# Measuring Informativeness Gap of (Mis)Calibrated Predictors

**Yiding Feng**[*] **Wei Tang**[†]

## Abstract

In many applications, decision-makers must choose between multiple predictive models that may all be miscalibrated. Which model (i.e., predictor) is more "useful" in downstream decision tasks? To answer this, our first contribution introduces the notion of the *informativeness gap* between any two predictors, defined as the maximum normalized payoff advantage one predictor offers over the other across all decision-making tasks. Our framework strictly generalizes several existing notions: it subsumes U-Calibration (Kleinberg et al., 2023) and Calibration Decision Loss (Hu and Wu, 2024), which compare a miscalibrated predictor to its calibrated counterpart, and it recovers Blackwell informativeness (Blackwell, 1951, 1953) as a special case when both predictors are perfectly calibrated. Our second contribution is a dual characterization of the informativeness gap, which gives rise to a natural informativeness measure that can be viewed as a relaxed variant of the earth mover's distance (EMD) between two prediction distributions. We show that this measure satisfies natural desiderata: it is complete and sound, and it can be estimated sample-efficiently in the prediction-only access setting. Along the way, we also obtain novel combinatorial structural results when applying this measure to perfectly calibrated predictors. [3]

## 1 Introduction

Over the last decade, the machine learning predictors have grown remarkably powerful, especially with the rapid advancements in large-scale models such as large language models (LLMs). These predictors have demonstrated strong performance across a wide range of domains, providing high-quality predictions that are increasingly used by downstream decision-makers to inform their decisions. However, there are usually two key challenges that often hinder decision-makers from fully leveraging these predictions: (1) the underlying mechanisms used to generate predictions are frequently proprietary and opaque to external users; (2) due to limitations in training data or computational constraints in the training process, these predictors may exhibit biases and fail to accurately reflect the empirical frequencies of outcomes. To mitigate these challenges, one natural solution is to ensure that the predictions are calibrated.

A calibrated predictor regulates that predicted probabilities align with the true (conditional) probability of the outcome (Dawid, 1982; Foster and Vohra, 1998). For example, predictions of "80% likelihood" materialize approximately 80% outcome realizations of the time. It is well established that agents who naively best respond to perfectly calibrated predictions incur no regret (Foster and Vohra, 1998; Foster and Hart, 2021). With this desired property, a variety of calibration error metrics have been proposed to quantify how much a predictor deviates from perfect calibration – such as the Expected Calibration Error (ECE) (Foster and Vohra, 1997), the smooth calibration error (Foster and Hart,

[*]Hong Kong University of Science and Technology. Email: ydfeng@ust.hk

[†]Chinese University of Hong Kong. Email: weitang@cuhk.edu.hk

[3]The full version of this paper can be found at https://arxiv.org/pdf/2507.12094

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2018), and the distance to calibration (Blasiok et al., 2023). Remarkably, some "decision-theoretic" measures are proposed to directly quantify the decision-makers' regret when she best-responds to the predictor's forecasts. Kleinberg et al. (2023) introduced "U-Calibration" (UCal), a measure that captures the maximum payoff loss incurred by an agent who naively best responds to a miscalibrated predictor, compared to responding to a best calibrated predictor that provides a fixed prediction. Similarly, Hu and Wu (2024) proposed the "Calibration Decision Loss" (CDL) which quantifies the maximum payoff gap between naively responding to a miscalibrated predictor and responding to the true empirical frequencies of the outcomes associated with each prediction.

Both UCal and CDL therefore assess how far a predictor falls short of its own calibrated counterpart – they tell us "how much better this predictor would be if it were perfectly calibrated." In practice, however, the decision-maker is often handed two (or more) distinct predictors, each trained separately, each potentially miscalibrated in its own way, and must decide which one to rely on. Because neither predictor is necessarily the calibrated version of the other, existing metrics do not directly answer the questions like "Which of these two (possibly miscalibrated) predictors will yield the higher payoff? Is a perfectly calibrated predictor always more useful than a miscalibrated predictor?" The answers to these questions are far from obvious. To illustrate this subtlety, we begin with a simple weather-forecasting example.

**Example 1.1.** *Suppose the long-run probability of rain is 50%. We compare following predictors:*

- *The predictor $\nu$ always forecasts a $50\%$ chance of rain.*
- *The predictor $\mu_1$ forecasts four possible predictions $\{0\%, 49\%, 51\%, 100\%\}$, with a conditional prediction distribution constructed as*

| | $\mathbf{0\%}$ | $\mathbf{49\%}$ | $\mathbf{51\%}$ | $\mathbf{100\%}$ |
|---|---|---|---|---|
| **When it rains** | 0 | 0.0051 | 0.0049 | 0.99 |
| **When it does not rain** | 0.99 | 0.0049 | 0.0051 | 0 |

- *The predictor $\mu_2$ forecasts two possible predictions $\{1\%, 99\%\}$, with a conditional prediction distribution constructed as*

| | $\mathbf{1\%}$ | $\mathbf{99\%}$ |
|---|---|---|
| **When it rains** | 0 | 1 |
| **When it does not rain** | 1 | 0 |

In this example, we can see that although the predictor $\nu$ is perfectly calibrated but it contains no information about the realized outcome beyond the base rate (empirical rainy frequency). In contrast, although predictor $\mu_1$ miscalibrates slightly on the two middle predictions – it says $49\%$ (resp. $51\%$) but the true rain rate is $51\%$ (resp. $49\%$) – but is otherwise nearly perfectly accurate: it predicts $100\%$ almost surely when it rains, and $0\%$ almost surely when it does not. Because of this, any decision-maker who acts on predictor $\mu_1$'s predictions can make choices that more closely reflect the actual outcome, compared to relying solely on the base rate. In fact, and perhaps unsurprisingly, one can show that for every decision problem – no matter the payoff structure – the expected payoff under miscalibrated predictor $\mu_1$ is never lower (and often strictly higher) than under perfectly calibrated predictor $\nu$. This example demonstrates that a miscalibrated but more informative predictor can dominate a perfectly calibrated yet uninformative one.

However, this dominance is not guaranteed in general. If we slightly alter predictor $\mu_1$ to be predictor $\mu_2$ such that it miscalibrates on the two extreme predictions – i.e., predicting $99\%$ when it actually rains, and predicting $1\%$ when it actually does not rain – then its advantage no longer holds.[4] In

[4]In fact, one can show that the ECEs of predictors $\mu_1$ and $\mu_2$ have the same magnitude.

this case, in some decision problems, the predictor $\mu_2$ leads to strictly lower expected payoff than predictor $\nu$, despite still being more "informative" in some sense.

The above example highlights that not all miscalibrations are equal in terms of their impact on decision-making. This raises a natural question: Which predictor is more useful for decision-making problems? Notably, this question is not fully answered even when both predictors are perfectly calibrated. In particular, by viewing a predictor as an information structure, the Blackwell informativeness order offers a partial answer to this question (Blackwell, 1951, 1953). Intuitively, the Blackwell informativeness and its induced Blackwell order captures whether one perfectly calibrated predictor is always more useful than another in every decision problem, thereby inducing a partial order over the space of perfectly calibrated predictors. However, not every pair of perfectly calibrated predictors is comparable under the Blackwell order, let alone pairs of possibly miscalibrated predictors. This work aims to study the following questions:

> *Can we compare any two (possibly miscalibrated) predictors based on how "useful" they are to the decision-making problems?*
> *If they are not equally useful, can we quantify their gap, and is there a natural measure that satisfies common sense desiderata for characterizing this difference?*

### 1.1 Main Results

In this paper, we provide principled answers to both of the motivating questions. In line with most prior work in the calibration literature, we focus on predictors for binary outcomes.

**Informativeness gap.** As our first conceptual contribution, we introduce and study a new notion called the *informativeness gap*, denoted by $\textsc{InfoGap}[\cdot,\cdot]$. Given any two (possibly miscalibrated) predictors $\mu$ and $\nu$, the informativeness gap $\textsc{InfoGap}[\mu,\nu]$ quantifies the maximum payoff advantage that predictor $\mu$ offers over predictor $\nu$ across all *normalized* decision-making tasks. Here, the normalized decision-making tasks refer to those in which, for any action, the decision-maker's payoff difference over different outcomes are bounded by one (see the formal definition in Definition 2.2).[5]

$$\textsc{InfoGap}[\mu,\nu] \triangleq \sup_{\substack{\text{payoff-normalized}\\ \text{decision problem}}} \textsc{Payoff}[\mu] - \textsc{Payoff}[\nu] \ ,$$

where $\textsc{Payoff}[\mu]$ denotes the expected utility in a particular decision problem obtained by the decision-maker who naively best-responds to the predictions produced by predictor. This definition provides an operational analogue to Blackwell informativeness, which interprets the relative usefulness of two predictors through the decision-maker's maximum payoff difference on all possible decision problems.

**Characterizing $\textsc{InfoGap}[\cdot,\cdot]$ between perfectly calibrated predictors.** We begin by analyzing the informativeness gap $\textsc{InfoGap}[\mu,\nu]$ in the special but theoretically fundamental case where both predictors $\mu$ and $\nu$ are perfectly calibrated. Our main result provides a characterization of this gap, and interestingly, reveals that it closely resembles a relaxed variant of the well-known earth mover's distance (EMD), also known as Wasserstein distance, between probability distributions. Motivated by this connection, we introduce the quantity $\mathrm{REMD}[\cdot,\cdot]$, defined over probability distributions. We then illustrate how to interpret it as an informativeness measure that quantifies how much more useful predictor $\mu$ is compared to predictor $\nu$ in decision-making tasks.

**Definition 3.2 and Theorem 3.1 (Informal).** *For any two distributions $f_1, f_2$, define the* relaxed earth mover's distance $\mathrm{REMD}[f_1,f_2]$ *as* $\mathrm{REMD}[f_1,f_2] \triangleq \inf_{\pi\in\Pi(f_1,f_2)} \int_0^1 \left|\int_0^1 \pi(p,q)\cdot(p-q)\,\mathrm{d}q\right| \mathrm{d}p$ *and $\Pi(f_1,f_2)$ denotes the set of all couplings between two distributions. For any two perfectly calibrated predictors $\mu,\nu$, their informativeness gap $\textsc{InfoGap}[\mu,\nu]$ satisfies $\textsc{InfoGap}[\mu,\nu] = \mathrm{REMD}[f_\mu,f_\nu]$ where $f_\mu, f_\nu \in \Delta([0,1])$ denote the distributions over predictions generated by the two predictors $\mu,\nu$, respectively.*

[5] Alternatively, one could consider a *multiplicative* informativeness gap, defined as the maximum ratio between the payoffs under two predictors across all decision problems. However, we focus on and study the *additive* version, as it aligns more naturally with the decision-theoretic calibration literature, which primarily emphasizes additive regret.

While the informativeness gap $\textsc{InfoGap}[\cdot,\cdot]$ is defined over predictors, our relaxed earth mover's distance $\mathrm{REMD}[\cdot,\cdot]$ is defined over distributions. As we discuss below, $\mathrm{REMD}[\cdot,\cdot]$ admits interesting structural characterizations and may be of independent interest—even beyond the context of calibration or forecasting.[6]

**Characterizing $\textsc{InfoGap}[\cdot,\cdot]$ between miscalibrated predictors.** We now turn to the setting where both predictors may be miscalibrated. To capture the informativeness gap in the presence of miscalibration, it is essential to incorporate the true frequencies of outcomes conditional on predictions. Specifically, for any prediction $p \sim \mu$, we let $\kappa_\mu(p) \in [0,1]$ denote the true outcome frequency given the prediction. Our main result introduces a strict extension of the measure $\mathrm{REMD}[\cdot,\cdot]$. This generalized measure, denoted by $\mathrm{REMD}^{\textsc{MisC}}[f_1, f_2, \kappa_1, \kappa_2]$, takes as input two distributions and two corresponding outcome functions.

**Definition 4.1 and Theorem 4.1 (Informal).** *For any two distributions $f_1, f_2 \in \Delta([0,1])$ and two functions $\kappa_1, \kappa_2 : [0,1] \to [0,1]$, define* informativeness measure $\mathrm{REMD}^{\textsc{MisC}}[f_1, f_2, \kappa_1, \kappa_2]$ *as*

$$\mathrm{REMD}^{\textsc{MisC}}[f_1, f_2, \kappa, \kappa_2] \triangleq \inf_{\pi \in \overline{\Pi}(f_1, f_2)} \int_0^1 \left| \int_0^1 \pi(p,q)(p-q)\,\mathrm{d}q + (\kappa_1(p) - p) f_1(p) - (\kappa_2(p) - p) f_2(p) \right| \mathrm{d}p\,,$$

*where the set $\overline{\Pi}(\mu,\nu)$, referred to as the* flow set *is a strict superset of the coupling set $\Pi(f_1, f_2)$, which imposes "flow conservation" constraint over $\pi \in \Delta([0,1] \times [0,1])$.*

*For any two (possibly miscalibrated) predictors $\mu, \nu$, their informativeness gap $\textsc{InfoGap}[\mu,\nu]$ satisfies*

$$\textsc{InfoGap}[\mu,\nu] = \mathrm{REMD}^{\textsc{MisC}}[f_\mu, f_\nu, \kappa_\mu, \kappa_\nu]$$

*where $f_\mu, f_\nu$ (resp. $\kappa_\mu, \kappa_\nu$) are the prediction distributions (resp. true outcome frequencies) of predictors $\mu, \nu$, respectively.*

To understand how $\mathrm{REMD}^{\textsc{MisC}}[f_\mu, f_\nu, \kappa_\mu, \kappa_\nu]$ generalizes $\mathrm{REMD}[f_\mu, f_\nu]$, notice that when both predictors $\mu, \nu$ are perfectly calibrated, we have $\kappa_\mu(p) = \kappa_\nu(p) = p$ for all $p \in [0,1]$. In this case, the objective in $\mathrm{REMD}^{\textsc{MisC}}[f_\mu, f_\nu, \kappa_\mu, \kappa_\nu]$ reduces exactly to that of $\mathrm{REMD}[f_\mu, f_\nu]$. Moreover, the flow set $\overline{\Pi}(f_\mu, f_\nu)$ (see Section 4 for the formal definition) generalizes the standard coupling set $\Pi(f_\mu, f_\nu)$, and thus it represents a relaxed constraint.

Interestingly, by Theorem 3.1 and Theorem 4.1, we can see that when both predictors are perfectly calibrated, the two measures equal to each other: $\mathrm{REMD}^{\textsc{MisC}}[f_\mu, f_\nu, \kappa_\mu, \kappa_\nu] = \mathrm{REMD}[f_\mu, f_\nu]$. In other words, optimizing over the relaxed flow set is also sufficient to characterize the informativeness gap in the perfectly calibrated setting.

**Desiderata of our informativeness gap and informativeness measure.** Our proposed informativeness measure $\mathrm{REMD}^{\textsc{MisC}}[\cdot,\cdot,\cdot,\cdot]$ can be served as a tool for quantifying the informativeness gap between (possibly miscalibrated) predictors. By Theorem 4.1, $\mathrm{REMD}^{\textsc{MisC}}[f_\mu, f_\nu, \kappa_\mu, \kappa_\nu]$ is both complete and sound (in fact, it satisfies these criteria exactly), and thus it is consistent. By our definition, our proposed informativeness gap $\textsc{InfoGap}[\mu,\nu]$ and its equivalence representation, i.e., informativeness measure $\mathrm{REMD}^{\textsc{MisC}}[f_\mu, f_\nu, \kappa_\mu, \kappa_\nu]$, are also prediction-only accessible: notice that for any predictor $\mu$, its true conditional frequency $\kappa_\mu(p)$ can be computed using only the sample pair (prediction, realized outcome). Lastly, we also present sample complexity bounds for estimating the measure $\mathrm{REMD}^{\textsc{MisC}}[f_\mu, f_\nu, \kappa_\mu, \kappa_\nu]$ and thus the informativeness gap $\textsc{InfoGap}[\mu,\nu]$. The algorithm for this estimation utilizes our structural characterizations for $\mathrm{REMD}^{\textsc{MisC}}[f_\mu, f_\nu, \kappa_\mu, \kappa_\nu]$, and has a sample complexity of $\mathrm{poly}(1/\varepsilon)$ (see Theorem 5.1). Finally, another desiderata that is usually considered in the machine learning is the *continuity* of a measure: the measure should be continuous w.r.t. the prediction distribution. This raises a natural question: does there exist an informativeness measure that satisfy the above three desiderata and this additional continuity property simultaneously? Perhaps not surprisingly, the answer is no. We present a general impossibility result showing that no informativeness measure can satisfy consistency (i.e., completeness and soundness) and continuity simultaneously.

[6] Throughout the paper, we use $f_\mu$ and $f_\nu$ to denote the prediction distributions associated with predictors $\mu$ and $\nu$, respectively, and $f_1$ and $f_2$ to denote general distributions.

**Proposition 4.9 (Informal).** *For any informativeness measure over predictors, at least one of the following must fail: it is complete, it is sound, and it is continuous.*

## References


David Blackwell. 1951. Comparison of Experiments. *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability* (1951), 930–102.

David Blackwell. 1953. Equivalent comparisons of experiments. *The annals of mathematical statistics* (1953), 265–272.

Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. 2023. When does optimizing a proper loss yield calibration? *Advances in Neural Information Processing Systems* 36 (2023), 72071–72095.

A Philip Dawid. 1982. The well-calibrated Bayesian. *Journal of the American statistical Association* 77, 379 (1982), 605–610.

Dean P Foster and Sergiu Hart. 2018. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games and Economic Behavior* 109 (2018), 271–293.

Dean P Foster and Sergiu Hart. 2021. Forecast hedging and calibration. *Journal of Political Economy* 129, 12 (2021), 3447–3490.

Dean P Foster and Rakesh V Vohra. 1997. Calibrated learning and correlated equilibrium. *Games and Economic Behavior* 21, 1-2 (1997), 40–55.

Dean P Foster and Rakesh V Vohra. 1998. Asymptotic calibration. *Biometrika* 85, 2 (1998), 379–390.

Lunjia Hu and Yifan Wu. 2024. Predict to minimize swap regret for all payoff-bounded tasks. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 244–263.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. 2023. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 5143–5145.