How to Evaluate Behavioral Models

Greg d'Eon[1], **Sophie Greenwood**[1,2*], **Kevin Leyton-Brown**[1], **James R. Wright**[3]

[1]University of British Columbia
[2]Cornell University
[3]University of Alberta
gregdeon@cs.ubc.ca, sjgreenwood@cs.cornell.edu, kevinlb@cs.ubc.ca, jrwright@ualberta.ca

Abstract

Researchers building behavioral models, such as behavioral game theorists, use experimental data to evaluate predictive models of human behavior. However, there is little agreement about which loss function should be used in evaluations, with error rate, negative log-likelihood, cross-entropy, Brier score, and squared L2 error all being common choices. We attempt to offer a principled answer to the question of which loss functions should be used for this task, formalizing axioms that we argue loss functions should satisfy. We construct a family of loss functions, which we dub "diagonal bounded Bregman divergences", that satisfy all of these axioms. These rule out many loss functions used in practice, but notably include squared L2 error; we thus recommend its use for evaluating behavioral models.

1 Introduction

Theoretical models of decision-making are often poor descriptions of behavior in practice. As a prime example, classic economic models such as Nash equilibrium fail to describe salient aspects of human behavior: people often choose dominated actions (Goeree and Holt 2001) and fail to account for others' strategic decision making (Kneeland 2015). In response to such failures, fields such as behavioral game theory aim to develop interpretable models that can predict human responses to strategic situations. Such models are helpful to cognitive scientists, for learning how humans think when confronted with economic or strategic choices; to designers of economic systems, for tuning these systems to perform better in practice; and to designers of cooperative AI agents, for enabling these agents to effectively coordinate their behavior with humans (Hu et al. 2020; Carroll et al. 2019).

However, evaluating the quality of such a model on a dataset requires a loss function. Researchers working in behavioral game theory have made a wide variety of different choices about precisely which loss function to use for such evaluations, with error rate, negative log-likelihood, cross-entropy, and (at least two notions of) mean-squared error all being common choices. Clearly, the choice is a substantive one, as different losses will disagree about the quality of a prediction. Which loss function should they use?

In this paper, we attempt to answer this question with a first-principles argument. Though we are motivated by behavioral game theory—and so it is the basis of our examples—our argument depends only on four key characteristics of this field. First, there is some mapping of interest from settings to *distributions* over *finite sets of discrete outcomes* (e.g., the distribution of human decisions in strategic situations). Second, it is possible to collect *multiple samples* from this mapping for any given setting (e.g., by running an experiment with multiple participants). Third, a researcher seeks a *predictive model* of this mapping, which can predict the distribution of unseen data. Fourth, this model must also be *interpretable*, having few parameters whose values can be inspected and understood, and so it cannot generally represent the true mapping perfectly. Our arguments can therefore be extended to other domains that share these characteristics; we give several examples at the end of this paper.

From these characteristics, we argue that loss functions should satisfy five key axioms. The first two, which we call *alignment* axioms, ensure that the loss function induces a correct preference ordering over predictions. These axioms, *sample Pareto-alignment* and *distributional Pareto-alignment*, ensure that the loss function penalizes predictions that are clearly worse (on a given dataset or in expectation over realizations of this data, respectively). The other three, *interpretability* axioms, relate the numerical value of the loss to a prediction's quality. *Empirical distribution sufficiency* requires that the loss be invariant to the number or order of the observations; *counterfactual Pareto-regularity* ensures that the loss appropriately respects changes in the data, and *zero minimum* gives the loss an interpretable optimum.

We show that it is possible to satisfy all of these axioms: we identify an entire family of loss functions that do so, which we dub "diagonal bounded Bregman divergences". Exactly one widely used loss function, the squared L2 error between the predicted and empirical distributions, belongs to this set; we show how each of the other common loss functions violates at least one axiom. In particular, the entire class of scoring rules,[1] a class of loss functions with celebrated alignment

[*]Work done while at the University of British Columbia.

[1]The term "scoring rule" has multiple definitions in the literature. We use a standard definition (e.g., Savage 1971; Gneiting and Raftery 2007) that a scoring rule computes a loss separately for each observation, then takes the mean of these losses (Definition 2.1). Other authors (e.g., Abernethy and Frongillo 2012) use the term to

properties, all fail our interpretability axioms, making them suitable for training models but not evaluating them.

The statistician's view: the likelihood principle. It might seem that the problem of choosing a loss function is a straightforward application of statistical inference: given a dataset and a model class that induces a set of probability distributions, we seek to understand how well each distribution describes the data. Then, the standard statistics textbook argument is that we should use the likelihood of the data to evaluate each of these predicted distributions. This argument is known as the "likelihood principle" (e.g., Berger and Wolpert 1988): if the data was generated by one of the predicted distributions, then likelihood is a sufficient statistic for this distribution. The catch is that this argument relies on the assumption that the model class is "well-specified", containing a model that outputs the true generating distribution. This is not usually the case when evaluating interpretable models, which typically approximate behavior rather than to predict it perfectly. We elaborate further on the problem of evaluating misspecified models when presenting our alignment axioms.

The forecaster's view: scoring rules. Another closely related problem is that of evaluating probabilistic forecasts of future events. Work in this field generally uses *scoring rules* (e.g., Gneiting and Raftery 2007), a class of loss functions that evaluate predictions independently on each observation. Axiomatic characterizations from this literature agree that losses should be *proper*— the expected loss should be minimized by the true distribution, an axiom that we refer to in our analysis as "distributionally proper" —but diverge beyond this point: negative log-likelihood is the only proper scoring rule that satisfies a locality axiom (McCarthy 1956), and two different neutrality axioms characterize Brier score (Selten 1998) and the spherical score (Jose 2009). Our work differs in that we propose axioms that address critical problems that arise when evaluating behavioral models, without being concerned that we are left with an entire class of loss functions.

Some authors have proposed stronger alternatives to propriety. Instead of simply requiring that the correct prediction minimize the expected loss, others have considered lower-bounding the loss of incorrect predictions (Friedman 1983; Nau 1985; Haghtalab, Musco, and Waggoner 2019), maximizing the loss of a naive prediction (Li et al. 2022), or ensuring that it also receives a lower loss in finite samples with high probability (Haghtalab, Musco, and Waggoner 2019). These axioms focus on identifying correct predictions, while we focus on comparing and evaluating incorrect predictions.

The field of property elicitation extends the definition of propriety in a different way, aiming to construct loss functions whose expectations are minimized at other summary statistics of a distribution; propriety is the special case of eliciting the mean. Of particular interest here is work on eliciting multiple properties (Lambert, Pennock, and Shoham 2008; Fissler and Ziegel 2019), as their "accuracy rewarding" and "order sensitivity" axioms are similar to our alignment axioms. We discuss this relationship further in Section 3.

Evaluating model classes. Our axioms are concerned with evaluating individual predictions. Fudenberg et al. (2022) tackle the related problem of evaluating a *model class*, considering the cross-validation performance of a training algorithm that selects a model from this class. They formalize a *completeness* metric, which transforms an existing loss, giving a score of 100% to an algorithm with the best possible cross-validation performance and 0% to a baseline algorithm. Their work complements ours: their completeness measure can be applied to any loss function, but they do not claim how this loss should behave on individual datasets. We thus recommend that researchers evaluating a model class should apply completeness to a loss that satisfies our alignment axioms.

2 Setup and Existing Losses

We now give a formal description of the problem. We start by making a simplification. While researchers generally collect data and evaluate models on many different settings (e.g., games) at once, reporting a model's aggregate performance across these settings, we focus on evaluation in a *single* setting. However, this simplified analysis is useful: any loss that behaves appropriately on an arbitrary number of settings must behave appropriately in the special case of a single setting, so all of the loss functions we disqualify are also unsuitable for multiple settings. We discuss the multiple-setting case in detail in the appendix, where we provide straightforward extensions of our axioms and results.

We model a single scenario as follows. Let $A = \{1, \dots, d\}$ be a fixed set of choices available to the decision maker being modelled (e.g., actions available to experiment participants), and let $\Delta(A)$ be the set of distributions over these choices, i.e., the $(d-1)$-dimensional simplex. We assume that there exists a fixed but unknown true distribution $p \in \Delta(A)$ of behavior, where the randomness in p captures both differences between individuals and randomness in their behavior. An analyst can collect a dataset consisting of n independent, identical draws from p, which we denote $y \sim p^n$, representing actions taken by distinct actors (for example, different participants in a psychology experiment). We denote the set of all such datasets by $\mathcal{D}(A) = \bigcup_{n=1}^{\infty} A^n$.

The analyst is equipped with a model class, which induces a set of predicted distributions $\mathcal{F} \subseteq \Delta(A)$. As this model class is interpretable (e.g., a parametric model with few parameters), this inequality can generally be strict, and $\mathcal{F}$ does not generally include the true distribution p. Their goal is then to choose a model from this class that is good at predicting the distribution of behavior on unseen data.[2] To make their choice, the analyst relies on a loss function $L : \Delta(A) \times \mathcal{D}(A) \to \mathbb{R}$ representing preferences over these predictions: that is, $L(f, y) < L(g, y)$ if and only if f is a better description of the data than g. Note that our analysis can easily be modified to handle objective functions that are expressed in a "positive" sense: for example, it is equivalent to maximize accuracy or minimize error rate.

We pause to define some additional notation. For any dataset $y \in \mathcal{D}(A)$, let $n(y)$ denote the number of observations in y (or simply n, when y is clear from context), and

refer to any arbitrary loss function. Of course, our results on scoring rules only apply to the former, more restrictive definition.

[2]We use the terms "model" and "prediction" interchangeably, as the model is only used to predict behavior in a single scenario.

let $\bar{p}(y) \in \Delta(A)$ be its empirical distribution: that is, for all $a \in A$, $\bar{p}(y)_a = \sum_{i=1}^{n(y)} \mathbb{1}_{\{y_i=a\}} / n(y)$. Lastly, for any action $a \in A$, let $e_a \in \Delta(A)$ denote a point mass distribution on a.

While behavioral game theorists broadly take this approach of evaluating their models with *some* loss function, they largely disagree about precisely *which* loss function to use; in fact, it is not uncommon for a single paper to use multiple different losses while analyzing different experiments. To illustrate this disagreement, we give seven examples of losses that are common in the literature.

First, one common choice is the **error rate** (Fudenberg and Liang 2019; García-Pola, Iriberri, and Kovářík 2020). It is especially common when $\mathcal{F}$ consists only of deterministic predictions, which assign probability to one action.

$$L_{\text{Err}}(f, y) = \sum_{a=1}^{d} \bar{p}(y)_a (1 - f_a),$$

It is similar to **mean absolute error** (MAE) (Camerer, Ho, and Chong 2004; Levin and Zhang 2019).

$$L_{\text{MAE}}(f, y) = \|f - \bar{p}(y)\|_1 = \sum_{a=1}^{d} |f_a - \bar{p}(y)_a|.$$

These two losses are attractive because of their clearly defined scale, with a loss of 0 being achieved by a prediction that never makes mistakes (error rate) or matches the data perfectly (MAE), and a maximum loss of 1 or 2, respectively, by a prediction that is never correct.

Next, several common losses are based on the likelihood of the data, given the prediction. Perhaps the most common choice of loss in all of behavioral game theory is **negative log-likelihood** (NLL) (McKelvey and Palfrey 1992; Stahl and Wilson 1995; Wright and Leyton-Brown 2017).

$$L_{\text{NLL}}(f, y) = -n \sum_{a=1}^{d} \bar{p}(y)_a \log(f_a).$$

Cross-entropy (Kolumbus and Noti 2019) differs from NLL by a factor of n, and **KL divergence** further subtracts the entropy of the dataset.

$$L_{\text{CE}}(f, y) = \tfrac{1}{n} L_{\text{NLL}}(f, a),$$
$$L_{\text{KL}}(f, y) = -\sum_{a=1}^{d} \bar{p}(y)_a \log(\tfrac{f_a}{\bar{p}(y)_a}).$$

All three of these options are rooted in statistics: they make up the core of many statistical hypothesis tests, and all three of them agree with the likelihood principle.

Two more losses originate from regression problems and forecasting. One is the **Brier score**, frequently referred to as mean-squared error or mean-squared deviation (Camerer, Ho, and Chong 2004; Golman, Bhatia, and Kane 2019).

$$L_{\text{Brier}}(f, y) = \tfrac{1}{n} \sum_{i=1}^{n} \|f - e_{y_i}\|_2^2.$$

A small modification is the **squared L2 error**, which is often also called MSE or MSD (Camerer, Ho, and Chong 2003; Selten and Chmura 2008).

$$L_{\text{L2}}(f, y) = \|f - \bar{p}(y)\|_2^2 = \sum_{a=1}^{d} (f_a - \bar{p}(y)_a)^2.$$

Both are natural options for researchers familiar with regression problems, where it is typical to optimize a least-squares objective. They also have roots in forecasting, as the Brier score was originally introduced for evaluating weather forecasts (Brier 1950). We avoid the common but ambiguous term "mean-squared error" to avoid confusion.

Finally, a unifying definition that ties together many losses is the concept of a scoring rule.

Definition 2.1. *(Gneiting and Raftery 2007, page 2.) A* scoring rule *is a function $S : \Delta(A) \times A \to \mathbb{R}$ that maps a prediction $f \in \Delta(A)$ and a single outcome $a \in A$ to a score $S(f, a)$. By averaging these scores over the dataset, every scoring rule S induces a loss function $L_S(f, y) = \frac{1}{n} \sum_{i=1}^{n} S(f, y_i) = \sum_{a \in A} \bar{p}(y)_a S(f, a)$.*

Scoring rules are popular due to their simple functional form, which evaluates the prediction independently on each observation. Their alignment properties are also the subject of several celebrated results (Savage 1971; Gneiting and Raftery 2007), which we describe in detail in Section 5. Error rate, negative log-likelihood, cross-entropy, and Brier score are scoring rules; MAE, KL, and squared L2 are not.

3 Formalizing an Ideal Loss Function

Each loss function from the previous section captures the quality of a prediction on a dataset with a single number, inducing preferences over these predictions. Of course, these loss functions will not always agree with each other about how to order different predictions. Is each loss an equally acceptable choice? To answer this question, we turn to an axiomatic analysis, formalizing axioms that a loss function in a behavioral setting ought to obey. We aim to identify axioms that are as weak as possible, only disqualifying loss functions that exhibit clearly objectionable behavior.

Our axioms can be grouped according to two distinct roles that a loss function serves in describing the quality of a prediction. First, loss functions are used to compare models within a fixed experimental setting. This occurs both during training, when a modeller aims to minimize expected loss on future data; and when evaluating models on a given dataset, comparing losses to see which model achieves the best performance. Our *alignment* axioms address this case, requiring that the loss correctly orders predictions in cases where quality disparities are unambiguous; both are extensions of already standard *propriety* axioms. Second, loss functions are used to understand model performance more broadly; studies report losses and these values are interpreted as conveying information about how well a given model captured human behavior. Our *interpretability* axioms ensure that the loss can indeed be understood in this way, having a well-defined reference point and changing coherently as the data varies.

Alignment axioms. Our first alignment axiom pertains to the training process. While training a predictive model, a modeller's goal is to select a prediction that has low expected test loss over new, unseen data. Thus, if one model better fits the data than another, it should receive a lower expected loss.

What do we mean by "better"? Reasonable people disagree about many comparisons between models, but some are unarguable. For instance, a perfect prediction—one that exactly matches the data generating process—is better than an imperfect one. A standard axiom known as Propriety captures this intuition, requiring that a perfect prediction minimizes the expected loss. To distinguish it from a Sample Propriety axiom that will follow, we refer to it as *Distributional Propriety*.

Axiom (Distributional Propriety (DP))**.** *For all predictions $f \in \Delta(A)$ and all $n \geq 1$, $p \in \Delta(A)$, $f \neq p \implies \mathbb{E}_{y \sim p^n} L(p, y) < \mathbb{E}_{y \sim p^n} L(f, y)$.*

Unfortunately, Distributional Propriety is insufficient for interpretable models: there is often no model in a given class that is able to output an arbitrary distribution. We thus impose a stronger requirement that implies Distributional Propriety: that we should prefer one (potentially imperfect) prediction to another whenever the first is an unambiguously better fit. We formalize this idea with the notion of a Pareto improvement, which we will use extensively in what follows.

Definition 3.1 (Pareto improvement)**.** *Let $p, q, r \in \Delta(A)$ be three distributions. We say that q is a* Pareto improvement *over p with respect to r, denoted by $q \succ_r p$, if for all $a \in A$, either $p_a \leq q_a \leq r_a$ or $p_a \geq q_a \geq r_a$, and furthermore this inequality between p_a and q_a is strict for at least one a.*

In other words, q is a Pareto improvement over p if q is at least as close to r as p in every dimension, and strictly closer to r in some dimension. Then, if one prediction is a Pareto improvement over another with respect to the true distribution—i.e., its predicted probabilities are uniformly closer to the truth—it should receive a lower expected loss.

Axiom (Distributional Pareto-Alignment (DPA))**.** *For all predictions $f, g \in \Delta(A)$, $n \geq 1$, and $p \in \Delta(A)$, $f \succ_p g \implies \mathbb{E}_{y\sim p^n} L(f, y) < \mathbb{E}_{y\sim p^n} L(g, y)$.*

A similar axiom was proposed by Lambert, Pennock, and Shoham (2008) under the name "accuracy-rewarding", and by Fissler and Ziegel (2019) under the name "order sensitive". There is only one difference: in their settings, a prediction is a vector in $\mathbb{R}^d$, containing independent predictions for d different summary statistics of the dataset. Because our predictions lie on the simplex, they are not independent in this way: e.g., predicting that one action has a probability of 1 constrains the predictions for all other actions to be 0.

Next, we consider the situation where two models' predictions are compared to each other on a fixed dataset. This, too, is a fundamental step in behavioral modelling: to evaluate a proposed model, one must compare its predictions to other existing models on *some* dataset to understand whether their proposal better captures human behavior. Here, if one model fits the data better than another, it should receive a lower loss.

As with DP, it is standard to insist that the loss must be minimized when the empirical distribution is reported.

Axiom (Sample Propriety (SP))**.** *For all predictions $f \in \Delta(A)$ and sampled datasets $y \in \mathcal{D}(A)$, $f \neq \bar{p}(y) \implies L(\bar{p}(y), y) < L(f, y)$.*

As above, though, Sample Propriety is insufficient for interpretable models. In this case, it is necessary to prefer predictions that are clearly closer to the empirical distribution, accurately reflecting improvements even away from the optimum. We capture this intuition with a second alignment axiom, which we refer to as Sample Pareto-Alignment.

Axiom (Sample Pareto-Alignment (SPA))**.** *For all predictions $f, g \in \Delta(A)$ and sampled datasets $y \in \mathcal{D}(A)$, $f \succ_{\bar{p}(y)} g \implies L(f, y) < L(g, y)$.*

In the same way as DPA implies DP, SPA implies SP.

Interpretability axioms. Our alignment axioms constrain how the loss may vary as the prediction varies. Our next axioms constrain how the loss may vary as the data varies. Such constraints are important for ensuring that loss represents an understandable measurement of a prediction's quality.

Because it is possible to evaluate a model on multiple observations, one simple way that the data could be changed is simply by observing the same empirical distribution with a different set of observations. This could happen if an experimenter made the same observations in a different order, or collected twice as many observations. Since each observation is independent (e.g., representing an independent trial with a distinct participant), we argue that the loss should be unaffected by such changes to the data.

Axiom (Empirical Distribution Sufficiency (EDS))**.** *For all datasets $y, y' \in \mathcal{D}(A)$ and predictions $f \in \Delta(A)$, $\bar{p}(y) = \bar{p}(y') \implies L(f, y) = L(f, y')$.*

This implies a weaker axiom of *exchangeability*—that permuting the observations does not affect the loss—which is a standard assumption in statistics (e.g., Easton 1989).

What if the dataset varies in a more substantial way? For example, one might replicate an experiment with another group of participants, producing a new set of observations for the same setting, or run slight variations to an experiment to assess their impact on the quality of a model (Goeree and Holt 2001). In both cases, it would be undesirable if the change in the data could cause the prediction to clearly decrease in quality, but be awarded a better loss.

As with varying predictions, there are many ways in which datasets could vary for which reasonable people could disagree about whether the same prediction ought to receive a higher or lower loss. However, we can again leverage the insight that Pareto improvements are unambiguously better: holding a prediction fixed, if the empirical probabilities of the data are brought closer to the predictions for at least some actions and further in none, it is clear that this dataset is better described by the prediction. In such cases, we require that the loss must also improve.

Axiom (Counterfactual Pareto-Regularity (CPR))**.** *Let $f \in \Delta(A)$ be a fixed prediction. Suppose that $y, y' \in \mathcal{D}(A)$ are two datasets of equal size, where $n(y) = n(y')$. Then $\bar{p}(y) \succ_f \bar{p}(y') \implies L(f, y) < L(f, y')$.*

Note that this axiom leverages the discrete outcome space, as it does not obviously generalize to arbitrary distributions.

Up to this point, all of the axioms have only described equalities or inequalities between certain pairs of losses. None have constrained the precise numerical values of the losses: indeed, if L satisfies all of these axioms, then any positive affine transformation $aL + b$ (with $a > 0$) does too. This leaves users with a choice of how to set these two degrees of freedom. We propose to use this freedom to constrain the minimum loss, requiring that a perfect prediction achieves a loss of zero (which must be the loss function's minimum, by SP). This makes the loss easier to interpret: when the analyst has multiple observations in the same setting, it removes the possibility for irreducible error, where even a perfect prediction could get a positive loss.

Axiom (Zero-Minimum (ZM))**.** *For all $y \in \mathcal{D}(A)$, $L(\bar{p}(y), y) = 0$.*

ZM is admittedly the most subjective of our axioms: for

example, on some problems, it might be reasonable to anchor the loss to a different baseline, such as a uniform random prediction. However, its addition is inconsequential when analyzing existing loss functions: in Section 5, we show that each commonly used loss that violates ZM also violates CPR.

4 Diagonal Bounded Bregman Divergences

With these desiderata in mind, the obvious question is: are there loss functions that satisfy all of our axioms? In this section, we provide a positive answer. We first appeal to existing results to show that even asking for a subset of the axioms gives these loss functions considerable structure: Bregman divergences are essentially the only losses that satisfy SP, DP, EDS, and ZM. Narrowing down this class further, we identify a family of losses, which we coin *diagonal bounded Bregman divergences*, that each satisfy our whole set of axioms (SPA, DPA, CPR, EDS, and ZM).

Let us now make these claims more precise. We first define a Bregman divergence. Let $\overline{\mathbb{R}}$ denote the extended real numbers $\mathbb{R} \cup \{\pm\infty\}$, and adopt the convention that $0 \cdot \infty = 0$.

Definition 4.1. *Let $B : C \to \mathbb{R}$ be a closed and proper strictly convex function on a convex set $C \subseteq \mathbb{R}^k$. Then a* subgradient *of B is a function $dB : C \to \overline{\mathbb{R}}^k$ such that*

$$B(x) - B(x_0) \geq dB(x_0)^T(x - x_0)$$

for all $x_0, x \in C$. If B is also differentiable, it has a unique subgradient ∇B on the interior of C.

Definition 4.2. *Given a closed and proper strictly convex function $B : C \to \mathbb{R}$ and subgradient dB of B, the* Bregman divergence *$\nabla_{(B,dB)} : C \times C \to \overline{\mathbb{R}}_{\geq 0}$ of B and dB is*

$$\nabla_{(B,dB)}(p, q) = B(p) - B(q) - dB(q)^T(p - q).$$

We now leverage existing work from the field of property elicitation. Abernethy and Frongillo (2012) show that essentially all loss functions satisfying DP are equivalent to Bregman divergences between a summary statistic of the dataset and the prediction, up to a translation by a function of the data. This immediately yields the following result.

Theorem 4.3 (Corollary of Theorem 11 of Abernethy and Frongillo (2012), informal)**.** *For any n, under mild technical conditions, a loss function L that satisfies DP must be of the form $L(f, y) = \nabla_{(B,dB)}(\rho(y), f) + c(y)$ for some closed and proper strictly convex function B, subgradient dB of B, translation $c : A^n \to \mathbb{R}$, and summary statistic $\rho : A^n \to \Delta(A)$, where $\mathbb{E}_{y \sim p^n}\, \rho(y) = p$ for all p.*

We extend this result, showing that the SP and ZM axioms additionally determine c and ρ, and that the EDS axiom removes the dependence on n. In other words, essentially every loss function satisfying DP, SP, ZM, and EDS is a Bregman divergence between the empirical distribution and the prediction.

Theorem 4.4 (Informal)**.** *Under mild technical conditions, a loss function L satisfies SP and DP if and only if $L(f, y) = \nabla_{(\mathcal{B}(n), d\mathcal{B}(n))}(\bar{p}(y), f) + c(y)$ for some family of closed and proper strictly convex functions $\mathcal{B}$ with subgradients $d\mathcal{B}$ and some translation c. Additionally, L satisfies ZM if and only if $c(y) = 0$ for all y, and L further satisfies EDS if and only if there is some convex function B and subgradient dB such that $\mathcal{B}(n) = B$ and $d\mathcal{B}(n) = dB$ for all n.*

We defer a formal statement and proof of Theorem 4.4 to the appendix, as describing the technical conditions on L takes care. The proof obtains L satisfying DP from Theorem 11 of Abernethy and Frongillo (2012), then applies standard facts about Bregman divergences to show that the additional axioms constrain ρ, c, and $\mathcal{B}$ as described. The reverse direction follows from standard observations from convex analysis.

However, not all Bregman divergences satisfy our remaining axioms SPA, DPA, and CPR. For example, taking $B(f) = \sum_{a=1}^{d} f_a \log f_a$ recovers the KL divergence; we will show in Section 5 that this does not satisfy SPA. Our main result is that all of our axioms are satisfied by the restricted set of *diagonal bounded Bregman divergences*.

Definition 4.5 (Diagonal bounded Bregman divergence (DBBD))**.** *Let $b : [0, 1] \to \mathbb{R}$ be a continuously differentiable convex function where b' is bounded on $[0, 1]$. Let $B_b(x) = \sum_i b(x_i)$ for $x \in [0, 1]^d$. Then, a* diagonal bounded Bregman divergence *is a loss function $L : \Delta(A) \times \mathcal{D}(A) \to \mathbb{R}$, where $L(f, y) = \nabla_{(B_b, \nabla B_b)}(\bar{p}(y), f)$.*

Theorem 4.6. *If L is a DBBD, then L satisfies SPA, DPA, EDS, CPR, and ZM.*

We again defer the proof to the appendix. Briefly, EDS is trivial; ZM follows from Theorem 4.4; SPA, DPA, and CPR leverage the diagonal structure and convexity of B_b.

5 Evaluating Existing Loss Functions

We now revisit the loss functions introduced in Section 2. It is straightforward to see that squared L2 error is a DBBD (with $b(x) = x^2$) and so it satisfies all of the axioms. Each other loss function violates at least one axiom (Table 1). We give an example for each loss below, showing that each axiom violation leads to undesirable results under reasonable conditions. We also demonstrate many of these axiom violations on real behavioral data in the appendix.

Error rate. Error rate violates every axiom except EDS. We show that error rate violates both SP and ZM with the following example.

Example 5.1. *Consider a game in which a player can choose between two actions, "defect" and "cooperate". Suppose that in the true distribution of human play, two-thirds of players defect: $p = (2/3, 1/3)$. In an experiment with 10 distinct participants, an analyst finds that 6 chose to defect, while the remaining 4 chose to cooperate, yielding an empirical distribution of $\bar{p}(y) = (0.6, 0.4)$. Letting $(f, 1 - f)$ be a prediction in this setting, the error rate on this dataset is*

$$L_{Err}(f, y) = 1 - 0.6f - 0.4(1 - f) = 0.6 - 0.2f.$$

This expression is minimized by the prediction $f = 1$, which has an error rate of 0.4. In particular, this prediction achieves a lower error rate than reporting the empirical distribution, which has an error rate of $L_{Err}(\bar{p}(y), y) = 0.48$.

Axiom	Error rate	MAE	NLL	Cross-entropy	KL divergence	Brier score	Squared L2 error
Sample Pareto-Alignment (SPA)	–	✓	–	–	–	✓	✓
Sample Propriety (SP)	–	✓	✓	✓	✓	✓	✓
Distributional Pareto-Alignment (DPA)	–	–	–	–	–	✓	✓
Distributional Propriety (DP)	–	–	✓	✓	✓	✓	✓
Empirical Distribution Sufficiency (EDS)	✓	✓	–	✓	✓	✓	✓
Counterfactual Pareto-Regularity (CPR)	–	✓	–	–	–	–	✓
Zero Minimum (ZM)	–	✓	–	–	✓	–	✓

Table 1: Existing losses and their status under the axioms.

This example illustrates a general problem: for any dataset, the error rate is minimized by predicting the mode, giving more credit to predictions that overestimate the probability of the most likely action.

Mean absolute error. MAE satisfies both SPA and ZM, but does not satisfy DPA or DP. In some cases, a model that predicts the true population distribution gets worse expected MAE on unseen data than an incorrect prediction.

Example 5.2. *Suppose, as in Example 5.1, that the true distribution is $p = (2/3, 1/3)$. However, now suppose that the dataset is not yet available; all that is known is that it consists of 10 independent observations sampled from p. Then, the expected loss of predicting $(f, 1-f)$ is $2\,\mathbb{E}_{y\sim p^{10}}\,|f - \bar{p}(y)_D|$, where $10\bar{p}(y)_D$, the number of participants that defect, is a Binomial random variable with parameters $n = 10, p = 2/3$. This expected loss is minimized by predicting the median of $\bar{p}(y)_D$, which is 0.7. In particular, this prediction receives an expected loss of 0.235, which is lower than the expected loss of 0.243 achieved by predicting the true distribution.*

This example, too, generalizes: in any setting with two actions, the expected loss is minimized by reporting the median of the empirical probability distribution, which is generally not equal to p. In other words, if a model is designed to minimize expected loss, MAE fails to elicit the true distribution.

Negative log-likelihood. NLL is the only loss that violates EDS, which we show in the following example.

Example 5.3. *A second experimenter attempts to reproduce the results from Example 5.1. They first fit a model to the existing dataset y, which has an empirical distribution of $\bar{p}(y) = (0.6, 0.4)$. Their model fits perfectly, returning the exact empirical distribution and getting a negative log-likelihood of $L_{NLL}(\bar{p}(y), y) = 2.9$. They then collect their own dataset y', re-running the experiment with a different set of 20 participants; they find that 12 defect and 8 cooperate, resulting in the same empirical distribution. Although their model still fits the data perfectly, they are surprised to see that it now receives a higher loss of $L_{NLL}(\bar{p}(y'), y') = 5.8$.*

In general, negative log-likelihood scales linearly with the number of observations in the dataset, as it takes a sum over the observations rather than an average.

Cross-entropy and Brier score. We group the next two losses together as they suffer from the same key issue: they violate both CPR and ZM.

Example 5.4. *Undeterred, our experimenter from Example 5.3 considers different loss functions. Using Brier score and cross-entropy to evaluate their perfect model on the original dataset, they obtain losses of*

$$L_{Brier}(\bar{p}(y), y) = 0.48; \quad L_{CE}(\bar{p}(y), y) = 0.29.$$

They collect a third dataset y''; these 10 participants are quite different, with 9 defecting and only one cooperating. They are surprised to find that, despite failing to predict this new dataset perfectly, their model receives lower losses of

$$L_{Brier}(\bar{p}(y), y'') = 0.36; \quad L_{CE}(\bar{p}(y), y'') = 0.24.$$

The first dataset in this example demonstrates violations of ZM: there is no indication that the model has made a perfect prediction, leaving it unclear to the experimenter whether there is room for improvement. In general, the both losses have a non-zero minimum as long as the dataset has two distinct observations. The second dataset shows violations of CPR: it intuitively appears that the model is now better, even though it no longer outputs the correct distribution.

KL divergence. The KL divergence is a translated version of cross-entropy that satisfies ZM, but not SPA, DPA, or CPR. The key issue is that KL divergence gives infinite losses at the boundary. That is, when a model predicts that an action has zero probability of being selected, but the action is observed in the data, that model will have an infinite KL divergence. This leads to situations such as the following.

Example 5.5. *Now, suppose that there are three actions, with a true distribution of $p = (0.001, 0.199, 0.8)$, and that among 100 participants we observe $y = (1, 19, 80)$, yielding an empirical distribution of $\bar{p}(y) = (0.01, 0.19, 0.80)$. Consider comparing two predictions on this dataset: the very coarse prediction of $f = (0, 1, 0)$ and the far more precise $f' = (0, 0.2, 0.8)$. Although f' is a better prediction, as it is closer to $\bar{p}(y)$ than f on both the second and third actions, both receive equal losses of $L_{KL}(f, y) = L_{KL}(f', y) = \infty$.*

In general, when every action appears at least once in the dataset, KL divergence assesses every prediction that places 0 probability on any action as equally bad, and considers all

of these predictions to be worse than any prediction having full support. This is a serious problem, as it is common for every action to be played at least once in sufficiently large behavioral datasets. This makes it difficult to evaluate classical economic predictions, such as Nash equilibrium, which assign 0 probability to many actions. To avoid this issue, some researchers (e.g., Stahl and Wilson 1994) perturb the predictions of such models to yield finite losses, but in doing so introduce an important new parameter and sacrifice the ability to evaluate the original models.

Scoring rules. Recall that error rate, cross-entropy, negative log-likelihood, and Brier score each violated the ZM and CPR axioms. It turns out that these failures are common to all scoring rules, implying that scoring rules should not be used to report model performance.

Proposition 5.6. *Every scoring rule that satisfies SPA violates ZM. Moreover, no scoring rule satisfies CPR.*

We defer the proof to the appendix. Intuitively, since scoring rules must consider each sample independently, they must treat every sample as if it were the entire dataset. Then, in order to satisfy SPA, scoring rules must give positive losses to every nondeterministic prediction, causing them to violate the ZM axiom. Moreover, scoring rules are linear in the empirical probabilities $\bar{p}(y)$ (Definition 2.1). Any such linear function is minimized at one of its boundaries, meaning that it is not uniquely minimized at $\bar{p}(y) = f$ unless $\bar{p}(y)$ is a unit vector; hence, all scoring rules violate CPR.

However, scoring rules do not necessarily violate the alignment axioms. In fact, for every Bregman divergence, there is a scoring rule that gives the same difference in losses between any two predictions on every dataset. For example, this relationship holds between the Brier score and squared L2 error. To state this fact more generally, we recall a classic result characterizing the set of scoring rules satisfying DP.

Theorem 5.7. *(Gneiting and Raftery 2007, Theorem 1.) A scoring rule satisfies DP if and only if there exists a strictly convex function $B : \Delta(A) \to \mathbb{R}$ and subgradient dB such that, for all $f \in \Delta(A)$ and $a \in A$,*

$$S(f, a) = -B(f) - dB(f)^T(e_a - f).$$

Furthermore, every such scoring rule satisfies SP.

Now, suppose that $L(f, y) = \nabla_{(B,dB)}(\bar{p}(y), f)$ is a Bregman divergence, and consider the alternative loss $L'(f, y) = L(f, y) + c(y)$, where $c(y)$ is an arbitrary function that depends only on the data. This additive shift maintains the difference in losses between any two models on every dataset, and it is straightforward to show that it does not affect the status of any of the alignment axioms. In particular, setting $c(y) = -B(\bar{p}(y))$ makes $L'(f, y)$ a scoring rule.

What's more, these scoring rules are computationally easier to minimize than their corresponding DBBDs. Scoring rules can be computed without explicitly calculating $\bar{p}(y)$, making them ideal for large datasets, as the loss can be evaluated without loading the entire dataset into memory at once. Therefore, we do not recommend against the use of scoring rules for model training—it may often be a good idea! We simply argue that researchers should use a corresponding DBBD when evaluating model performance.

6 Conclusions

Our goal in this paper was to identify suitable loss functions for evaluating behavioral models. We took an axiomatic approach, developing axioms describing alignment and interpretability properties that such a loss function should satisfy. We showed that almost all of the loss functions used in the field of behavioral game theory, including the entire class of scoring rules, violate at least one of these axioms. However, it is indeed possible to construct loss functions that satisfy all of our axioms: we identified a large class—the diagonal bounded Bregman divergences—that does. Thus, we advocate that behavioral modelling work use one of these loss functions, with the squared L2 error as a natural incumbent.

Although our motivation comes from behavioral game theory, recall that our arguments rely only on four characteristics of the field: the existence of a mapping from settings to finite, discrete distributions; the ability to obtain multiple observations for any setting; the goal of finding predictive models; and the need for these models to be interpretable. Thus, our work provides guidance not only to behavioral game theorists, but to other researchers whose fields share these characteristics. We are aware of examples in behavioral economics (Plonsky et al. 2019; Agrawal, Peterson, and Griffiths 2020) and further afield in psychology (Busemeyer and Townsend 1993) and operations research (Hensher and Ton 2000; Brenner, Wu, and Amin 2022), and believe that there are yet more potential applications in political science and ecology. We hope that our axiomatic view can help researchers across these disparate areas evaluate and interpret the performance of their models.

Limitations and Future Work. All four of the characteristics played a role in our analysis: finite discrete distributions allowed us to formalize CPR; multiple observations motivated EDS and ZM; predictive models motivated DP and DPA; and interpretable models motivated DPA and SPA. This makes it clear that DP and DPA are not intended for descriptive modelling work, which focuses only on in-sample fit, and that DPA and SPA are unnecessary for evaluating high-capacity uninterpretable models such as deep neural nets, where propriety is sufficient. The impact of our interpretability axioms is also limited, as they are not well motivated for modelling continuous distributions, such as energy consumption or climate variables, or in cases where only one sample can be observed, such as forecasting precipitation types. It would be valuable to extend our results to these fields by developing suitable analogues of our axioms, lifting the need for discrete distributions or finding principled ways to aggregate similar observations.

Is it possible to make a theoretical argument for a *single* best loss function? If so, the path forward is to identify additional desirable axioms for loss functions in behavioral research. For example, on "rock-paper-scissors" experiments, one might insist that loss functions be agnostic to the actions' identities, ensuring that they do not treat "rock" differently from "paper" or "scissors". Making compelling arguments for new axioms and understanding how they narrow down the space of permissible losses—indeed, whether any remain at all—is a valuable direction for future work.

Acknowledgements

Thanks to Frederik Kunstner and Victor Sanches Portella for helpful discussions. This work was funded by an NSERC CGS-D scholarship, an NSERC USRA award, an NSERC Discovery Grant, a DND/NSERC Discovery Grant Supplement, a CIFAR Canada AI Research Chair (Alberta Machine Intelligence Institute), awards from Facebook Research and Amazon Research, and DARPA award FA8750-19-2-0222, CFDA #12.910 (Air Force Research Laboratory).

References

Abernethy, J. D.; and Frongillo, R. M. 2012. A Characterization of Scoring Rules for Linear Properties. *Conference on Learning Theory*.

Agrawal, M.; Peterson, J. C.; and Griffiths, T. L. 2020. Scaling up psychology via Scientific Regret Minimization. *Proceedings of the National Academy of Sciences*, 117(16): 8825–8835.

Berger, J. O.; and Wolpert, R. L. 1988. *The likelihood principle*. Institute of Mathematical Statistics.

Brenner, A.; Wu, M.; and Amin, S. 2022. Interpretable Machine Learning Models for Modal Split Prediction in Transportation Systems. In *IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 901–908.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3.

Busemeyer, J. R.; and Townsend, J. T. 1993. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3): 432.

Camerer, C.; Ho, T.; and Chong, J.-K. 2003. A cognitive hierarchy theory of one-shot games: Some preliminary results. Levine's bibliography, UCLA Department of Economics.

Camerer, C. F.; Ho, T.-H.; and Chong, J.-K. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3): 861–898.

Carroll, M.; Shah, R.; Ho, M. K.; Griffiths, T. L.; Seshia, S. A.; Abbeel, P.; and Dragan, A. 2019. On the Utility of Learning about Humans for Human-AI Coordination. *Advances in neural information processing systems*.

Easton, M. L. 1989. Finite de Finetti style theorems. In *Group invariance in applications in statistics*, volume 1, 108–121. Institute of Mathematical Statistics.

Fissler, T.; and Ziegel, J. F. 2019. Order-sensitivity and equivariance of scoring functions. *Electronic Journal of Statistics*, 13(1): 1166 – 1211.

Friedman, D. 1983. Effective Scoring Rules for Probabilistic Forecasts. *Management Science*, 29(4): 447–454.

Fudenberg, D.; Kleinberg, J.; Liang, A.; and Mullainathan, S. 2022. Measuring the completeness of economic models. *Journal of Political Economy*, 130(4): 956–990.

Fudenberg, D.; and Liang, A. 2019. Predicting and understanding initial play. *American Economic Review*, 109(12): 4112–41.

García-Pola, B.; Iriberri, N.; and Kovářík, J. 2020. Non-equilibrium play in centipede games. *Games and Economic Behavior*, 120: 391–433.

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

Goeree, J. K.; and Holt, C. A. 2001. Ten Little Treasures of Game Theory and Ten Intuitive Contradictions. *American Economic Review*, 91(5): 1402–1422.

Golman, R.; Bhatia, S.; and Kane, P. 2019. The dual accumulator model of strategic deliberation and decision making. *Psychological review*.

Haghtalab, N.; Musco, C.; and Waggoner, B. 2019. Toward a Characterization of Loss Functions for Distribution Learning. In *Advances in Neural Information Processing Systems*, 7237–7246.

Hensher, D. A.; and Ton, T. T. 2000. A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. *Transportation Research Part E: Logistics and Transportation Review*, 36(3): 155–172.

Hu, H.; Lerer, A.; Peysakhovich, A.; and Foerster, J. 2020. "Other-Play"' for Zero-Shot Coordination. In *Proceedings of the 37th International Conference on Machine Learning*.

Jose, V. R. 2009. A characterization for the spherical scoring rule. *Theory and Decision*, 66(3): 263–281.

Kneeland, T. 2015. Identifying higher-order rationality. *Econometrica*, 83(5): 2065–2079.

Kolumbus, Y.; and Noti, G. 2019. Neural networks for predicting human interactions in repeated games. *arXiv preprint arXiv:1911.03233*.

Lambert, N.; Pennock, D.; and Shoham, Y. 2008. Eliciting properties of probability distributions. *Proceedings of the ACM Conference on Electronic Commerce*, 129–138.

Levin, D.; and Zhang, L. 2019. Bridging Level-K to Nash Equilibrium. *Review of Economics and Statistics*, 104: 1329–1340.

Li, Y.; Hartline, J. D.; Shan, L.; and Wu, Y. 2022. Optimization of Scoring Rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, 988–989.

McCarthy, J. 1956. Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42(9): 654.

McKelvey, R. D.; and Palfrey, T. R. 1992. An experimental study of the centipede game. *Econometrica*, 60: 803–836.

Nau, R. F. 1985. Should Scoring Rules be "Effective"? *Management Science*, 31(5): 527–535.

Plonsky, O.; Apel, R.; Ert, E.; Tennenholtz, M.; Bourgin, D.; Peterson, J. C.; Reichman, D.; Griffiths, T. L.; Russell, S. J.; Carter, E. C.; Cavanagh, J. F.; and Erev, I. 2019. Predicting human decisions with behavioral theories and machine learning. *CoRR*, abs/1904.06866.

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336): 783–801.

Selten, R. 1998. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1): 43–61.

Selten, R.; and Chmura, T. 2008. Stationary Concepts for Experimental 2×2-Games. *The American Economic Review*, 98(3): 938–966.

Stahl, D. O.; and Wilson, P. W. 1994. Experimental evidence on players' models of other players. *Journal of economic behavior & organization*, 25(3): 309–327.

Stahl, D. O.; and Wilson, P. W. 1995. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, 10: 218–254.

Wright, J. R.; and Leyton-Brown, K. 2017. Predicting human behavior in unrepeated, simultaneous-move games. *Games and Economic Behavior*, 106: 16–37.