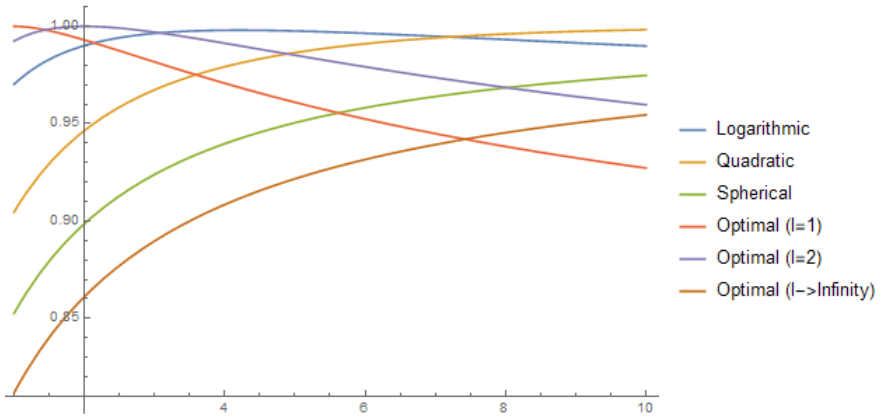
$$\left(\frac{\mathrm{Ind}^\ell(g_{\ell,\mathrm{Opt}})}{\mathrm{Ind}^\ell(g)}\right)^{1/\ell}.$$

Figure 2, which evaluates this expression for a selection of scoring rules and values of ℓ, reveals some interesting patterns. Of the hs, logarithmic, quadratic, and spherical scoring rules, the hs scoring rule is the best one for the smallest values of ℓ and is in fact near-optimal for $\ell = 2$. The logarithmic rule is the best one for somewhat larger values of ℓ and is in near-optimal for $\ell \approx 4$. For larger values of ℓ, the quadratic scoring rule is best, and is near-optimal for $\ell \approx 16$. For even larger values of ℓ, the spherical scoring rule is the best of the four. This pattern suggests that for any given proper scoring rule there is a trade-off between incentivizing precision at low and at high values of ℓ; it would be interesting to explore this further.

Below is a continuous version of Figure 2. The chart shows how the numbers above vary as ℓ ranges from 1 to 200.



And below is a zoomed-in version where ℓ ranges from 1 to 10.



6 Almost-Optimal Incentivization Indices with Polynomial Respectful Scoring Rules

The main result of this section is the following theorem, stating that polynomial,[9] respectful proper scoring rules suffice to get arbitrarily close to the optimal incentivization index.

Theorem 6.1. *For $\ell \geq 1$ and $\varepsilon > 0$, there exists a respectful polynomial normalized proper scoring rule f satisfying $\mathsf{Ind}^{\ell}(f) \leq \mathsf{Ind}^{\ell}(g_{\ell,\mathrm{Opt}}) + \varepsilon$.*

The proof of Theorem 6.1 uses ideas from the Weierstrass approximation theorem. However, the Weierstrass approximation theorem gives a particular measure of "distance" between two functions, which does not translate to these functions having similar incentivization indices. So one challenge of the proof is ensuring convergence of a sequence of polynomials to $g_{\ell,\mathrm{Opt}}$ *in a measure related to* Ind^{ℓ}. A second challenge is to ensure that all polynomials in this sequence are themselves proper, respectful scoring rules. Like previous technical sections, we include a few concrete lemmas to give a sense of our proof outline.

For example, one step in our proof is to characterize all *analytic* proper scoring rules (that is, proper scoring rules that have a Taylor expansion which converges on their entire domain $(0, 1)$).

[9]To be clear, when we say a scoring rule $f(\cdot)$ is polynomial, we mean simply that $f(\cdot)$ is a polynomial function.

A necessary condition to be analytic is to be infinitely differentiable, which rules of the form $g_{\ell,\text{Opt}}$ are not, for any fixed ℓ. We therefore seek to approximate such scoring rules with polynomial scoring rules (which are analytic), which are also respectful and proper.

Theorem 6.2. *Let $f : (0,1) \to \mathbb{R}$ be analytic. Then f is a proper scoring rule if and only if f is nonconstant, $f'(x) \geq 0$ everywhere, and*

$$f(x) = c_0 + \sum_{k>0 \text{ odd}} c_k(2k+1-2kx)\left(x - \frac{1}{2}\right)^k$$

for some $c_0, c_1, c_3, c_5, \cdots \in \mathbb{R}$.

As an example to help parse Theorem 6.2, the quadratic scoring rule has $c_1 < 0$, and $c_i = 0$ for all other i. Using Theorem 6.2, we can conclude the following about $(R^f)''$ for any proper scoring rule f:

Lemma 6.3. *Let $f : (0,1) \to \mathbb{R}$ be analytic. Then f is a proper scoring rule if and only if $(R^f)''$ is not uniformly zero, nonnegative everywhere, and can be written as*

$$(R^f)''(x) = \sum_{k \geq 0 \text{ even}} d_k \left(x - \frac{1}{2}\right)^k.$$

Lemma 6.3 provides clean conditions on what functions $(R^f)''$ are safe to use in our sequence of approximations, and our proof follows by following a Weierstrass approximation-type argument while keeping track of these conditions. The rest of the details for the proof of Theorem 6.1 can be found in Appendix G in the supplement.

7 Conclusion

We propose a simple model where an expert can expend costly effort to refine their prediction, and study the effectiveness of different scoring rules in incentivizing the expert to form a precise belief. Our first main result (Theorem 3.3) identifies the existence of a closed-form incentivization index: scoring rules with a lower index incentivize the expert to be more accurate. Our second main result (Theorem 4.1) identifies the unique optimal scoring rule with respect to this index. Section 5 then uses the incentivization index to compare common scoring rules (including our newly-found optimal ones), and Section 6 shows that one can get arbitrarily close to the optimal incentivization index with polynomial scoring rules.

Our model is mathematically simple to describe, and yet it captures realistic settings surprisingly well (see Section 1 and Appendix A in the supplement). As such, there are many interesting directions for future work. For example:

- Our work considers a globally-adaptive expert, and establishes that they behave nearly identically to a locally-adaptive expert. What about a non-adaptive expert, who must decide a priori how many flips to make before seeing their results?
- Our work considers a principal who wishes to minimize expected error. What if instead the principal wishes to optimize other objectives? In particular, are there objectives that are optimized by simpler rules (such as quadratic, logarithmic, etc.)?
- Our work considers optimal scoring rules for the incentivization index, and shows that polynomial scoring rules approach the optimum. Do *exceptionally* simple scoring rules (such as quadratic, logarithmic, etc.) guarantee a good approximation to the optimal incentivization index for all ℓ?

References

[1] Shalev Ben-David and Eric Blais. 2020. A New Minimax Theorem for Randomized Algorithms. *CoRR* abs/2002.10802 (2020). arXiv:2002.10802 https://arxiv.org/abs/2002.10802

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

[3] Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. 2015. Optimum Statistical Estimation with Strategic Data Sources. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*. 280–296. http://proceedings.mlr.press/v40/Cai15.html

[4] Yiling Chen, Nicole Immorlica, Brendan Lucier, Vasilis Syrgkanis, and Juba Ziani. 2018. Optimal Data Acquisition for Statistical Estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*, Éva Tardos, Edith Elkind, and Rakesh Vohra (Eds.). ACM, 27–44. https://doi.org/10.1145/3219166.3219195

[5] Yiling Chen and Shuran Zheng. 2019. Prior-free Data Acquisition for Accurate Statistical Estimation. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, Anna Karlin, Nicole Immorlica, and Ramesh Johari (Eds.). ACM, 659–677. https://doi.org/10.1145/3328526.3329564

[6] Robert T. Clemen. 2002. Incentive contracts and strictly proper scoring rules. *Test* 11, 1 (01 Jun 2002), 167–189.

[7] Alexander Philip Dawid and Monica Musio. 2014. Theory and applications of proper scoring rules. *METRON* 72, 2 (01 Aug 2014), 169–183.

[8] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

[9] I. J. Good. 1952. Rational Decisions. *Journal of the Royal Statistical Society: Series B (Methodological)* 14, 1 (1952), 107–114.

[10] Jason D. Hartline, Yingkai Li, Liren Shan, and Yifan Wu. 2020. Optimization of Scoring Rules. *CoRR* abs/2007.02905 (2020). https://arxiv.org/abs/2007.02905

[11] Yang Liu and Yiling Chen. 2016. A Bandit Framework for Strategic Regression. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 1813–1821. http://papers.nips.cc/paper/6190-a-bandit-framework-for-strategic-regression

[12] J McCarthy. 1956. Measures of the Value of Information. *Proceedings of the National Academy of Sciences of the United States of America* 42, 9 (9 1956), 654–5.

[13] Kent Osband. 1989. Optimal Forecasting Incentives. *Journal of Political Economy* 97, 5 (1989), 1091–1112.

[14] Tim Roughgarden and Okke Schrijvers. 2017. Online Prediction with Selfish Experts. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 1300–1310. http://papers.nips.cc/paper/6729-online-prediction-with-selfish-experts

[15] Leonard J. Savage. 1971. Elicitation of Personal Probabilities and Expectations. *J. Amer. Statist. Assoc.* 66, 336 (1971), 783–801.

[16] Elias Tsakas. 2019. Robust Scoring Rules. *SSRN* (20 Jul 2019).

[17] K. Weierstrass. 1885. Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen. *Verl. d. Kgl. Akad. d. Wiss. Berlin* 2 (1885), 633–639.

[18] R. L. Winkler, Javier Muñoz, José L. Cervera, José M. Bernardo, Gail Blattenberger, Joseph B. Kadane, Dennis V. Lindley, Allan H. Murphy, Robert M. Oliver, and David Ríos-Insua. 1996. Scoring rules and the evaluation of probabilities. *Test* 5, 1 (01 Jun 1996), 1–60.