Learning What Evaluators Value: A Reliable Approach to Modeling Evaluator Preferences

Madeline Celi Kitch
Carnegie Mellon University
Pittsburgh, PA 15213
mckitch@cs.cmu.edu

Nihar B. Shah
Carnegie Mellon University
Pittsburgh, PA 15213
nihars@cs.cmu.edu

Abstract

In many machine learning applications, human and LLM evaluators use assessments of relevant criteria to create an overall evaluation for an item or individual. In applications like admissions, committees assess candidates on attributes such as test scores, GPA, and research experience to evaluate their overall fit for the program. In medical care, clinicians use patient reports of symptoms to consider preliminary diagnoses and assess risks. Each case involves mapping measurable criteria to an overall evaluation—a process that reflects the evaluator's underlying preferences. We focus on the fundamental issue of learning these preferences. Many applications of this problem make specific modeling assumptions on evaluator preferences that may be substantially violated in the real world. We make the minimal assumption that the preference function is coordinate-wise non-decreasing, which is reasonable in a large number of evaluation settings. We theoretically characterize the severity of model mismatch for many common assumptions and show that it can lead to catastrophic effects for learning evaluator preferences and other important downstream tasks. We then present an algorithm for learning evaluators' preferences that is robust to model mismatch. We prove theoretically that our algorithm can learn any monotonic preference function without sacrificing performance when the linearity assumption holds. Evaluations of our algorithm with synthetic simulations and real-world data confirm its ability to learn preferences robustly and illustrate key aspects of LLM and human preferences.

1 Introduction

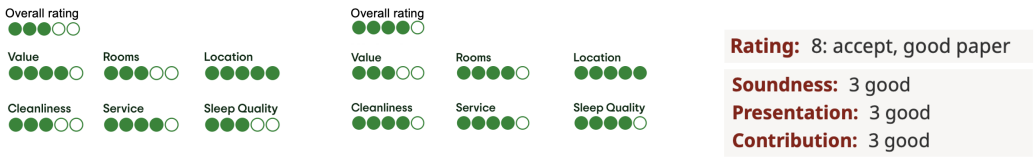
In many areas of life, we use evaluations to help guide decision-making. Doctors evaluate patients in order to recommend the best course of treatment; (LLM) reviewers evaluate academic papers to determine areas of improvement and conference acceptance; everyday individuals evaluate products and hotels that help inform others on what to buy and where to stay.

Often, to make their evaluations clearer and methodical, evaluators provide assessments of relevant criteria or aspects alongside their overall rating or recommendation. Doctors record patients' clinical signs and symptoms alongside their assessments and treatment recommendations; reviewers may be asked for ratings of paper aspects in addition to an overall score (see Figure 1).

Such evaluations serve many purposes: they illustrate what factors went into the evaluator's decision-making and ensure that they consider multiple dimensions of the item or individual prior to reaching their conclusion. Consequently, it is harder for evaluators to simply "go with their gut" and miss a potentially relevant factor.

In this paper, we are interested in the fundamental question: *How do evaluators map criteria evaluations to their overall scores or recommendations?*

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Example Tripadvisor reviews (b) Example ICLR 2024 review

Figure 1: Evaluators provide criteria evaluations alongside an overall score or recommendation. Our focus is on learning how evaluators map assessments of criteria (e.g., cleanliness or rooms or soundness) to their overall scores (e.g., a stay rated 4 out of 5 stars, or a review says 8: accept).

We will call this mapping the evaluators' *preference function*. Learning these preferences is useful in several ways: it can help us understand what evaluators care about, reduce disparities from idiosyncratic evaluation preferences, and align rule-based assessments with people's beliefs and/or expertise (we explain each use more fully in Appendix A).

For these and other reasons, many researchers and practitioners have tried to learn evaluator preferences. However, they typically assume the preference function is linear, or some extension thereof (e.g., log-linear or linear with interactions). Though these simplifying assumptions have benefits—such as increased interpretability and lower sample complexity—they also imply certain assumptions on people's preferences that may or may not make sense. What's more, people often fail to check the necessary model assumptions in practice. Jones et al. (2025) surveyed papers where researchers assumed a linear model and needed to check the linearity assumption; they found that less than 20% did so directly. Furthermore, others have begun to critique the use of simple parametric functions on philosophical grounds, arguing that such models often cannot capture the essential components of people's preferences Keswani et al. (2025).

Our approach. Given the potential issues arising due to model mismatch, we do not assume a parametric model. Instead, we make the minimal assumption that preferences are isotonic or non-decreasing in each criterion. This model is well-suited for evaluation problems: while one may not know how much an increase in a criteria score affects the overall assessment, it is often reasonable to assume the direction. For instance, all else constant, a paper with a greater degree of novelty should not be rated lower, and a psychiatric patient presenting with more severe symptoms should not be deemed lower risk.

We ask two main questions. First, to what extent model mismatch is problematic for understanding evaluator preferences. Second, we ask if an alternative algorithm could learn preferences robustly and reliably. To conclude, we empirically analyze our algorithm on synthetic and real-world preference data.

Our contributions: Here are the key contributions of this paper.

Problems with common model assumptions. We establish strong negative results for commonly used models when true preferences are more complex. For any linear model, generalized additive model, or generalized additive model with interactions, there exist isotonic preference functions for which the estimation error is nearly as large as randomly guessing outputs.

Algorithmic solution and theoretical guarantees. We develop a regularized isotonic regression algorithm with cross-validation that interpolates between isotonic and linear regression. We prove it learns any isotonic preference function and is never worse than linear regression (up to constant factors). Our theoretical techniques extend naturally to other parametric assumptions.

Empirical results. On synthetic preference data, our algorithm matches linear regression for linear preferences and substantially outperforms it for nonlinear ones (see Appendix D.1). Given preference data on hotel ratings (Antognini and Faltings, 2020), our algorithm always achieves lower prediction and reducible error (a proxy for estimation error) than linear regression, and results in reductions of up to 69% in reducible error. Finally, we provide a case study demonstrating how our method can be used in practice. We compare preferences of human and LLM evaluators using a dataset of human and LLM (GPT-4o and Llama 3.1 70b) generated scientific reviews from Yu et al. (2026). We

test both preference alignment (i.e., how close the estimated mapping from criteria to overall scores of an LLM is to that of human evaluators) and how consistent the preferences of a given LLM are across reviews. On both dimensions, GPT significantly outperforms Llama, though given the limited datasets we test on, this finding is meant to highlight the utility of our method rather than the relative benefit of using either LLM for evaluations.

2 Related work

As we are interested in learning isotonic (preference) functions, our work relates to the literature on isotonic regression. When using a least squares estimator, a series of papers prove theoretical upper bounds on isotonic regression in the case of one variable or more variables (Chatterjee et al., 2014, 2015; Fang et al., 2020; Han et al., 2019). We complement this by proving bias results, demonstrating that a broad class of common model assumptions can lead to worst-case minimax error. Another key thread in the literature is the natural adaptivity of the least squares estimator when the function is piecewise-constant Chatterjee et al. (2015); Fang et al. (2020); Han et al. (2019) or lower-dimensional Chatterjee et al. (2015). Our work illustrates an alternative way of ensuring adaptivity via cross-validation and regularization. By applying a key oracle inequality from Vaart et al. (2006), we show that our algorithm recovers parametric rates for linear models.

As our data often comes from people, another set of relevant work is on preference aggregation and learning. A key setting in this literature is learning from pairwise comparisons (Shah and Wainwright, 2018; Heckel et al., 2016; Wang et al., 2019; Bradley and Terry, 1952; Shah et al., 2016b,a). As in our setting, researchers often assume specific parametric models (i.e., Bradley-Terry-Luce or Thurstone). For the problem of estimating comparison probabilities, Shah et al. (2016a) establishes strong bias results if these assumptions do not hold, and introduces a natural non-parametric assumption that can still recover minimax optimal parametric rates. Our work is similar in spirit, but focuses on learning a *function* given cardinal feedback rather than simply overall scores, rankings, or comparison probabilities given cardinal feedback. Additionally, in the field of participatory design, there has been increasing interest in understanding human preferences (as functions) in order to incorporate their beliefs in algorithmic design (Lee et al., 2019; Keswani et al., 2025; Freedman et al., 2020; Noothigattu et al., 2018). Most of this work assumes a linear model (Keswani et al., 2025; Lee et al., 2019) or another simple parametric form such as weighted p-means (Pardeshi et al., 2024), despite the significant issues described in Keswani et al. (2025). When applied to this literature, we provide a method for learning what individuals care about without missing morally relevant components of their beliefs. Noothigattu et al. (2021) is closest to our work, focusing specifically on learning evaluator preferences, but from a social choice theoretic perspective. Given a set of parametric loss functions for empirical risk minimization, they show that only one choice of parameters satisfies three natural axioms. Moreover, the algorithm they present has been used broadly in several ML/AI conferences as well as admissions processes, underscoring the applicability and importance of our setting. In contrast to Noothigattu et al. (2021), we take a statistical approach: our goal is to find a function that best fits the data. This perspective is beneficial in situations where reliably estimating the value or outcome (i.e., clinicians assessing patient risks, companies predicting product ratings) is of primary importance. Our method provides practitioners with statistical guarantees and avoids overfitting on smaller datasets.

Lastly, our work relates to applied literature on commensuration bias and participatory design. Most relevant is philosopher Carole Lee's work on commensuration bias. In the context of the scientific review process, Lee argues that commensuration, the process of "converting distinct qualities into a single metric," can be significantly inconsistent from one evaluation to the next (Lee, 2015). This is largely due to reviewer idiosyncrasy; different evaluators tend to hold different beliefs about the commensuration process (for instance, evaluator 1 may believe two criteria interact, while evaluator 2 does not). Consequently, the assessment and resulting outcome or recommendation for an item or individual becomes contingent on the particular evaluator they are assigned, violating the philosophical imperative of equal treatment. This problem, which she terms *commensuration bias*, is evident in a variety of scientific review processes. Our approach can help to mitigate commensuration bias: Practitioners can use the community preference function to flag assessments where the criteria-to-overall score mapping deviates from the norm. Taking a further look at such assessments can help ensure that the outcome or recommendations ultimately provided reflect what the community of evaluators cares about. Finally, within computer science, there has been interest in

understanding (mathematically) human preferences so as to incorporate their beliefs in algorithmic design (Lee et al., 2019; Keswani et al., 2025; Freedman et al., 2020). At a high level, the aim is to learn individual preference functions, and then aggregate them via an axiomatic approach (Lee et al., 2019; Noothigattu et al., 2018); most work assumes a linear model (Keswani et al., 2025; Lee et al., 2019) or another simple parametric form such as weighted p-means (Pardeshi et al., 2024). This is despite the apparent issues with doing so, as described in Keswani et al. (2025). When applied to this literature, we provide a method for learning what individuals care about without missing morally relevant components of their beliefs.

3 Model overview

In this section, we will provide a formal overview of the evaluation setting, our learning problem, and the common model assumptions we consider. Recall that evaluators take as input assessments of individual criteria or aspects and output an overall score or assessment (see Figure 1). In our ICLR example, the criteria correspond to aspects of the paper (soundness, presentation, and contribution); the overall score corresponds to the rating (e.g., 8 out of 10). Each criterion is given scores corresponding to integer values from 1 (poor) to 4 (excellent). We use $d \geq 2$ to denote the number of relevant criteria, and $m \geq 2$ to denote the number of values a score can have (e.g., for ICLR, there are 3 criteria and 4 possible scores per criterion). Let $[m] = \{1, 2, \ldots, m\}$ be the set of integers between 1 and m. Then the vector of criteria scores, x, is contained in $\mathcal{X} = [m]^d$. Further, because overall scores commonly take on values in a bounded range (e.g., papers are scored from 1-10), we assume that the preference function maps to some bounded interval. For technical convenience, we rescale the outputs to $[0, 1]$. Therefore, the evaluator preference function f takes as input a vector of criteria scores $x \in \mathcal{X}$ and outputs a scalar value $f(x) \in [0, 1]$.

Recall that we assume only that the preference function is isotonic, or non-decreasing in each criterion. Formally, we define the set of isotonic preferences with range $\mathcal{Y}$, to be

$$\mathcal{F}_{\texttt{ISO}}(\mathcal{Y}) := \{f : \mathcal{X} \to \mathcal{Y} \mid f(x) \geq f(x') \text{ whenever } x_i \geq x'_i \; \forall i \in [d]\}, \tag{1}$$

and we let $\mathcal{F}_{\texttt{ISO}} := \mathcal{F}_{\texttt{ISO}}([0, 1])$. As discussed in Section 1, in many evaluation settings, any realistic preference function is coordinate-wise non-decreasing; hence, it is reasonable to assume that $\mathcal{F}_{\texttt{ISO}}$ contains the set of *all* possible evaluator preferences.

3.1 Learning preference functions

Our primary goal is to learn evaluator preferences. To accomplish this, we use evaluation data consisting of criteria and overall scores, which serve as the function inputs and outputs, respectively. Let there be $N \geq 2$ total assessments, indexed by j, and let $y^{(j)}$ be the overall score or recommendation for item j. The set of all N evaluations is represented by a sequence of tuples $\{x^{(j)}, y^{(j)}\}_{j \in [N]}$ satisfying

$$y^{(j)} = f(x^{(j)}) + w^{(j)} \tag{2}$$

where $w^{(j)}$ is independent noise, capturing the fact that the preference applied to each sample may be different due to context or differences in evaluator opinions. The $\{x^{(j)}\}_{j=1}^N$ are independent and identically distributed from some distribution P with support on $\mathcal{X}$ and inducing a set of probabilities $p(x) \in [0, 1]$. Our goal is to learn f as well as possible over the set of typical criteria scores (captured by the distribution P). We measure this by how well our estimated function, $\widehat{f}$, predicts the output of f, when given a new set of criteria scores $x \sim P$. Let $\|g\|^2_{L_2(P)} = \mathbb{E}_{x \sim P} \|g(x)\|_2^2$ be the expected L_2 norm of a function with inputs randomly sampled from P. Then we can define our measure, estimation error, as

$$R(\widehat{f}, f) = \mathbb{E}\|f(x) - \widehat{f}\|^2_{L_2(P)} \tag{3}$$

where the expectation is taken over the randomness of the sample data. Since f takes values in $[0, 1]$, any estimator $\widehat{f}$ constrained to the same range satisfies $R(\widehat{f}, f) \leq 1$.

3.2 Common model assumptions

We consider a set of common model assumptions, the simplest of which is the linear model. The set of non-decreasing linear models with range $\mathcal{Y}$ is defined as

$$\mathcal{F}_{\texttt{LIN}}(\mathcal{Y}) := \left\{ f : \mathcal{X} \to \mathcal{Y} \;\middle|\; f(x) = a^\top x + b,\; a \in \mathbb{R}^d_{\geq 0},\; b \in \mathbb{R} \right\}. \tag{4}$$

For brevity, we set $\mathcal{F}_{\texttt{LIN}}([0,1]) = \mathcal{F}_{\texttt{LIN}}$. One can extend this function class by allowing for transformations of each input and the final output. Our main results apply to a larger class, $\mathcal{F}_{\texttt{GENINT}}$, which also allows for interactions among criteria. The name GENINT comes from the fact that this class includes all generalized linear models (linear models with non-decreasing output transformations) and encompasses those with interactions between criteria. Let $\mathcal{P}([d])$ denote the power set of $[d]$. Then $f \in \mathcal{F}_{\texttt{GENINT}}$ if there exists a non-decreasing transformation of outputs $F : \mathbb{R} \to \mathbb{R}$, set of non-decreasing transformations of inputs $h_i : [m] \to \mathbb{R}$ for each i, and coefficients $\{\gamma_I\}_{I \in \mathcal{P}([d])}$ such that for all $x \in \mathcal{X}$:

$$f(x) = F\Big(\sum_{I \in \mathcal{P}([d])} \gamma_I \prod_{i \in I} h_i(x_i)\Big). \tag{5}$$

We use the convention $\prod_{i \in \emptyset}(\cdot) = 1$ so that $\gamma_\emptyset$ serves as an intercept. Notably, $\mathcal{F}_{\texttt{GENINT}}$ contains $\mathcal{F}_{\texttt{LIN}}$, weighted p-means functions used in social choice (Pardeshi et al., 2024), as well as generalized linear models, and (log-)linear models with interactions.

4 Problems with common model assumptions

While we may not believe preferences to follow a model assumption on philosophical grounds (e.g., the inability to capture interactions), the more important question is whether this mismatch meaningfully impacts the outcomes practitioners care about. We investigate consequences for understanding evaluator preferences next, as well as two other common downstream applications or uses—ranking items and estimating the weight or worth of different criteria (these latter two results are discussed in Appendix B). In particular, in this section, we ask: *How well commonly used models capture complex evaluator preferences?*

We assume that the researcher assumes the function is contained in $\mathcal{F}_{\texttt{GENINT}}$ (defined in Equation (5)), which contains linear, generalized linear models, and generalized linear models with criteria interactions. We care about minimizing estimation error defined by Equation (3). Given a set of "worst-case" true preferences, if the researcher outputs an estimated preference function in $\mathcal{F}_{\texttt{GENINT}}$, we ask what the minimum estimation error they could possibly achieve is. That is, if we had access to an oracle that could tell us the best estimation of a given preference function in $\mathcal{F}_{\texttt{GENINT}}$, what is the worst estimation error this oracle could ever incur?

Proposition 4.1. Recall that $\|\widehat{f} - f\|^2_{L_2(P)} \leq 1$ for any $\widehat{f}$ with outputs in $[0,1]$. There exists some universal constant C such that for all $m \geq 4$ and $d \geq 2$,

$$\sup_{f \in \mathcal{F}_{\texttt{ISO}}} \inf_{\widehat{f} \in \mathcal{F}_{\texttt{GENINT}}} \|\widehat{f} - f\|^2_{L_2(P)} \geq C$$

where P is the uniform distribution over $\mathcal{X}$.

We will now discuss what this result means for researchers and practitioners interested in learning preference functions. For any (reasonable) choice of estimator, $\widehat{f}$, the risk is upper-bounded by 1. This means that the worst-case error (e.g., that incurred by a random function over $[0,1]$) is a constant factor. One might expect that as the number of values criterion scores take on, m, increases, or the number of possible interactions (up to 2^d), increase, the additional expressivity of the function class will enable us to better approximate the preference function. Yet, we prove that assuming an incorrect model can lead to error on the same order as random outputs. Our result holds not just for a simple linear model, but also for those allowing for unique transformations of each criterion input, transformations of the output, and as many as 2^d terms corresponding to interactions between criteria. The implication is that one can have a worst-case estimation error even if they employ many of the extensions of linear models used in practice. Hence, even if a researcher or practitioner assumes a very general class of models, they may still end up learning preferences quite poorly.

The full proof of Proposition 4.1 is in Appendix B; we include a proof sketch below.

Proof sketch: We begin by showing that there are ordinal relations over inputs that can be satisfied by functions in $\mathcal{F}_{\texttt{ISO}}$ but not $\mathcal{F}_{\texttt{GENINT}}$. To illustrate how this can arise, consider the case of a linear model. If a model is linear, $f(1,2) > f(2,1)$ implies that coefficient for x_2 is greater than the coefficient for x_1. Therefore, $f(2,3) > f(3,2)$ and in general for any $i > j$, $f(j,i) > f(i,j)$. This example illustrates that the relative value of the coefficients can impose a constraint on the relative value of the outputs. We prove that this same type of issue can arise even for model classes as expressive as

Algorithm 1 Our Algorithm

Require: Dataset of size N, set of regularization parameters Λ containing 0 and ∞

1: Randomly partition $S = [N]$ into a training set S_0 of size $3N/4$ and validation set $S_1 = S \setminus S_0$
2: **for** each $\lambda \in \Lambda$ **do**
3: Compute $(\widehat{f}_\lambda, \widehat{g}_{\texttt{LIN}}) = \texttt{RLS}(S_0, \lambda)$ by solving Equation (6) ▷ solve regularized least squares
4: $\widehat{f}_\lambda \leftarrow \textsc{PostProcess}(\widehat{f}_\lambda, \widehat{g}_{\texttt{LIN}}, S_0, \lambda)$ ▷ truncate and interpolate, Protocol 5.1
5: $R_{\texttt{VAL}}(\lambda) = \sum_{j \in S_1} (\widehat{f}_\lambda(x^{(j)}) - y^{(j)})^2$ ▷ compute validation risk
6: $\lambda^\star \leftarrow \text{argmin}_{\lambda \in \Lambda} R_{\texttt{VAL}}(\lambda)$
7: Repeat Steps 3–5 with dataset S and parameter $\lambda^\star$ to obtain $f_{\texttt{CV}}$. ▷ refit on full dataset
8: **return** $f_{\texttt{CV}}$

$\mathcal{F}_{\texttt{GENINT}}$. If there are such ordinal results that any $\widehat{f} \in \mathcal{F}_{\texttt{GENINT}}$ cannot satisfy, and the true preference function, f does, then $\widehat{f}$ must induce a different ranking over the pairs of inputs than f. If the outputs of f on those input values are non-equal, this leads to some amount of error. When there are enough instances of such sets of inputs, this leads to a very large error. We show that the number of such subsets of inputs is $\Omega(m^d)$. This gives rise to our constant error result. □

5 Our algorithm and theoretical guarantees

In this section, we will describe our algorithmic solution and its theoretical guarantees. Though we focus specifically on the linear assumption, the key techniques we use—cross-validation and regularization—as well as much of our theoretical analysis are applicable to developing robust algorithms for any parametric model choice.

Our goal in developing an algorithm is two-fold: (1) we want to be able to learn what evaluators care about, even if the researcher's assumption of linearity does not hold; (2) we do not want to make the researcher worse off if their assumption does hold. The most natural baseline is non-negative least squares (hereafter *linear regression*), which imposes non-negativity on coefficients to ensure monotonicity and performs similarly to unconstrained regression in practice. Our algorithm should ideally match its performance when linearity holds. By Proposition 4.1, even generalized additive models with interactions can poorly capture more complex evaluator preferences, leading to a constant level of error. Thus, we cannot rely on linear regression alone. Yet if the model truly is linear (as the researcher assumes), it will be easier to learn through linear regression (when we assume linearity) than isotonic regression (when we assume it is merely isotonic). As neither isotonic nor linear regression alone suffices, this motivates the development of a new approach, which we detail in the following section.

5.1 Overview of our algorithm

At a high level, our solution is to solve a regularized least squares problem over isotonic functions. We weigh both how far the observed preferences are from the estimated function (i.e., the fit on preference data) and how far that function is from any linear one (i.e., the amount of model mismatch). Our algorithm is presented as Algorithm 1. We will now interpret the algorithm and the associated sub-routines: regularized isotonic regression, post-processing, and cross-validation.

Regularization. We test a set of regularization weights, denoted by λ, that allows us to interpolate between linear regression ($\lambda = \infty$) and isotonic regression ($\lambda = 0$). Recall that $\mathcal{F}_{\texttt{ISO}}(\mathbb{R})$ and $\mathcal{F}_{\texttt{LIN}}(\mathbb{R})$ are the set of isotonic and linear functions with outputs in $\mathbb{R}$ (see Definitions 1 and 4, respectively). Given $S' \subseteq [N]$ with unique inputs $\mathcal{X}_{S'}$, the core optimization problem is:

$$\texttt{RLS}(S', \lambda) = \min_{\widehat{f} \in \mathcal{F}_{\texttt{ISO}}(\mathbb{R})} \min_{g_{\texttt{LIN}} \in \mathcal{F}_{\texttt{LIN}}(\mathbb{R})} \frac{1}{|S'|} \sum_{j \in S'} \left(\widehat{f}(x^{(j)}) - y^{(j)} \right)^2 + \frac{\lambda}{|\mathcal{X}_{S'}|} \sum_{x \in \mathcal{X}_{S'}} \left(\widehat{f}(x) - g_{\texttt{LIN}}(x) \right)^2. \quad (6)$$

We write $(\widehat{f}, \widehat{g}_{\texttt{LIN}}) = \texttt{RLS}(S', \lambda)$ to denote the outer and inner minimizers of problem (6) respectively.

Protocol 5.1 Post-Processing

Require: Fitted functions $\widehat{f}$ and $\widehat{g}_{\texttt{LIN}}$, training set S', parameter λ

$\widehat{f}(x) \leftarrow \max\Big(0, \min(\widehat{f}(x), 1)\Big)$ for all $x \in \mathcal{X}_{S'}$ ▷ truncate to $[0,1]$

for all $x \notin \mathcal{X}_{S'}$ **do**

  $A(x) \leftarrow \{\widehat{f}(z) : z \in S',\ z \gtrsim x\} \cup \{\max_{z \in S'} \widehat{f}(z)\}$

  $B(x) \leftarrow \{\widehat{f}(z) : z \in S',\ z \lesssim x\} \cup \{\min_{z \in S'} \widehat{f}(z)\}$

  if $\lambda \neq \infty$ **then**

    $\widehat{f}(x) \leftarrow \frac{1}{2}\big(\min A(x) + \max B(x)\big)$ ▷ interpolation

  else

    $\widehat{f}(x) \leftarrow \max\big(0,\ \min(\widehat{g}_{\texttt{LIN}}(x),\ 1)\big)$ ▷ linear extrapolation

return $\widehat{f}$

Post-processing: truncation and interpolation. Ensuring a well-defined preference function in $\mathcal{F}_{\texttt{ISO}}$ as output requires a few additional steps. Because of noise and the regularization component, the outputted function $\widehat{f}$ of Equation (6) may map to values outside of $[0,1]$. Hence, we truncate each regression function $\widehat{f}_\lambda$ to have outputs in $[0,1]$. To estimate values or data-points not seen in training, we use interpolation. For $\lambda = \infty$, we can simply extrapolate our (truncated) linear model. Otherwise, we set the output to be the average of the smallest and largest values it can be while preserving monotonicity and range constraints. These steps are illustrated fully in Protocol 5.1.

Cross-validation. The final regularization weight is carefully chosen from a set of weights, Λ. Cross-validation addresses the fact that, a priori, we do not know a suitable choice of λ. For nonlinear data, it is likely to be smaller. For linear data or data with fewer samples, it is likely to be larger. Cross-validation allows us to carefully pick λ based on how well a set of regression functions $\{\widehat{f}_\lambda\}_{\lambda \in \Lambda}$ fit the true function on an independent set of data. For simplicity, we use hold-out cross-validation and divide the dataset into training and validation subsets (denoted by S_0 and S_1) with a 75-25 split. However, our results hold when the split assigns any constant fraction of the data to both subsets, and one could also employ K-fold cross-validation.

5.2 Theoretical guarantees

In this section, we provide theoretical results to support both goals (1) and (2) discussed in Section 5. Since the re-use of data in step 7 is challenging to analyze, we will consider the estimator $f_{\texttt{CV}} = \widehat{f}_{\lambda^\star}$ before retraining. In practice, however, retraining tends only to improve performance.

Theorem 5.1. Let $f_{\texttt{LIN}}$ be the non-negative least squares estimator solving Equation (6) with $\lambda = \infty$ for the training set S_0, and let $f_{\texttt{CV}} = \widehat{f}_{\lambda^\star}$. Let the additive noise terms, $w^{(j)}$, be independent samples from $\mathcal{N}(0,1)$. Then the following properties hold:

(a) There exists a constant $C > 0$ such that for any distribution P over $\mathcal{X}$,

$$\sup_{f \in \mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{CV}}, f) \leq C \log(|\Lambda|) \sup_{f \in \mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{LIN}}, f).$$

(b) Assume that for some $p_{\texttt{min}} > 0$, $\min_{x \in \mathcal{X}} P(x) \geq \frac{p_{\texttt{min}}}{|\mathcal{X}|}$. Let $\gamma_d = (d^2 + d + 1)/2$ for $d \geq 3$ and $\gamma_2 = 9/2$. Then there exists a constant $C_{d,m,p_{\texttt{min}}} > 0$ depending only on d, m and $p_{\texttt{min}}$ such that

$$\sup_{f \in \mathcal{F}_{\texttt{ISO}}} R(f_{\texttt{CV}}, f) \leq C_{d,m,p_{\texttt{min}}} \log(|\Lambda|) N^{-1/d} \log^{\gamma_d} N$$

Part (a) of the theorem tells us that given a linear preference function, the risk of our algorithm is upper-bounded by the risk of $f_{\texttt{LIN}}$ up to a constant factor depending logarithmically on $|\Lambda|$. Hence, the asymptotic dependence on N, m, and d is just as good for our algorithm as non-negative least squares. Part (b) of the theorem states that if all criteria scores have some positive probability of being observed, the risk of our algorithm decays with $N^{-1/d}$ (plus constant and logarithmic factors). The assumption on $p_{\texttt{min}}$ is minimal (it can be arbitrarily small). As the number of samples increases, this

bound becomes tighter with error vanishing completely as $N \to \infty$. Together, these results illustrate that if one assumes a linear model, they cannot be meaningfully worse off from using our algorithm. And, on the flip side, they may avoid disastrous consequences (as illustrated in Proposition 4.1) if the true preference function is more complex. For synthetic simulations supporting these claims, see Appendix D.1. The full proof of Theorem 5.1 is in Appendix C, and we will provide a brief proof sketch below.

Proof sketch: Proving this result has two general components. Firstly, we use an existing oracle inequality from Vaart et al. (2006) that upper bounds the risk of the cross-validation estimator in terms of the risk of any of the $\widehat{f}_\lambda$ and an additive $\log|\Lambda|/N$ term (up to constant factors). To prove part (a), it suffices to show that $\sup_{f \in \mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{LIN}}, f) = \Omega(1/N)$. This is a well-known lower bound on linear estimators which we re-derive for completeness. Since we include a truncated linear regression estimator for cross-validation, this bound implies that using cross-validation leads to the same asymptotic dependence on d, N and m. To prove part (b), it suffices to prove a $O(N^{-1/d}\log^{\gamma_d} N)$ upper-bound on the risk of isotonic regression. We do so by extending existing results on isotonic regression sample complexity from Han et al. (2019) to our setting, which has a discrete instead of continuous space for random inputs. □

6 Empirical results

In this section, we present two empirical results: a comparison of our algorithm and linear regression real-world preference data, and a qualitative study of LLM and human evaluation preferences.

6.1 Human reviews on Tripadvisor

Our data is of user ratings of hotels on Tripadvisor obtained from the HotelRec dataset (Antognini and Faltings, 2020). Evaluations are comprised of ratings of six criteria (value, rooms, location, cleanliness, service, sleep quality) and overall scores, each rated 1–5 stars (see Figure 1 (a)).

We assess two quantities. First, prediction error: how well $\widehat{f}$ predicts overall scores. Second, since prediction error conflates noise with estimation error, we report *reducible error* — the gap between prediction error and the lowest error achievable by *any* preference function on the test set,

$$\text{Irreducible error} = \frac{1}{|S_2|} \min_{\widetilde{f} \in \mathcal{F}_{\texttt{ISO}}} \sum_{j \in S_2} \left(\widetilde{f}(x^{(j)}) - y^{(j)}\right)^2, \tag{7}$$

$$\text{Reducible error} = \text{Prediction error} - \text{Irreducible error}. \tag{8}$$

Irreducible error proxies for noise; reducible error proxies for estimation error. Since $\widetilde{f}$ minimizes empirical rather than population error, reducible error has a slight negative bias, so if anything it understates the benefit of our algorithm.

6.1.1 Results

In general we find our algorithm to be slightly better then linear regression in terms of prediction error, but significantly better in terms of reducible error (at learning evaluator preferences). Figure 2 shows the prediction error for isotonic regression compared with linear regression. Our algorithm has lower prediction error for each of the traveler groups, and, as indicated by the error bars (computed with prediction error), these differences are statistically significant. The benefit of our algorithm becomes more apparent when looking at reducible error. *For all Tripadvisor datasets, our algorithm results in over a 50% reduction in reducible error compared with linear regression, and as much as a 69% reduction.* Thus, these results further instantiate our theoretical claims in Proposition 4.1 and Theorem 5.1: our algorithm never does much worse than linear regression, can be significantly better.

6.2 LLM and human peer review dataset

Increasingly, large language models are being considered in addition to or in place of human evaluators in areas such as medicine, hiring, and peer review. Given this fact, it is of interest to understand LLM evaluator preferences and how they differ from human ones. In this section, we provide a qualitative

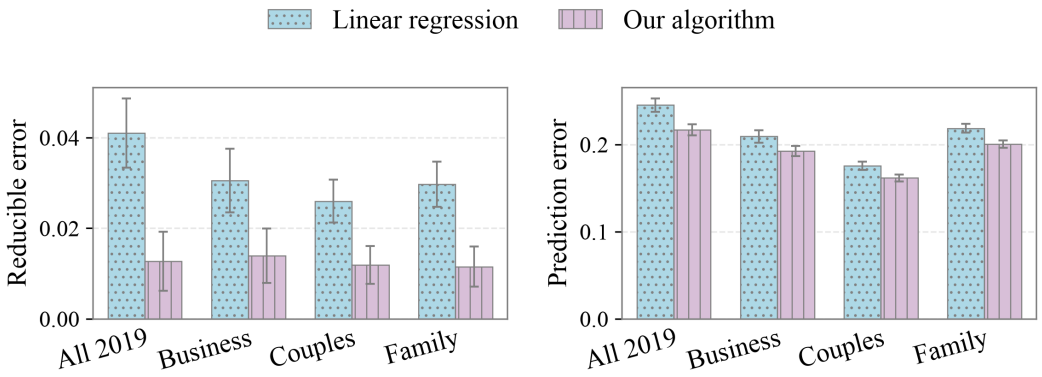


Figure 2: Results of our algorithm relative to linear regression on Tripadvisor rating data. Error bars represent 95% confidence intervals and are computed via the standard deviation of prediction error.

analysis of LLM and human preferences on peer review; this work serves as a case study of how our algorithm may be used in practice.

Our dataset comes from (Yu et al., 2026) and contains peer reviews for papers from ICLR 2024 written by human evaluators, and two LLMs: GPT-4o (OpenAI et al., 2024) and Llama 3.1 70b (Grattafiori et al., 2024). The reviews all contain ratings on three criteria (soundness, presentation, and contribution) scored 1–4, alongside an overall rating scored 1–10 (see Figure 1 (b) for an example).

We assess two qualities: *consistency*, measured via irreducible error ((7)) on each dataset, and *preference misalignment*—the expected disagreement between human and LLM preference functions,

$$\text{Preference misalignment} = \mathbb{E}\,\|\widehat{f}_{\texttt{Human}} - \widehat{f}_{\texttt{LLM}}\|^2_{L_2(P)}, \tag{9}$$

where P is the empirical distribution over criteria scores in the human test set. See Appendix D.3 for additional discussion on methodology. In contrast to the human peer review process, only one evaluator—GPT-4o or Llama—makes all of the evaluations in their respective datasets. Consequently, we may expect LLM preferences to be more consistent than human ones. Regarding preference alignment, in settings such as hiring or resource allocation, minimizing preference misalignment may help to ensure that decisions reflect human needs and values. We test how each of these qualities varies based on the type of LLM.

6.2.1 Results

We present the measurements of both qualities in Figure 3. Our first quantity—irreducible error—captures the level of noise, and tells us how consistent the preferences of human evaluators or a given LLM are across reviews. The second quantity—preference misalignment—concerns the expected discrepancy between human and LLM evaluation preferences. Overall, we find GPT-4o to be significantly better than Llama 3.1 70b on both metrics. *The irreducible error or noise level of GPT is nearly a quarter that of human evaluators*, while the irreducible error of Llama and Human reviews are comparable. Additionally, the *preferences of GPT are significantly closer to human ones* as measured by our distance function. As illustrated by Figure 3, the measures of preference alignment for GPT and Llama differ by a factor of more than five. Though this analysis is not comprehensive enough to draw any conclusions, our case study provides interesting insights, suggesting that preference alignment and consistency may vary widely depending on the specific LLM.

7 Discussion

We will conclude by discussing limitations, broader impacts, and directions for future research. While our assumption on coordinate-wise monotonicity is natural for most evaluation settings, there are situations or aspects for which the direction of a marginal change in criteria is unclear a priori. In these cases, one could extend our techniques to use other non-parametric function classes. Furthermore, for

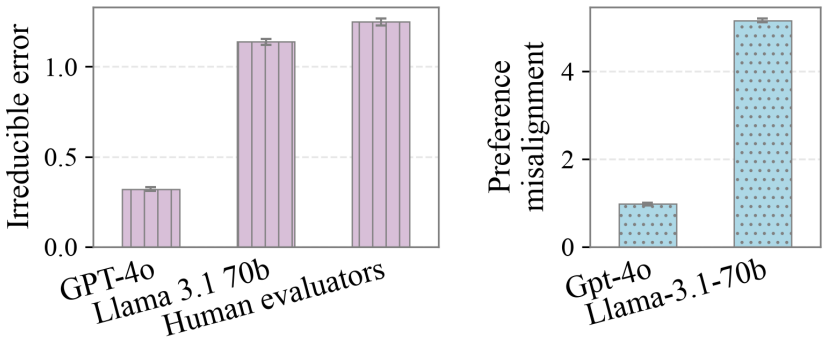


Figure 3: Consistency and preference alignment for LLM (and human) reviewers on the ICLR dataset which provides scores on the 1-10 scale. Error bars for irreducible error reflect 95% bootstrap confidence intervals (percentile method); error bars for preference alignment are computed via the standard error of the mean.

applications in which carefully accounting for the diversity of perspectives is important, employing the axiomatic empirical risk minimization approach of Noothigattu et al. (2021) may be more appropriate. And, as with any work on learning preferences, the insights gained can serve both beneficial ends—improving hiring, peer review, and medical outcomes—and potentially harmful ones; responsible deployment therefore requires careful consideration of context and use case. For future work, one could include non-quantitative inputs such as free-form review summaries to increase the accuracy of learned preferences. Lastly, given the importance of LLMs, one could apply our method to investigate the determinants of and solutions to preference misalignment between human and LLM evaluators.

Acknowledgments

This work was funded through grants NSF 1942124 and ONR N000142512346. We thank R. Ravi for very helpful discussions. We thank Jacqueline Lane for thoughtful discussions and for letting us run the algorithms on a dataset of aerospace and robotics design evaluations. We also thank our anonymous reviewer for catching a mistake in an earlier version of Theorem 5.1.

References

Diego Antognini and Boi Faltings. HotelRec: a Novel Very Large-Scale Hotel Recommendation Dataset, February 2020. URL http://arxiv.org/abs/2002.06854. arXiv:2002.06854 [cs].

Ralph Bradley and Milton Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, December 1952. URL https://www.jstor.org/stable/2334029.

Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On Risk Bounds in Isotonic and Other Shape Restricted Regression Problems, December 2014. URL http://arxiv.org/abs/1311.3765. arXiv:1311.3765 [math] version: 2.

Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On matrix estimation under monotonicity constraints, November 2015. URL http://arxiv.org/abs/1506.03430. arXiv:1506.03430 [math].

Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation, June 2020. URL http://arxiv.org/abs/1903.01395. arXiv:1903.01395 [math].

Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. Adapting a Kidney Exchange Algorithm to Align with Human Values. *Artificial Intelligence*, 283:103261, June 2020. ISSN 00043702. doi: 10.1016/j.artint.2020.10326110.1145/3278721.3278727. URL http://arxiv.org/abs/2005.09755. arXiv:2005.09755 [cs].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J. Samworth. Isotonic Regression in General Dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019. ISSN 0090-5364. URL https://www.jstor.org/stable/26784035.

Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active Ranking from Pairwise Comparisons and when Parametric Assumptions Don't Help, September 2016. URL http://arxiv.org/abs/1606.08842. arXiv:1606.08842 [cs].

Lee Jones, Adrian Barnett, and Dimitrios Vagenas. Common misconceptions held by health researchers when interpreting linear regression assumptions, a cross-sectional study. *PLOS ONE*, 20(6):e0299617, June 2025. ISSN 1932-6203. doi: 10.1371/journal.pone.0299617. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0299617.

Vijay Keswani, Vincent Conitzer, Walter Sinnott-Armstrong, Breanna K. Nguyen, Hoda Heidari, and Jana Schaich Borg. Can AI Model the Complexities of Human Moral Decision-making? A Qualitative Study of Kidney Allocation Decisions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–17, New York, NY, USA, April 2025. Association for Computing Machinery. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3714167. URL https://dl.acm.org/doi/10.1145/3706598.3714167.

Jacqueline Lane, Zoe Szajnfarber, Jason Crusan, Michael Menietti, and Karim Lakhani. Architectural Knowledge and Evaluations: When Does Feasibility Drive Technological Innovation? *Academy of Management Proceedings*, 2023, August 2023. doi: 10.5465/AMPROC.2023.6bp.

Carole J. Lee. Commensuration Bias in Peer Review. *Philosophy of Science*, 82(5):1272–1283, 2015. doi: 10.1086/683652.

Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.*, 3 (CSCW):181:1–181:35, November 2019. doi: 10.1145/3359283. URL https://dl.acm.org/doi/10.1145/3359283.

Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A Voting-Based System for Ethical Decision Making, December 2018. URL http://arxiv.org/abs/1709.06692. arXiv:1709.06692 [cs].

Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Loss Functions, Axioms, and Peer Review. *Journal of Artificial Intelligence Research*, 70:1481–1515, April 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12554. URL https://jair.org/index.php/jair/article/view/12554.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. GPT-4 Technical Report, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Kanad S. Pardeshi, Itai Shapira, Ariel D. Procaccia, and Aarti Singh. Learning Social Welfare Functions. *Advances in Neural Information Processing Systems*, 37:41733–41766, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/499b12df1531fe8ee0febcf08381f3a4-Abstract-Conference.html.

Davor Runje and Sharath M. Shankaranarayana. Constrained Monotonic Neural Networks, May 2023. URL http://arxiv.org/abs/2205.11775. arXiv:2205.11775 [cs.LG].

Nihar B Shah and Martin J Wainwright. Simple, Robust and Optimal Ranking from Pairwise Comparisons. *Journal of Machine Learning Research*, 2018.

Nihar B. Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J. Wainwright. Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues, September 2016a. URL http://arxiv.org/abs/1510.05610. arXiv:1510.05610 [stat].

Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. Feeling the Bern: Adaptive Estimators for Bernoulli Probabilities of Pairwise Comparisons, March 2016b. URL http://arxiv.org/abs/1603.06881. arXiv:1603.06881 [cs].

A.W. van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Risk Modeling*, 24(3):351–371, 2006. URL https://ideas.repec.org//a/bpj/strimo/v24y2006i3p21n3.html.

Jingyan Wang, Nihar B. Shah, and R. Ravi. Stretching the Effectiveness of MLE from Accuracy to Bias for Pairwise Comparisons, June 2019. URL http://arxiv.org/abs/1906.04066. arXiv:1906.04066 [cs].

Yihong Wu. Lecture notes on: Information-theoretic methods for high-dimensional statistics, January 2020.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is Your Paper Being Reviewed by an LLM? Benchmarking AI Text Detection in Peer Review, February 2026. URL http://arxiv.org/abs/2502.19614. arXiv:2502.19614 [cs].

Appendices

A Supplementary material for Section 1

We explain briefly three use cases for preference functions in the real-world. Here we will describe them each in more detail.

- Practitioners and researchers can use the preference function to understand what evaluators' care about. Say, for instance, someone wants to understand the relative value of novelty in a paper. Such a learned preference function can illustrate how improving the rating of novelty (all else constant) alters the overall recommendation of the paper, and how this depends on the other criteria evaluations. An increase in novelty may be more beneficial if the paper also has a high level of clarity, supporting the notion that novel ideas are most useful when they are well-motivated and presented. Further, the benefit may increase as the level of novelty increases, aligning with the idea that particularly novel ideas are especially notable.
- The preference function can reduce discrepancies arising from idiosyncratic evaluation preferences. Often there are multiple evaluators with different preferences. Since every evaluator applies their own unique preferences, overall assessments often depend on the specific preferences of the evaluator to whom an item or individual is assigned. Since such assessments impact decision-making in the real-world, this inequity violates the philosophical principle of equal treatment. To alleviate such discrepancies, we can learn the preference function representing the consensus or collective preferences of the diversity of evaluators. Such a collective or community-level preference function can flag assessments where the mapping of criteria to overall scores deviates significantly from the norm. Then, practitioners can critically asses the flagged evaluation and determine whether the subsequent recommendation or assessment should be altered to align more with collective practice.
- We can use evaluator preferences to increasing alignment in rule-based evaluations. There are many situations where rule-based evaluations or algorithms replace human judgment. This can be for multiple reasons, including the fact that evaluations can be costly in terms of time and financial resources. One example is preliminary diagnostic tests of anxiety and depression (PHQ-9 and GAD-7). In these settings, if we are apply the learned evaluation function instead of the current algorithm (e.g., the predicted psychiatric diagnosis given patient responses), this could allow for the evaluation to more closely align with evaluator beliefs and/or expertise.

B Supplementary material for Section 4

B.1 Proof of Proposition 4.1

For our proof, it is convenient to express the preference function $f : [m]^d \to [0,1]$ by a tensor $T^* \in [0,1]^{m^d}$ defined such that for all $(i_1, \ldots, i_d) \in \mathcal{X}$, $T^*_{i_1,\ldots,i_d} = f(i_1, \ldots, i_d)$ and then to prove the analogous result for the tensor versions of the estimator and true preference function. Let $\mathbb{C}^d_{\texttt{ISO}}$ be the set of isotonic tensors representing all functions in $\mathcal{F}_{\texttt{ISO}}$ and $\mathbb{C}^d_{\texttt{GENINT}}$ be the set of tensors representing all functions in $\mathcal{F}_{\texttt{GENINT}}$. In particular,

$$\mathbb{C}^d_{\texttt{ISO}} := \{T \in [0,1]^{m^d} | T_{i_1,\ldots,i_d} \geq T_{i'_1,\ldots,i'_d} \text{ whenever } i_k \geq i'_k \forall k \in [d]\}.$$

As for the definition of $\mathcal{F}_{\texttt{GENINT}}$, we define $\mathbb{C}^d_{\texttt{GENINT}}$ to be the set of tensors that can be induced by a specific equation. For tensors the analog of the equation for functions in $\mathcal{F}_{\texttt{GENINT}}$, Equation (5), replaces the set of non-decreasing transformations over inputs with a set of non-decreasing vectors.

Let the set of transformations on the outputs be captured by the set of vectors $\{a^{(k)}\}_{k=1}^d$ with increasing coordinates. As for the definition of $\mathcal{F}_{\texttt{GENINT}}$ let the coefficients of the criteria interactions be denoted by $\{\gamma_S\}_{S \in \mathcal{P}([d]}$ and F be a non-decreasing transformation. We say that T is in a member of tensor analog of $\mathcal{F}_{\texttt{GENINT}}$, denoted as $\mathbb{C}^d_{\texttt{INT}}$, if for some $\{a^{(k)}\}_{k=1}^d$, $\{\gamma_S\}_{S \in \mathcal{P}([d]}$, and F:

$$T_{i_1,i_2,\ldots,i_d} = F\left(\sum_{S\in\mathcal{P}([d])} \gamma_S \prod_{k\in S} a_{i_k}^{(k)}\right) \text{ for all pairs } (i_1,i_2,\ldots,i_d)\in[m]^d. \tag{10}$$

By construction, for any two functions $f, \widehat{f}$ mapping $[m]^d \to [0,1]$ with tensor representations T^*, T where P is the uniform distribution, $\frac{1}{m^d}\|T-T^*\|_F^2 = \|f-\widehat{f}\|_{L_2(P)}^2$ where P is the uniform distribution. Using this fact and our definitions of the relevant classes of tensors are defined, we can note that Proposition 4.1 is equivalent to the Proposition B.1.

Proposition B.1. There exists some universal constant C such that for all $m \geq 4$, $d \geq 2$, there exists a tensor $T^* \in \mathbb{C}_{\texttt{ISO}}^d$ such that

$$\frac{1}{m^d}\inf_{T\in\mathbb{C}_{\texttt{GENINT}}^d} \|T-T^*\|_F^2 \geq C.$$

Proof. We will prove this by constructing an explicit T^* such that the proposition holds. To do this, we will first solve the case when $d=2$ and then extend the result to cover the case when $d\geq 3$.

When $d=2$, $f\in\mathcal{F}_{\texttt{ISO}}$ and $\widehat{f}\in\mathcal{F}_{\texttt{GENINT}}$ can be expressed as matrices $M\in\mathbb{C}_{\texttt{ISO}}^2$ and $\widehat{M}\in\mathbb{C}_{\texttt{GENINT}}^2$, respectively. To construct such a counterexample, we will show that there is a set of inequalities over entries that M can satisfy but $\widehat{M}$ cannot. This implies that over that subset of indices, $\widehat{M}$ must incur some amount of error. We then construct a matrix such that the set of those inequalities has size of order m^d to prove our bias result. The specific ordering and proof of the impossibility for any $\widehat{M}\in\mathbb{C}_{\texttt{GENINT}}^2$ is the subject of Lemma B.2 which we will prove at the end of this section.

Lemma B.2. Assume that $\widehat{M}\in\mathbb{C}_{\texttt{GENINT}}(F)$ for some non-decreasing F and $m=4$. Then the following inequalities cannot all hold:

$$\widehat{M}_{12} < \widehat{M}_{21} < \widehat{M}_{31} < \widehat{M}_{22} \tag{11}$$
$$\widehat{M}_{22} < \widehat{M}_{13} < \widehat{M}_{23} < \widehat{M}_{32} \tag{12}$$
$$\widehat{M}_{31} < \widehat{M}_{22} < \widehat{M}_{32} < \widehat{M}_{41} \tag{13}$$
$$\widehat{M}_{23} < \widehat{M}_{32} < \widehat{M}_{42} < \widehat{M}_{33} \tag{14}$$

Taking Lemma B.2 as given, we will construct a general $m\times m$ matrix M and showing that it satisfies

$$\frac{1}{m^2}\inf_{\widehat{M}\in\mathbb{C}_{\texttt{GENINT}}^2} \|\widehat{M}-M\|_F^2 \geq C$$

for some universal constant $C>0$. We will do this by using a 4×4 matrix, B, defined as follows

$$B = \frac{1}{12}\begin{pmatrix} 1 & 2 & 6 & 12 \\ 3 & 5 & 7 & 12 \\ 4 & 8 & 11 & 12 \\ 9 & 10 & 12 & 12 \end{pmatrix}$$

Note that $B\in\mathbb{C}_{\texttt{BISO}}$. Additionally, it satisfies the inequalities in Lemma B.2. These inequalities for our specific choice of B evaluate to

$$\begin{aligned} B_{12} < B_{21} < B_{31} < B_{22} &\implies 2/12 < 3/12 < 4/12 < 5/12 \\ B_{22} < B_{13} < B_{23} < B_{32} &\implies 5/12 < 6/12 < 7/12 < 8/12 \\ B_{31} < B_{22} < B_{32} < B_{41} &\implies 4/12 < 5/12 < 8/12 < 9/12 \\ B_{23} < B_{32} < B_{42} < B_{33} &\implies 7/12 < 8/12 < 10/12 < 11/12 \end{aligned}$$

Define M to have entries $M_{ij} = B_{\lceil (4i)/m\rceil,\lceil (4j)/m\rceil}$ for all $(i,j)\in[m]^2$ and assume that $m/4$ is an integer (we will later show that this is without loss of generality). Therefore, we can write M as

$$M = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}$$

where A_{kl} for $(k,l) \in [4]^2$ are $n/4 \times n/4$ constant matrices where each entry is equal to B_{kl}. We can partition M into $m^2/16$ sub-matrices that are all equal to B. For instance, let each sub-matrix be defined by having entry i, j of each of the blocks $A_{11}, \ldots, A_{44}$ (this means that the sub-matrix has rows $\{i, n/4+i, 2n/4+i, 3n/4+i\}$ and columns $\{j, n/4+j, 2n/4+j, 3n/4+j\}$).

Let $\widehat{M} \in \mathbb{C}^2_{\texttt{GENINT}}$ be any estimate for M. By construction, any sub-matrix of a $\widehat{M}$ must also be in $\mathbb{C}^2_{\texttt{GENINT}}$. Let B' be the estimate of B for any of the sub-matrices defined above. By Lemma B.2, it must be the case that one of the four inequalities does not hold on B'.

This implies that at least one of the ordinal relations between the entries of B must be flipped or hold with equality. Since each relevant entry (those in the first 3 columns) is at least $1/12$ apart, we have that $\|B - B'\|_F^2 \geq 2 \cdot \frac{1}{24^2} = \frac{1}{288}$. This is because squared error is minimized when the change is the same for both entries contained in the ordinal relation. Since the total change is at least $1/12$, they individually must change at least by $1/24$.

Summing over the $m^2/16$ sub-matrices equal to B, we have that $\|M - \widehat{M}\|_F^2 \geq m^2/(16 \cdot 288)$. As this inequality holds for all F, a, b we have that for the M defined above

$$\frac{1}{m^2} \inf_{\widehat{M} \in \mathbb{C}^2_{\texttt{GENINT}}} \|M - \widehat{M}\|_F^2 \geq 1/(16 \cdot 288)$$

Note that the assumption that $m/4$ is an integer is without loss of generality. If $m/4$ was not an integer, we could make the same argument for the first $4\lfloor m/4 \rfloor$ rows and columns (corresponding to more than a quarter of all entries) and fill the remaining rows and columns with all 1. This would give us error at least $\frac{1}{4}m^2/(16 \cdot 288)$. Setting $C = \frac{1}{4}/(16 \cdot 288)$, then the above inequality holds when $d = 2$ and $m \geq 4$.

In order to extend our result to cases when $d > 2$, we will use Lemma B.3 which tells us that when fixing $d-2$ coordinates for a $T \in \mathbb{C}^d_{\texttt{GENINT}}$, the resulting matrix must be in $\mathbb{C}^2_{\texttt{GENINT}}$. This allows us to extend the bias result for $d = 2$ to the case when $d > 2$.

Lemma B.3. Assume that $T \in \mathbb{C}^d_{\texttt{GENINT}}$ and $d \geq 3$. Then the matrix $\widehat{M}$ formed by fixing any $d-2$ indices is in $\mathbb{C}^2_{\texttt{GENINT}}$.

Let M be a $m \times m$ matrix such Proposition B.1 holds (we have just proved the existence of one). Set $T^*_{i_1,i_2,\ldots,i_d} = M_{i_1,i_2}$ for any $(i_1, \ldots, i_d) \in [m]^d$ so that it is constant with respect to all but the first two indices. Consider the approximation of T^* by any $T \in \mathbb{C}_{\texttt{GENINT}}(F)$ for some non-decreasing F. For simplicity of notation, we will use $T_{\cdot,i_{3:d}}$ to denote the matrix form by fixing the last $d-2$ indices of the tensor. By the definition of the Frobenius norm and our construction of T^*:

$$\begin{aligned}
\|T - T^*\|_F^2 &= \sum_{i_1=1}^m \sum_{i_2=1}^m \cdots \sum_{i_d=1}^m (T^*_{i_1,i_2,\ldots,i_d} - T_{i_1,i_2,\ldots,i_d})^2 \\
&= \sum_{i_3=1}^m \sum_{i_4=1}^m \cdots \sum_{i_d=1}^m (M - T_{\cdot,i_{3:d}})^2
\end{aligned}$$

By our choice of M we have that for some universal constant $C > 0$, $\inf_{\widehat{M} \in \mathbb{C}^2_{\texttt{GENINT}}} \|M - \widehat{M}\|_F^2 \geq Cm^2$. Thus, multiplying this by all m^{d-2} different pairs $(i_3, \ldots, i_d) \in [m]^{d-2}$, we have our claimed result

$$\frac{1}{m^d}\|T - T^*\|_F^2 \geq C.$$

□

Now all that is left is to prove the remaining lemmas, Lemma B.2 and Lemma B.3.

Proof of Lemma B.2. Our goal is to show that the series of inequalities is not compatible with any set of parameters and non-decreasing function F. In this proof we will leverage the fact that the ordering over the entries has to be determined by the relations of the inputs to F. From the first two inequalities we learn something about the relative values of $(\gamma_1 + \gamma_{12}b_1)$, $(\gamma_1 + \gamma_{12}b_2)$, $(\gamma_1 + \gamma_{12}b_2)$ such both

inequalities can hold at the same time. The second two inequalities are constructed to be of the same form, but provide different requirements on the relative values of $\gamma_1 + \gamma_{12} b_j$ for $j \in \{1, 2, 3\}$. We can show that these requirements are not compatible with a single set of parameters, allowing us to reach a contradiction.

Assume for the sake of contradiction that $\widehat{M} \in \mathbb{C}_{\texttt{GENINT}}$ for some $F, a, b, c, \gamma_1, \gamma_2, \gamma_{12}$. Because our argument pertains to the ordinal relationship between entries, we can assume WLOG that $c = 0$.

The first inequality implies that for some $F, a, b, \gamma_1, \gamma_2, \gamma_{12}$

$$\begin{aligned} F(\gamma_1 a_1 + \gamma_2 b_2 + \gamma_{12} a_1 b_2) &< F(\gamma_1 a_2 + \gamma_2 b_1 + \gamma_{12} a_2 b_1) \\ &< F(\gamma_1 a_3 + \gamma_2 b_1 + \gamma_{12} a_3 b_1) \\ &< F(\gamma_1 a_2 + \gamma_2 b_2 + \gamma_{12} a_2 b_2). \end{aligned}$$

As F is non-decreasing, the ordinal relationship of the inputs is the same as the outputs. Consequently,

$$\begin{aligned} \gamma_1 a_1 + \gamma_2 b_2 + \gamma_{12} a_1 b_2 \gamma_1 a_2 + \gamma_2 b_1 + \gamma_{12} a_2 b_1 \\ &< \gamma_1 a_3 + \gamma_2 b_1 + \gamma_{12} a_3 b_1 \\ &< \gamma_1 a_2 + \gamma_2 b_2 + \gamma_{12} a_2 b_2. \end{aligned}$$

Since the difference between the outer two terms is larger than that between the inner two terms we have that

$$\begin{aligned} &(\gamma_1 a_2 + \gamma_2 b_2 + \gamma_{12} a_2 b_2) - (\gamma_1 a_1 + \gamma_2 b_2 + \gamma_{12} a_1 b_2) \\ &> (\gamma_1 a_3 + \gamma_2 b_1 + \gamma_{12} a_3 b_1) - (\gamma_1 a_2 + \gamma_2 b_1 + \gamma_{12} a_2 b_1)\,. \end{aligned}$$

Rearranging, we have that

$$(\gamma_1 + \gamma_{12} b_2)(a_2 - a_1) > (\gamma_1 + \gamma_{12} b_1)(a_3 - a_2).$$

Following the same argument the remaining three inequalities imply:

$$\begin{aligned} (\gamma_1 + \gamma_{12} b_3)(a_2 - a_1) &< (\gamma_1 + \gamma_{12} b_2)(a_3 - a_2) \\ (\gamma_1 + \gamma_{12} b_1)(a_4 - a_3) &> (\gamma_1 + \gamma_{12} b_2)(a_3 - a_2) \\ (\gamma_1 + \gamma_{12} b_2)(a_4 - a_3) &< (\gamma_1 + \gamma_{12} b_3)(a_3 - a_2) \end{aligned}$$

There are two cases.

Case 1: Both $(\gamma_1 + \gamma_{12} b_3)$ and $(\gamma_1 + \gamma_{12} b_2)$ are non-zero. This allows us to combine the first two inequalities to get

$$\frac{\gamma_1 + \gamma_{12} b_1}{\gamma_1 + \gamma_{12} b_2}(a_3 - a_2) < a_2 - a_1 < \frac{\gamma_1 + \gamma_{12} b_2}{\gamma_1 + \gamma_{12} b_3}(a_3 - a_2)$$

Because $a_3 - a_2$ is non-negative by assumption, in order for this to hold, we then require that $\frac{\gamma_1 + \gamma_{12} b_2}{\gamma_1 + \gamma_{12} b_3} > \frac{\gamma_1 + \gamma_{12} b_1}{\gamma_1 + \gamma_{12} b_2}$. Likewise, we can derive from the last two inequalities that

$$\frac{\gamma_1 + \gamma_{12} b_2}{\gamma_1 + \gamma_{12} b_3}(a_4 - a_3) < a_3 - a_2 < \frac{\gamma_1 + \gamma_{12} b_1}{\gamma_1 + \gamma_{12} b_2}(a_4 - a_3)$$

which require that $\frac{\gamma_1 + \gamma_{12} b_2}{\gamma_1 + \gamma_{12} b_3} < \frac{\gamma_1 + \gamma_{12} b_1}{\gamma_1 + \gamma_{12} b_2}$, a contradiction.

Case 2: If one of $(\gamma_1 + \gamma_{12} b_3)$ and $(\gamma_1 + \gamma_{12} b_2)$ are zero, we will show this implies many of the entries are tied. If $(\gamma_1 + \gamma_{12} b_j) = 0$ for $j \in \{2, 3\}$, we will show that the jth column of $\widehat{M}$ must be constant. For any $i \in [m]$:

$$\begin{aligned} \widehat{M}_{ij} &= F(\gamma_1 a_i + \gamma_2 b_j + \gamma_{12} a_i b_j) \\ &= F((\gamma_1 + \gamma_{12} b_j) a_i + \gamma_2 b_j) \\ &= F(\gamma_2 b_j) \end{aligned}$$

Since the value of $\widehat{M}_{ij}$ does not depend on i, this would imply that the jth column is constant. For $j = 3$, we would have $\widehat{M}_{23} = \widehat{M}_{33}$, and for $j = 2$ we would have $\widehat{M}_{12} = \widehat{M}_{22}$. Either case gives us a contradiction since those inequalities are strict in $\widehat{M}$.

Since both cases lead to a contradiction, the four inequalities cannot hold simultaneously for some $\widehat{M} \in \mathbb{C}^2_{\texttt{GENINT}}$.

Proof of Lemma B.3. By assumption $T \in \mathbb{C}_{\texttt{GENINT}}(F)$ for some function F. Therefore, T is induced by Equation (10) for some scalars $\{\gamma_S\}_{S \in \mathcal{P}([d])}$, non-decreasing vectors $a^{(1)}, \ldots, a^{(d)} \in \mathbb{R}^m$, and non-decreasing function F.

WLOG we will fix $i_3, \ldots, i_d$ (the same arguments hold for fixing any $d-2$ values of the criteria scores). Consider the matrix $\widehat{M} = T_{\cdot, i_{3:d}} \in \mathbb{R}^{m \times m}$. Observe that we can re-write the input for F for an arbitrary set of indices $(i_1, i_2) \in [m]^2$ as:

$$\begin{aligned}
\sum_{S \in \mathcal{P}([d])} \gamma_S \prod_{k \in S} a_{i_k}^{(k)} &= \sum_{S \in \mathcal{P}([d]): |S \cap \{1,2\}| = 0} \gamma_S \prod_{k \in S} a_{i_k}^{(k)} + \sum_{S \in \mathcal{P}([d]): 1 \in S, 2 \notin S} \gamma_S \prod_{k \in S} a_{i_k}^{(k)} \\
&+ \sum_{S \in \mathcal{P}([d]): 1 \notin S, 2 \in S} \gamma_S \prod_{k \in S} a_{i_k}^{(k)} + \sum_{S \in \mathcal{P}([d]): 1 \in S, 2 \in S} \gamma_S \prod_{k \in S} a_{i_k}^{(k)} \\
&= \left(\sum_{S \in \mathcal{P}([d]): |S \cap \{1,2\}| = 0} \gamma_S \prod_{k \in S} a_{i_k}^{(k)} \right) + a_{i_1}^{(1)} \left(\sum_{S \in \mathcal{P}([d]): 1 \in S, 2 \notin S} \gamma_S \prod_{k \in S \setminus \{1\}} a_{i_k}^{(k)} \right) \\
&+ a_{i_2}^{(2)} \left(\sum_{S \in \mathcal{P}([d]): 1 \notin S, 2 \in S} \gamma_S \prod_{k \in S \setminus \{2\}} a_{i_k}^{(k)} \right) \\
&+ a_{i_1}^{(1)} a_{i_2}^{(2)} \left(\sum_{S \in \mathcal{P}([d]): 1 \in S, 2 \in S} \gamma_S \prod_{k \in S \setminus \{1,2\}} a_{i_k}^{(k)} \right)
\end{aligned}$$

Note that all components in parenthesis are constant with respect to i_1 and i_2. For each element S of the power set $\{1, 2, (1,2), 0\}$ we see that there is a constant that multiplies $\prod_{k \in S} a_{i_k}^{(k)}$. Set $a = a^{(1)}$, $b = b^{(2)}$ and the constants $c, \gamma_1, \gamma_2, \gamma_{12}$ to be the values in parenthesis

$$\begin{aligned}
c &= \sum_{S \in \mathcal{P}([d]): |S \cap \{1,2\}| = 0} \gamma_S \prod_{k \in S} a_{i_k}^{(k)} \\
\gamma_1 &= \sum_{S \in \mathcal{P}([d]): 1 \in S, 2 \notin S} \gamma_S \prod_{k \in S \setminus \{1\}} a_{i_k}^{(k)} \\
\gamma_2 &= \sum_{S \in \mathcal{P}([d]): 1 \notin S, 2 \in S} \gamma_S \prod_{k \in S \setminus \{2\}} a_{i_k}^{(k)} \\
\gamma_{12} &= \sum_{S \in \mathcal{P}([d]): 1 \in S, 2 \in S} \gamma_S \prod_{k \in S \setminus \{1,2\}} a_{i_k}^{(k)}.
\end{aligned}$$

Then is it easy to see that for all $(i_1, i_2) \in [m]^2$,

$$\widehat{M}_{i_1, i_2} = T_{i_1, i_2, \ldots, i_d} = F(\gamma_1 a_{i_1} + \gamma_2 b_{i_2} + \gamma_{12} a_{i_1} b_{i_2} + c).$$

Hence $M \in \mathbb{C}_{\texttt{GENINT}}(F)$ as desired. □

□

B.2 Ranking items

In many evaluation settings, it is of interest to create a ranking or prioritization among a set of items, individuals, or options. In resource allocation problems, we may care about identifying individuals or communities who would benefit the most from investment. In admissions and hiring processes, there is typically a fixed number of applicants who can receive offers (i.e., only the top k are selected). Concretely, Lee et al. (2019) applies learned individual preferences to rank prioritization of food bank recipients.

In this section, we consider the issues that could arise when using an estimated preference function to rank items when the researcher incorrectly assumes a model class. We consider a broad class

of models. In particular, we consider a linear model where there is a set of non-decreasing transformations applied to each of the criteria inputs, and a transformation applied to the output of the weighted combination of the (transformed) inputs. The set of models allowing for such input and output transformations are called Generalized Additive Models of GAMs, denoted by the class $\mathcal{F}_{\texttt{GAM}}$. We say that a function $f \in \mathcal{F}_{\texttt{GAM}}$ if there exists a set of input transformations $h_i : [m] \to \mathbb{R}$ for all $i \in [d]$, monotonic transformation of outputs F, and set of coefficients $\{\gamma_i\}_{i\in[d]}$ such that for all $x \in \mathcal{X}$

$$f(x) = F\left(\sum_{i\in[d]} \gamma_i h_i(x)\right). \tag{15}$$

To measure the ability of the estimated preference function to rank items, we consider the rankings over all possible sets of criteria scores $x \in \mathcal{X}$ induced by the estimated function $\widehat{f}$, denoted by $\pi_{\widehat{f}}$. We compare $\pi_{\widehat{f}}$ to the ranking over sets of criteria scores induced by the true preference function f, denoted by π_f. The distance between the two rankings is measured by a normalized Kendall tau distance. Normalized Kendall tau distance sums the ranking error over all unique pairs of inputs, and then divides by the total number of input pairs. We consider a variant of normalized Kendall tau distance that accounts for ties only when they occur in $\pi_{\widehat{f}}$ and the pairs are ranked in π_f. Otherwise, any constant function $\widehat{f}$ would have a distance of zero. The error for inputs $x \neq x' \in \mathcal{X}$ is

$$\overline{K}(x, x'; \pi_{\widehat{f}}, \pi_f) = \begin{cases} 1 & \text{relative order is different under } \widehat{f} \text{ and } f \\ \frac{1}{2} & \text{tied in } \widehat{f}, \text{ ranked in } f \\ 0 & \text{otherwise.} \end{cases}$$

And, the total distance is

$$K(\pi_{\widehat{f}}, \pi_f) = \frac{1}{\binom{m^d}{2}} \sum_{\text{unique } (x,x')\in\mathcal{X}^2} \overline{K}(x, x'; \pi_{\widehat{f}}, \pi_f).$$

The worst-case error of $\widehat{f}$ is 1 when all pairs are ranked incorrectly. We show in Proposition B.4 that there exist isotonic preference functions such that for any $\widehat{f} \in \mathcal{F}_{\texttt{GAM}}$, this worst-case error is incurred (up to constant factors).

Proposition B.4. There exists some universal constant $C > 0$ such that for all $m \geq 3$ and $d \geq 2$, there exists a $f \in \mathcal{F}_{\texttt{ISO}}$ such that

$$\inf_{\widehat{f}\in\mathcal{F}_{\texttt{GAM}}} K(\pi_{\widehat{f}}, \pi_f) \geq C.$$

Our result implies that when there is a model mismatch it is possible that a constant fraction of all pairs of criteria scores will be ranked incorrectly. In hiring and peer review, an incorrect ranking means that a better paper or applicant would be ranked *below* a worse one. Combined with our worst-case result on learning evaluator preferences Proposition 4.1, our findings underscore theoretically substantial issues that could result from many common model assumptions.

B.2.1 Proof of Proposition B.4

Our goal is to show that there exists a preference function $f \in \mathcal{F}_{\texttt{ISO}}$ such that when approximated by any function $\widehat{f} \in \mathcal{F}_{\texttt{GAM}}$, there is a worst-case Kendall-tau error. As in the proof of Proposition 4.1 we will use the same tensor representation of the preference function f and prove the analogous result for the tensor representations of any estimator $\widehat{f} \in \mathcal{F}_{\texttt{GAM}}$ and a worst-case function $f \in \mathcal{F}_{\texttt{ISO}}$. Recall that the class $\mathbb{C}^d_{\texttt{ISO}}$ represents all functions in $\mathcal{F}_{\texttt{ISO}}$. Further, we can express all functions in $\mathcal{F}_{\texttt{GAM}}$ by the set of tensors. For notational simplicity we will represent indices both as the list of their components (for instance, $(i_1, \ldots, i_d)$) and simply the vector i. We say that a tensor corresponds to an $f \in \mathcal{F}_{\texttt{GAM}}$, denoted by $T \in \mathbb{C}^d_{\texttt{GENLIN}}$, if there exists a set of vectors $\{a^{(k)}\}_{k=1}^d \subset \mathbb{R}^m$ (encoding the coordinate-wise transformations), a non-decreasing F, and a $c \in \mathbb{R}$ such that for all $i \in \mathcal{X}$:

$$T_i = F\left(\sum_{k\in[m]} a^{(k)}_{i_k} + c\right) \tag{16}$$

We will confirm that any $f \in \mathcal{F}_{\texttt{GAM}}$ has a corresponding $T \in \mathbb{C}^d_{\texttt{GENLIN}}$. For any $f \in \mathcal{F}_{\texttt{GAM}}$ expressed by Equation (15), this is accomplished by setting $F = F$, $b = c$, and for all $k \in [d$ and $x_k \in m$, $a^{(k)}_{x_k} = a_k h_k(x_k)$. Therefore, the set of permutations π_f for $f \in \mathcal{F}_{\texttt{GAM}}$ are contained in those of π^*_T for $T^* \in \mathbb{C}^d_{\texttt{GENLIN}}$. Let the corresponding tensors of f and $\widehat{f}$ be T^* and T, respectively. Then

$$\begin{aligned} K(\pi_{\widehat{f}}, \pi_f) &= K(\pi_T, \pi^*_T) \\ &= \frac{1}{\binom{|\mathcal{X}|}{2}} \sum_{\text{unique } i \neq i' \in \mathcal{X}} \overline{K}(i, i', \pi_T, \pi^*_T). \end{aligned}$$

Therefore, since $\mathbb{C}_{\texttt{GENLIN}}$ contains the tensor representations of all $\mathcal{F}_{\texttt{GAM}}$, the statement of Proposition B.4 is implied by the following proposition

Proposition B.5. There exists some universal constant $C > 0$ such that for all $m \geq 3$ and $d \geq 2$, there exists a $T^* \in \mathbb{C}^d_{\texttt{ISO}}$ such that

$$\frac{1}{\binom{|\mathcal{X}|}{2}} \inf_{\widehat{M} \in \mathbb{C}^d_{\texttt{GENLIN}}} K(\pi_{T^*}, \pi_T) \geq C.$$

Proof of Proposition B.5. We will prove this by constructing an explicit T^* that satisfies the conditions of Proposition B.5. Let's assume that $m/3$ is an integer and use the approximation of $\binom{|\mathcal{X}|}{2}$ by m^{2d} (this is WLOG up to constant factors). As in the proof of Proposition B.1, we will start with the case when $d = 2$ and then extend our construction to cases when $d > 2$. At a high-level, our goal is to show that there are a series of inequalities that cannot be satisfied by any $T \in \mathbb{C}^d_{\texttt{GENLIN}}$, and to show that the number of such inequalities between pairs of entries leads to a worst-case Kendall tau distance. Since there are a number of lemma's involved in this proof, we will leave the proof of each lemma to the end of the section.

When $d = 2$, we can represent f and $\widehat{f}$ by matrices $f \in \mathbb{C}^2_{\texttt{ISO}}$ and $\widehat{f} \in \mathbb{C}^2_{\texttt{GENLIN}}$, respectively. First we show that such a set of inequalities over entries exists in a 3×3 matrix.

Lemma B.6. Assume that $\widehat{M} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ for some F and that $m = 3$. Then the following inequalities cannot both hold:

$$\widehat{M}_{12} < \widehat{M}_{21} < \widehat{M}_{31} < \widehat{M}_{22}$$

and

$$\widehat{M}_{22} < \widehat{M}_{13} < \widehat{M}_{23} < \widehat{M}_{32}.$$

Let $B = \frac{1}{10}\begin{pmatrix} 1 & 2 & 6 \\ 3 & 5 & 7 \\ 4 & 8 & 9 \end{pmatrix}$ be a 3×3 bivariate isotonic matrix that satisfies the ordinal relations in Lemma B.6. We will prove our result for a matrix M which can be written as

$$M = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} + \begin{pmatrix} E & E & E \\ E & E & E \\ E & E & E \end{pmatrix}$$

where A_{kl} for $(k, l) \in [3]^2$ are $n/3 \times m/3$ constant matrices where each entry is equal to B_{kl}. We can partition M into $m^2/9$ sub-matrices that are all equal to B and $E \in \mathbb{R}^{m/3 \times m/3}$ and entries $E_{ij} = \frac{1}{10m} i + \frac{1}{10m^2} j$. Here the additional terms E create ordering within each of the blocks and ensure that each entry is unique. Explicitly, $M_{ij} = B_{\lceil (3i)/m \rceil, \lceil (3j)/m \rceil} + \epsilon_{ij}$ where

$$\varepsilon_{ij} = \frac{1}{10m^2}\left(i - \frac{n}{3}(\lceil ni/3 \rceil - 1)\right) + \frac{1}{10m}\left(j - \frac{n}{3}(\lceil mj/3 \rceil - 1)\right).$$

We will now verify that $M \in \mathbb{C}_{\texttt{BISO}}$ by confirming that it is isotonic and has entries between 0 and 1. First we will show that it is isotonic. Each submatrix A_{kl} is isotonic by construction. Let (i, j) and (i', j') be entries from different sub-matrices such that $i \geq i'$ and $j \geq j'$ where are least one inequality holds strictly. Then

$$M_{ij} - M_{i'j'} \geq \frac{1}{10} - \frac{m}{3}\left(\frac{1}{10m^2} + \frac{1}{10m}\right) > \frac{7}{120}.$$

By construction, $M_{ij} \geq M_{11} \geq 0.1 > 0$. Additionally,

$$M_{ij} \leq M_{nn} = 0.9 + \frac{1}{10m^2}(m/3) + \frac{1}{10m}(m/3) < 0.9 + \frac{7}{120} < 1.$$

Hence $M \in \mathbb{C}_{\texttt{BISO}}$. We will now lower-bound the Kendall tau distance from M to any matrix $\widehat{M} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ for some non-decreasing F.

We can partition M into $L = m^2/9$ sub-matrices with ordering the same as B. Let each sub-matrix be defined rows $\{i, m/3+i, 2m/3+i\}$ and columns $\{j, m/3+j, 2m/3+j\}$. Let $S_1, \ldots, S_L$ denote the set of indices that correspond to each sub-matrix. We can rearrange the terms in the Kendall tau distance to get:

$$\begin{aligned}
K(\pi_M, \pi_{\widehat{M}}) &= \sum_{\text{unique } (i,j),(i',j') \in [m]^2} \overline{K}_{(i,j),(i',j')}(\pi_{\widehat{M}}, \pi_M) \\
&= \sum_{\ell=1}^{L} \sum_{(i,j) \neq (i',j') \in S_\ell} \overline{K}_{(i,j),(i',j')}(\pi_{\widehat{M}}, \pi_M) \\
&+ \sum_{\ell=2}^{L} \sum_{\ell=1}^{\ell-1} \sum_{(i,j) \in S_\ell, (i',j') \in S_{\ell'}} \overline{K}_{(i,j),(i',j')}(\pi_{\widehat{M}}, \pi_M)
\end{aligned}$$

Define K_ℓ to be the distance within the sub-matrix $\ell \in [L]$, and $K_{\ell,\ell'}$ to be the distance between the permutations corresponding to sub-matrices ℓ and $\ell' \in [L]$.

$$\begin{aligned}
K_{\ell,\ell'} &= \sum_{(i,j) \in S_\ell, (i',j') \in S_{\ell'}} \overline{K}_{(i,j),(i',j')}(\pi_{\widehat{M}}, \pi_M) \\
K_\ell &= \sum_{(i,j) \neq (i',j') \in S_\ell} \overline{K}_{(i,j),(i',j')}(\pi_{\widehat{M}}, \pi_M)
\end{aligned}$$

The remainder of this proof will be devoted to showing that $K_{\ell,\ell'} \geq 1$ and $K_\ell \geq 1/2$. If this holds when we have

$$K(\pi_M, \pi_{\widehat{M}}) \geq L + \binom{L}{2} = m^2/9 + \frac{m^2/9(m^2/9-1)}{2} \geq m^4/162.$$

Case 1: No ties and incorrect ordering. If there exists some pair of entries $(i,j) \in S_\ell$ and $(i',j') \in S_{\ell'}$ where the relative order in $\widehat{M}$ is different from in M, then we have $K_{\ell,\ell'} \geq 1$. Likewise, if there exists some pair of entries $(i,j) \neq (i',j') \in S_\ell$ where the relative order in $\widehat{M}$ is different from in M, then we have $K_\ell \geq 1$.

Case 2: If all entries have the same ordering up to ties, we now need to show that there are at least two ties between non-equal entries. First we will consider $K_{\ell,\ell'}$. Each submatrix can be expressed by $B + E_{ij}$ for some $(i,j) \in [m/3]^2$. By construction, the values of E_{ij} are unique. Therefore, $M_{S_\ell} > M_{S_{\ell'}}$ or $M_{S_\ell} < M_{S_{\ell'}}$ depending on the relative value of E_{ij} for which the submatrices are defined. WLOG assume $M_{S_\ell} > M_{S_{\ell'}}$. In order for there to be zero error, it must be true that for the approximation $\widehat{M}$:

$$\widehat{M}_{S_{\ell'},12} \leq \widehat{M}_{S_\ell,12} \leq \widehat{M}_{S_{\ell'},21} \leq \widehat{M}_{S_\ell,21} \leq \widehat{M}_{S_{\ell'},31} \leq \widehat{M}_{S_\ell,31} \leq \widehat{M}_{S_{\ell'},22} \leq \widehat{M}_{S_\ell,22}$$

and

$$\widehat{M}_{S_{\ell'},22} \leq \widehat{M}_{S_\ell,22} \leq \widehat{M}_{S_{\ell'},13} \leq \widehat{M}_{S_\ell,13} \leq \widehat{M}_{S_{\ell'},23} \leq \widehat{M}_{S_\ell,23} \leq \widehat{M}_{S_{\ell'},32} \leq \widehat{M}_{S_\ell,32}.$$

It is easy to show that any sub-matrix of a matrix $\widehat{M} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ is also in $\mathbb{C}_{\texttt{GENLIN}}(F)$. This implies that $\widehat{M}_{S_{\ell'}}, \widehat{M}_{S_\ell} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ satisfy the ordinal relations in Lemma B.6 with weak inequalities.

Let us consider two cases. First, for errors within blocks. If $\widehat{M}_{S_\ell}$ has the same ordering as $M_{S_\ell} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ up to ties, then at least one of the consecutive entries must be tied. As such, $K_\ell \geq 1/2$. Otherwise, we know that there is an inversion implying $K_\ell \geq 1$. Consequently, we have

that for all ℓ, $K_\ell \geq 1/2$.

Now we will focus on errors between blocks. Again, if there is an inversion between items we have that $K_{\ell,\ell'} \geq 1$ trivially. If not, then we require that there is a tie between items within one of the blocks, which gives rise to at least two ties between the blocks. Thus $K_{\ell,\ell'} \geq 1$ even if there are no inversions.

Since we have show that $K_{\ell,\ell'} \geq 1$ and $K_\ell \geq 1/2$, this concludes our proof that

$$K(\pi_M, \pi_{\widehat{M}}) \geq L/2 + \binom{L}{2} = m^2/18 + \frac{m^2/9(m^2/9-1)}{2} = m^4/162,$$

showing that when $d = 2$, Proposition B.5 holds.

Now we will use the M constructed in the proof for the case when $d = 2$ to create a more general counterexample for the case when $d > 2$.

In the proof of Proposition B.1, we pick T^* to be constant with respect to the last $d-2$ indices. Hence, $T^*_{\cdot,i_{3:d}} = M$ where M is chosen to be a block matrix of the 3×3 matrix B whose relative ordinal relations cannot be satisfied by a generalized additive model. Let $A_{ij} \in [0,1]^{m/3 \times m/3}$ be the constant matrix taking on the value of B_{ij} for $(i,j) \in [3]^2$. Further, let $\otimes$ denote the tensor product. We constructed

$$T^* = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \otimes 1^{\otimes^{d-2}}.$$

This worked, because fixing $d-2$ of the last entries, we had $\Omega(m^2)$ Frobenius norm error and, summing over all such pairs, we had $\Omega(m^d)$ error. However, we cannot use this same approach for ranking error. If we summed the Kendall tau distance separately over each pair $(i_3, \ldots, i_d) \in [m]^{d-2}$, since the error for each pair is $\mathcal{O}(m^4)$, we would have total error of $\mathcal{O}(m^{d+2})$ which for $d > 2$ is asymptotically less than m^{2d} (the order of the Kendall tau distance).

Therefore, we instead need to compare across different values of $(i_3, \ldots, i_d) \in [m]^{d-2}$. To do this, and to ensure that we know the unique ground truth ranking of T^* we construct a special $m/3 \times m/3 \times m^{d-2}$ tensor of perturbations, E, to add to each of the nine $m/3 \times m/3$ blocks. Thus, now we can write

$$T^* = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \otimes 1^{\otimes^{d-2}} + \begin{pmatrix} E & E & E \\ E & E & E \\ E & E & E \end{pmatrix}.$$

To construct E, we will use the following lemma:

Lemma B.7. There exists a tensor $\widetilde{E} \in [0,1]^{m^d}$ that is isotonic and has unique entries.

Taking Lemma B.7 as given for now, let E be a normalized $\widetilde{E}$ such that it has values between $[0, 0.1)$ and contains only the first $m/3$ values of i_1 and i_2. In particular, for all $\widetilde{i} \in [m/3]^2 \times [m]^{d-2}$:

$$E_{\widetilde{i}} = \frac{1}{11} \widetilde{E}_{\widetilde{i}}.$$

Now, we will re-write this constructed T^* in an equivalent way which will help us in the proof. To do this, we will define a function g to map entries to the correct block matrix, and a function h to map entries to the component within the block matrix. In particular, we set $g(i) = \lceil 3i/n \rceil$ to partition the (i_1, i_2) pairs into the respective block (i.e. value of $B_{g(i_1),g(i_2)}$) and $h(i) = i - n/3(\lceil 3i/n \rceil - 1)$ to map the an index i to the associated entry within the block (i.e., perturbation value $E_{h(i_1),h(i_2),i_3,\ldots,i_d}$). For all $i \in [m]^d$:

$$T^*_{i_1,i_2,\ldots,i_d} = B_{g(i_1),g(i_2)} + E_{h(i_1),h(i_2),i_3,\ldots,i_d}.$$

We will confirm that T^* is in $\mathbb{C}^d_{\texttt{ISO}}$. First, it must have entries in $[0,1]$. B and E are non-negative so $T^* \geq 0$. Since B has entries between 0.1 and 0.9 and E has entries in $[0, 0.1)$:

$$\max_{i \in [m]^d} T^*_i \leq \max_{i \in [m]^d} B_{g(i_1),g(i_2)} + \max_{i \in [m]^d} E_{h(i_1),h(i_2),i_3,\ldots,i_d} < 0.9 + 0.1 = 1.$$

Second, T^* must be isotonic. Consider a non-equal set of indices $i = i_1, i_2, \ldots, i_d$ and $i' = i'_1, i'_2, \ldots, i'_d$ such that $i' \geq i$. If $g(i_1) = g(i'_1), g(i_2) = g(i'_2)$, then

$$T^*_{i'} - T^*_i = E_{h(i'_1),h(i'_2),i'_3,\ldots,i'_d} - E_{h(i_1),h(i_2),i_3,\ldots,i_d}.$$

Since E is isotonic, and h is increasing in i_1, i_2 given $g(i_1)$ and $g(i_2)$, this is strictly positive. The second case is that $g(i'_1) \neq g(i_1)$ and/or $g(i'_2) \neq g(i_2)$. Since g is weakly increasing, it must be the case that $g(i'_1) > g(i_1)$ or $g(i'_2) > g(i_2)$. Thus i is mapped to different entry of B which, by construction is at least 0.1 larger. Using the fact that entries in E are less than 0.1 apart:

$$T^*_{i'} - T^*_i > (B_{g(i'_1),g(i'_2)} - B_{g(i_1),g(i_2)}) - 0.1 = 0.$$

Thus, $T^* \in \mathbb{C}^d_{\texttt{ISO}}$ as desired.

Now, we want to lower-bound the Kendall tau distance between T^* and its approximation $T \in \mathbb{C}^d_{\texttt{GENLIN}}(F)$ for some non-decreasing F. Consider the partition of T^* based on the value of E an index maps to (ie the values of $h(i_1), h(i_2), i_3, \ldots, i_d$). There are $L = \frac{m^d}{9}$ total partitions. We will index these by $\ell \in [L]$ and use S_ℓ to refer to the set of indices that partition ℓ contains and $T_{S_\ell} \in \mathbb{R}^{3\times 3}$ to refer to the sub-matrix formed by just those elements. Note that

$$\begin{aligned} K(\pi_{T^*}, \pi_T) &= \sum_{\text{unique } i,i' \in [m]^d} \overline{K}(i, i', \pi_{T^*}, \pi_T) \\ &= \sum_{\ell \in [L]} \sum_{i \neq i' \in S_\ell} \overline{K}(i, i', \pi_{T^*}, \pi_T) + \sum_{\ell \neq \ell' \in [L]} \sum_{i \neq S_\ell, i' \in S_{\ell'}} \overline{K}(i, i', \pi_{T^*}, \pi_T) \end{aligned}$$

Let $K_\ell = \sum_{i \neq i' \in S_\ell} \overline{K}(i, i', \pi_{T^*}, \pi_T)$ and $K_{\ell,\ell'} = \sum_{i \neq S_\ell, i' \in S_{\ell'}} \overline{K}(i, i', \pi_{T^*}, \pi_T)$. Note that if for all $\ell, \ell' \neq \ell \in [L]$, $K_\ell \geq 1$ and $K_{\ell,\ell'} \geq 1$ then we would have

$$K(\pi_{T^*}, \pi_T) \geq L + \binom{L}{2} \geq L^2/2 = m^{2d}/162.$$

Thus, it suffices to show that $K_\ell \geq 1$ and $K_{\ell,\ell'} \geq 1$. There are two cases.

Case 1: No ties and an incorrect ordering. First consider what would happen for K_ℓ. This says that there are two indices $i \neq i' \in S_\ell$ such that the relative order in T^* is different from in T this implies that $\overline{K}(i, i', \pi_{T^*}, \pi_T) = 1$ and hence that $K_\ell \geq 1/2$. The same argument applies for $K_{\ell,\ell'}$: we have some $i \in S_\ell$ and $i' \in S_{\ell'}$ where the relative order in T^* is different from in T which implies $\overline{K}(i, i', \pi_{T^*}, \pi_T) = 1$ and hence that $K_{\ell,\ell'} \geq 1$. Now we are done with case 1.

Case 2: Same ordering up to ties. First consider K_ℓ. Let E_ℓ be the value of E that this partition corresponds to. Observe that $T^*_{S_\ell} = B + E_\ell$ by construction so that the ranking of entries in $T^*_{S_\ell}$ is the same as the ranking over entries in B. This implies that T_{S_ℓ} has the same ranking as B up to ties and consequently satisfies that relations in Lemma B.6 with weak inequalities.

Note that all indices $i \in S_\ell$ have the same values of $i_3, \ldots, i_d$. Thus T_{S_ℓ} is a sub-matrix of $T_{\cdot i_{3:d}}$ for some values $i_3, \ldots, i_d$ associated with the rows $\{h(i_1), m/3 + h(i_1), 2m/3 + h(i_1)\}$ and columns $\{h(i_2), m/3 + h(i_2), 2m/3 + h(i_2)\}$. Since sub-matrices of matrices in $\mathbb{C}^2_{\texttt{GENLIN}}$ are also in $\mathbb{C}^2_{\texttt{GENLIN}}$, $T_{S_\ell} \in \mathbb{C}^2_{\texttt{GENLIN}}$ if $T_{\cdot i_{3:d}} \in \mathbb{C}^2_{\texttt{GENLIN}}$. We will now show that this is true. Because $T \in \mathbb{C}^d_{\texttt{GENLIN}}(F)$, there exists some $a^{(1)}, \ldots, a^{(d)}$ such that for all $i_1, i_2 \in [d]^2$

$$T_{i_1 i_2 \ldots i_d} = F(\sum_{k \in [d]} a^{(k)}_{i_k} + c) = F(a^{(1)}_{i_1} + a^{(2)}_{i_2} + \sum_{k=3}^{d} a^{(k)}_{i_k} + c).$$

Setting $b \in \mathbb{R}^n$ to have entries $a^{(2)}_{i_2} + \sum_{k=3}^d a^{(k)}_{i_k} + c$ for all $i_2 \in [m]$ and $a^{(1)} = a^{(1)}$ we see that $T_{\cdot i_{3:d}} \in \mathbb{C}^2_{\texttt{GENLIN}}$ for some the same F, $a^{(1)}$ and $a^{(2)} = b$. As our argument does not depend on the value of ℓ, $T_{S_\ell} \in \mathbb{C}^2_{\texttt{GENLIN}}$ for all $\ell \in [L]$. As none of these entries are equal in T_{S_ℓ}, and at least two must be tied to not contradict Lemma B.2, it follows that $K_\ell \geq 1/2$.

Now, consider $K_{\ell,\ell'}$. Note that $T^*_{S_\ell} = B + E_\ell$ and $T^*_{S_{\ell'}} = B + E_{\ell'}$. Since the entries of E are unique, it follows that either $T^*_{S_\ell} > T^*_{S_{\ell'}}$ or $T^*_{S_\ell} < T^*_{S_{\ell'}}$. WLOG we will assume that $T^*_{S_\ell} > T^*_{S_{\ell'}}$. If all relations between entries in S_ℓ and $S_{\ell'}$ are satisfied up to ties, then:

$$T_{S_{\ell'},12} \leq T_{S_\ell,12} \leq T_{S_{\ell'},21} \leq T_{S_\ell,21} \leq T_{S_{\ell'},31} \leq T_{S_\ell,31} \leq T_{S_{\ell'},22} \leq T_{S_\ell,22}$$

and

$$T_{S_{\ell'},22} \leq T_{S_\ell,22} \leq T_{S_{\ell'},13} \leq T_{S_\ell,13} \leq T_{S_{\ell'},23} \leq T_{S_\ell,23} \leq T_{S_{\ell'},32} \leq T_{S_\ell,32}.$$

This implies that $T_{S_{\ell'}}, T_{S_\ell} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ satisfy the ordinal relations in Lemma B.6 with weak inequalities. Hence we know that at least one of the sub-blocks must have tied items, leading to two additional inter-block ties and $K_{\ell,\ell'} \geq 1$. This concludes the 2nd case. Since we have shown that for all $\ell, \ell' \neq \ell \in [L]$, $K_\ell \geq 1$ and $K_{\ell,\ell'} \geq 1$, we have

$$K(\pi_{T^*}, \pi_T) \geq L + \binom{L}{2} \geq L^2/2 = m^{2d}/162.$$

Since $\binom{|\mathcal{X}|}{2}$ is order m^{2d} we have shown that the statement of Proposition B.5 holds when $d = 2$ and $d > 2$, we have now completed the proof of Proposition B.5. All that is left is to prove the auxiliary lemmas.

Proof of Lemma B.6. Assume for the sake of contradiction that $\widehat{M} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ for some F, a, b, c. Without loss of generality, we can assume that $c = 0$. This is because we can set $\widetilde{b} = b_j + c$ such that for all $(i, j) \in [m]^2$, $f(a_i + \widetilde{b}_j) = f(a_i + b_j + c)$. Therefore, $\widehat{M} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ for some F, a, b, c implies that $\widehat{M} \in \mathbb{C}_{\texttt{GENLIN}}(F)$ for some F, a, b.

Since $\widehat{M}_{ij} = F(a_i + b_j)$, the first inequality implies:

$$F(a_1 + b_2) < F(a_2 + b_1) < F(a_3 + b_1) < F(a_2 + b_2).$$

Since F is non-decreasing, the ordering of the entries is the same as the ordering of the underlying linear model. Therefore, the inequalities in terms of the outputs of F hold for the inputs. Hence,

$$a_1 + b_2 < a_2 + b_1 < a_3 + b_1 < a_2 + b_2.$$

Lastly, since the difference between the first and last terms ($a_1 + b_2$ and $a_2 + b_2$) is greater than the difference between the middle two terms ($a_2 + b_1$ and $a_3 + b_1$) we have that

$$a_2 - a_1 > a_3 - a_2.$$

We will now make the same argument for the second inequality. The second inequality implies that

$$F(a_2 + b_2) < F(a_1 + b_3) < F(a_2 + b_3) < F(a_3 + b_2).$$

Following the same argument, we get that

$$a_2 - a_1 < a_3 - a_2$$

which is a contradiction. Therefore, both inequalities cannot hold at the same time. □

Proof of Lemma B.7. First, we will construct a tensor that has unique values. For all $i \in [m]^d$, let

$$T_i = \sum_{k=1}^{d} m^{k-1} i_k = i_1 + m i_2 + \cdots + m^{d-1} i_d$$

Since all indices i_k are between 1 and m, each entry of T can be represented by a unique value (base m) and thus is unique. Since the value of T_i is increasing in the value of each index, T is isotonic. Further, we have that for all i, $T_i \leq dm^d$. Dividing each entry by dm^d, we have the desired claim. □

□

B.3 Measuring criteria importance

Lastly, researchers often use regression coefficients on the criteria output by linear models to understand the relative importance of each criterion. This is useful information that can convey what aspects of an item or individual matter more, i.e., the contribution or presentation of a paper, or the level of service or location for a hotel. Thus, we ask: *how can assuming a linear model lead to issues when measuring relative criteria importance?*

To illustrate potential issues, consider a (simplified) setting of Lane et al. (2023). The authors are interested in understanding how evaluators "weigh" two different evaluation criteria —feasibility and novelty—and how these weights are dependent upon the type of expertise they have. For simplicity, we assume that each criterion is binary and that the researchers assume a linear model of the form $f(x_1, x_2) = a_1x_2 + a_2x_2$. Before discussing potential issues arising from model mismatch, we will introduce the key theoretical result we rely on in our analysis.

Proposition B.8. Let $d = m = 2$. Assume that we have access to data $\{x^{(j)}, y^{(j)}\}_{j\in[N]}$ where the $y^{(j)}$ are noiseless observations of the true preference function $f \in \mathcal{F}_{\texttt{ISO}}$. That is, for all $j \in [N]$, $y^{(j)} = f(x^{(j)})$. Let $a_1x_1 + a_2x_2$ be the OLS estimator fit on the data. There exist $f \in \mathcal{F}_{\texttt{ISO}}$ and sets of sample inputs $\{x^{(j)}\}_{j\in[N]}$ such that

(a) $a_1 < a_2$ when $f(x_1, x_2) > f(x_2, x_1)$ for all $x_1 > x_2$.
(b) $a_1 < a_2$ when $f(x_1, x_2) = f(x_2, x_1)$ for all $(x_1, x_2) \in \mathcal{X}$.

At a high level, part (a) implies that one can incorrectly assume criterion 1 is weighted more than criterion 2 when the reverse is true. Part (b) implies that one can find that one criterion is weighted above the other even when the underlying data-generating process is fully symmetric. We will prove Proposition B.8 at the end of this section. The key issue, however, arises from an imbalance in samples of different possible sets of criteria evaluations.

In Lane et al. (2023), as is common in many applications, the authors use the values of the regression coefficients, a_1, a_2 to infer the relative worth or weight of both criteria. They consider two groups of evaluators: multivalent evaluators, those who have expertise in both relevant technical domains, and univalent evaluators, those with expertise in only one of them. The authors find that univalent evaluators overweigh novelty more so than multivalent ones. By part (a) of Proposition B.8, it is possible that the researchers found that novelty (criteria 2) matters more to univalent evaluators, when in fact it matters less. By part (b), they may equally have concluded that feasibility (criteria) matters more to multivalent evaluators when they are fully indifferent. Consequently, the true story may actually that multivalent evaluators care *more* about novelty than univalent evaluators, not less.

B.3.1 Proof of Proposition B.8

Our proof relies on the following result.

Lemma B.9. Consider the setting of a function $f : \mathcal{X} \to \mathbb{R}$ fit by a linear model of the form, $f(x_1, x_2) = a_1x_1 + a_2x_2$ for some $(a_1, a_2) \in \mathbb{R}^2$. Given samples $\{x^{(j)}\}_{\ell=1}^N$ consider the least squares minimization problem:

$$\min_{(a_1,a_2)} \sum_{j=1}^{N} (f(x^{(j)}) - a_1x_1^{(j)} - a_2x_2^{(j)})^2.$$

Let N_{x_1,x_2} for $x \in \{0, 1\}^2$ be the number of observations corresponding to each input pair. Let $\alpha_{x_1,x_2} = f(x)$ the function value for those inputs. If at least two of $N_{1,1}, N_{0,1}, N_{1,0}$ are strictly positive, then the solutions are:

$$a_1 = \frac{(N_{0,1} + N_{1,1})(N_{1,0}\alpha_{1,0} + N_{1,1}\alpha_{1,1}) - N_{1,1}(N_{0,1}\alpha_{0,1} + N_{1,1}\alpha_{1,1})}{N_{1,1}(N_{0,1} + N_{1,0}) + N_{1,0}N_{0,1}}$$
$$a_2 = \frac{-N_{1,1}(N_{1,0}\alpha_{1,0} + N_{1,1}\alpha_{1,1}) + (N_{1,1} + N_{1,0})(N_{0,1}\alpha_{0,1} + N_{1,1}\alpha_{1,1})}{N_{1,1}(N_{0,1} + N_{1,0}) + N_{1,0}N_{0,1}}$$

Let us take this lemma as given for now (we will prove it at the end of this section).

Proof of part (a) : Consider the case when input pairs for all but $(0, 1)$ are observed. Set $\alpha_{1,0} = \frac{1}{4}$, $\alpha_{0,1} = 0$, and $\alpha_{1,1} = 1$.The denominator of the expressions for a_1 and a_2 are equal and strictly positive. Thus, it suffices to show that the difference of the numerators is negative. Then taking the difference of the numerators we get that

$$a_1 < a_2 \iff N_{1,0}N_{1,1}(2\alpha_{1,0} - \alpha_{1,1}) < 0.$$

Plugging in $\alpha_{1,0} = \frac{1}{4}$ and $\alpha_{1,1} = 1$, one can see that this inequality is satisfied.

Proof of part (b) : Assume that there is at least one sample of each criteria pair $x \in \{0,1\}^2$. Consider the case when $\alpha_{0,1} = \alpha_{1,0} = \alpha \in [0, \frac{1}{2})$, and $N_{1,0} > N_{0,1}$. Additionally, assume that $\alpha_{0,0} = 0$ and $\alpha_{1,1} = 1$. As for the proof of part (a), the denominator of the expressions for a_1 and a_2 are equal and strictly positive, so it suffices to show that the difference of the numerators is negative. Substituting in the values for $\alpha_{1,1}$ we get that

$$a_1 < a_2 \iff N_{1,1}(1-2\alpha)(N_{0,1} - N_{1,0}) < 0.$$

Since $\alpha < \frac{1}{2}$, $(1-2\alpha) > 0$. As $N_{1,0} > N_{0,1}$, $(N_{0,1} - N_{1,0}) < 0$. Hence the product is negative and $a_1 < a_2$ as claimed.

All that is left is to prove Lemma B.9.

Proof of Lemma B.9. Let $R_{\texttt{OLS}}(a_1, a_2)$ be the total squared error for any a_1, a_2 pair. Plugging in the values of f for the specific inputs we have:

$$\begin{aligned} R_{\texttt{OLS}}(a_1, a_2) &= \sum_{j \in [N]} \left(f(x_1^{(j)}, x_2^{(j)}) - a_1 x_1^{(j)} - a_2 x_2^{(j)} \right)^2 \\ &= N_{0,0}\alpha_{0,0}^2 + N_{1,0}\left(\alpha_{1,0} - a_1\right)^2 + N_{0,1}\left(\alpha_{0,1} - a_2\right)^2 + N_{1,1}\left(\alpha_{1,1} - a_1 - a_2\right)^2 \end{aligned}$$

The partial derivatives of $R_{\texttt{OLS}}$ with respect to each choice variable are

$$\begin{aligned} \frac{\partial R_{\texttt{OLS}}}{\partial a_1} &= -2N_{1,0}(\alpha_{1,0} - a_1) - 2N_{1,1}(\alpha_{1,1} - a_1 - a_2) \\ \frac{\partial R_{\texttt{OLS}}}{\partial a_2} &= -2N_{0,1}(\alpha_{0,1} - a_2) - 2N_{1,1}(\alpha_{1,1} - a_1 - a_2) \end{aligned}$$

Since $R_{\texttt{OLS}}$ is convex, it is minimized when both partials are zero. With some algebraic manipulation, we have that this is equivalent to the following two systems of equations:

$$\begin{aligned} (N_{1,0} + N_{1,1})a_1 + N_{1,1}a_2 &= N_{1,0}\alpha_{1,0} + N_{1,1}\alpha_{1,1} \\ N_{1,1}a_1 + (N_{0,1} + N_{1,1})a_2 &= N_{0,1}\alpha_{0,1} + N_{1,1}\alpha_{1,1} \end{aligned}$$

In matrix form, we can write this as

$$\begin{pmatrix} N_{1,0} + N_{1,1} & N_{1,1} \\ N_{1,1} & N_{1,1} + N_{0,1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} N_{1,0}\alpha_{1,0} + N_{1,1}\alpha_{1,1} \\ N_{0,1}\alpha_{0,1} + N_{1,1}\alpha_{1,1} \end{pmatrix}.$$

Using the formula for the inverse of a 2×2 matrix,

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} N_{1,0} + N_{1,1} & N_{1,1} \\ N_{1,1} & N_{1,1} + N_{0,1} \end{pmatrix}^{-1} \begin{pmatrix} N_{1,0}\alpha_{1,0} + N_{1,1}\alpha_{1,1} \\ N_{0,1}\alpha_{0,1} + N_{1,1}\alpha_{1,1} \end{pmatrix}.$$

Let $\widetilde{N} = N_{1,1}(N_{0,1}+N_{1,0})+N_{1,0}N_{0,1}$. Note that the matrix inverse is well-defined as its determinant is

$$N_{1,1}(N_{0,1} + N_{1,0}) + N_{1,0}N_{0,1} > \min\{N_{1,1}N_{0,1}, N_{1,1}N_{1,0}, N_{1,0}N_{0,1}\} > 0$$

by our assumption that at least two of $N_{1,1}, N_{0,1}, N_{1,0}$ are strictly positive. With a bit of algebra, we get that the estimated coefficients are

$$\begin{aligned} a_1 &= \frac{(N_{0,1} + N_{1,1})(N_{1,0}\alpha_{1,0} + N_{1,1}\alpha_{1,1}) - N_{1,1}(N_{0,1}\alpha_{0,1} + N_{1,1}\alpha_{1,1})}{\widetilde{N}} \\ a_2 &= \frac{-N_{1,1}(N_{1,0}\alpha_{1,0} + N_{1,1}\alpha_{1,1}) + (N_{1,1} + N_{1,0})(N_{0,1}\alpha_{0,1} + N_{1,1}\alpha_{1,1})}{\widetilde{N}}. \end{aligned}$$

□

C Supplementary material for Section 5

C.1 Proof of Theorem 5.1

We will prove (a) and (b) of Theorem 5.1 separately.

C.1.1 Proof of Theorem 5.1 (a)

Our goal to showing that there exists a universal constant $C > 0$ such that for any distribution P over $\mathcal{X}$

$$\sup_{f \in \mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{CV}}, f) \leq C \log |\Lambda| \sup_{f \in \mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{LIN}}, f).$$

We will do this by proving a $\Omega(1/N)$ lower bound on the risk of $f_{\texttt{LIN}}$, and show that, up to a constant factor, the error of $f_{\texttt{CV}}$ exceeds that of $f_{\texttt{LIN}}$ by an additive term of $O(\log |\Lambda|/N)$.

First, we will focus on the lower bound. It is well known that the risk of linear regression is $\Omega(1/N)$. However, for completeness, we will include a short proof. Define $\mathcal{F}_{\texttt{CONST}}$ to be the set of constant functions mapping $\mathcal{X}$ to $[0, 1]$. Since $\mathcal{F}_{\texttt{CONST}} \subset \mathcal{F}_{\texttt{LIN}}$ it follows that

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{\texttt{CONST}}} R(\widehat{f}, f) \leq \inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{\texttt{LIN}}} R(\widehat{f}, f).$$

Since it can be shown that $\mathcal{F}_{\texttt{CONST}}$ has a $\Omega(1/N)$ lower-bound, it suffices to prove a lower bound for $\mathcal{F}_{\texttt{CONST}}$.

Lower bound for learning constant functions. Let P_f be the probability distribution over (x, y) induced by a preference function f and $d_{\texttt{TV}}(P, Q)$ be the total variation distance between two distributions P and Q. By Assouad's lemma (e.g., (Wu, 2020) Theorem 10.2)

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{\texttt{CONST}}} R(\widehat{f}, f) \geq \sup_{f_0, f_1 \in \mathcal{F}_{\texttt{CONST}}} \frac{R(f_0, f_1)}{4} (1 - d_{\texttt{TV}}(P_{f_0}, P_{f_1})).$$

Let $f_0 = 0$ and $f_1 = \frac{1}{\sqrt{N}}$. Then $R(f_0, f_1) = \frac{1}{N}$. Further, by the affine invariance of total variation distance, and the fact that the sample mean is a sufficient statistic for normal random variables when the variance is known

$$d_{\texttt{TV}}\left(P_{f_0}, P_{f_1}\right) = d_{\texttt{TV}}\left(N(0, \frac{1}{N}), N\left(\frac{1}{\sqrt{N}}, \frac{1}{N}\right)\right) = d_{\texttt{TV}}\left(N(1,1), N(0,1)\right).$$

Note that $(1 - d_{\texttt{TV}}\left(N(1,1), N(0,1)\right)) \approx 0.7 \geq \frac{1}{2}$. This implies

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{\texttt{LIN}}} R(\widehat{f}, f) \geq \frac{1}{8N}.$$

Upper bound for our algorithm. To prove an upper bound for our algorithm, we will show that the risk of $f_{\texttt{CV}}$ can be bounded by the risk of f_∞. Then we will show that the risk of f_∞ is weakly less than that of non-negative least squares. To bound $R(f_{\texttt{CV}}, f)$ in terms of $R(f_\infty, f)$ we will use the following result:

Corollary C.1 (Corollary 4.2 of Vaart et al. (2006))**.** Consider the setting where we have a set of functions $f \in \mathcal{F}$ bounded by a constant $M \geq 1$ and are given access to data $(x^{(j)}, y^{(j)})_{j \in [N]}$ of independent and identically distributed samples where for all $j \in [N]$

$$y^{(j)} = f(x^{(j)}) + w^{(j)}$$

with $w^{(j)}$ independent sub-exponential noise with $\mathbb{E}[w^{(j)} | x^{(j)}] = 0$. Assume we divide the data into two random subsets: a dataset S_1 of size at least αN for validation and a dataset S_0 for training, where $\alpha > 0$. Suppose we fit a set of estimators $f_1, \ldots, f_K$ on data in S_0, calculate the mean squared error of each estimator on S_1, $\sum_{i \in S_1} (f_k(x_i) - y_i)^2$ for $k \in [K]$, and return $f_{\widetilde{k}}$ that minimizes the error on S_1. Then, there exists some constant $C_1 > 0$ such that for any $\delta \in (0, 1)$:

$$R(f_{\widetilde{k}}, f_0) \leq (1 + 2\delta) \inf_{k \in [K]} R(f_k, f_0) + C_1/N \log(1 + K) \frac{M^2}{\delta}.$$

This theorem gives us a bound on a cross-validation estimator that takes as input a set of K functions fitted by regression on the data in terms of the risk of the best method in hindsight. A brief remark is that in the Corollary presented in Vaart et al. (2006), they don't mention the dependencies of C_1. However, these dependencies can be attained by referring to Theorem 2.3 and Lemma 4.1 of the same paper from which the Corollary is derived (the proof is trivial). The constant factor depends only on the distribution of the noise term and linearly on $\frac{1}{\alpha}$, neither of which is relevant to our asymptotic analysis since they are fixed by our algorithm and the assumption that the noise is $\mathcal{N}(0,1)$.

In order to apply this result, we need to check that our cross-validation setting is equivalent to the setting of Corollary C.1 for a specific choice of α, K, M. Recall that we observe outputs from an unknown function $f \in \mathcal{F}_{\texttt{ISO}}$ with additive noise that is sub-exponential and with conditional mean zero (in our case, it is specifically $N(0,1)$) and impose that $\|f\|_\infty \leq 1$ (so we can set $M = 1$). Further, we test $|\Lambda|$ potential regression functions on a $\frac{1}{4}$ fraction of the data, satisfying the condition that the size of the validation set is at least a constant fraction of the total data ($\alpha = \frac{1}{4} > 0$). Lastly, we return $f_{\texttt{CV}}$ which minimizes the loss of $\{\widehat{f}_\lambda\}_{\lambda\in\Lambda}$ on the validation dataset. Thus, the setting of Theorem 5.1 satisfies the conditions of Corollary C.1 with $K = |\Lambda|$, $M = 1$ and $\alpha = \frac{1}{4}$. Letting $\delta = \frac{1}{2}$ and substituting in for the values of K, M and α the result of Corollary C.1 implies that

$$R(f_{\texttt{CV}}, f) \leq 2 \inf_{\lambda\in\Lambda} R(f_\lambda, f) + C_1/N \log(1 + |\Lambda|) \tag{17}$$

As $\infty \in \Lambda$, it follows that

$$\inf_{\lambda\in\Lambda} R(f_\lambda, f) \leq R(f_\infty, f).$$

When $\lambda = \infty$, we impose linearity on the regression function (prior to truncation). As the true f takes on values in $[0,1]$, $R(f_\infty, f) \leq R(f_{\texttt{LIN}}, f)$. This implies that

$$R(f_{\texttt{CV}}, f) \leq 2R(f_{\texttt{LIN}}, f) + C_1/N \log(1 + |\Lambda|).$$

To conclude, we simply want to show that for all $f \in \mathcal{F}_{\texttt{LIN}}$ that for some universal $C > 0$ that

$$2R(f_{\texttt{LIN}}, f) + C_1/N \log(1 + |\Lambda|) \leq C \log(|\Lambda|) \sup_{f\in\mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{LIN}}, f).$$

Observe that $2R(f_{\texttt{LIN}}, f) \leq \frac{2}{\log 2} \log(|\Lambda|) \sup_{f\in\mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{LIN}}, f)$ as by construction $|\Lambda| \geq 2$. Now, let us bound the second term in this expression. We know that $\log(1+|\Lambda|) \leq 2\log(|\Lambda|)$ as $1+|\Lambda| \leq |\Lambda|^2$. Furthermore,

$$C_1/N \leq 8C_1 \left(\frac{1}{8N}\right) \leq 8C_1 \sup_{f\in\mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{LIN}}, f).$$

As such,

$$C_1/N \log(1 + |\Lambda|) \leq 16C_1 \log(|\Lambda|) \sup_{f\in\mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{LIN}}, f).$$

Setting $C = 16C_1 + \frac{2}{\log 2}$ thus gives the desired result:

$$\sup_{f\in\mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{CV}}, f) \leq C \log(|\Lambda|) \sup_{f\in\mathcal{F}_{\texttt{LIN}}} R(f_{\texttt{LIN}}, f).$$

This concludes our proof of Theorem 5.1

C.1.2 Proof of Theorem 5.1 (b)

Recall that our goal is to prove that

$$\sup_{f\in\mathcal{F}_{\texttt{ISO}}} R(f_{\texttt{CV}}, f) \leq C_{d,m,p_{\min}} \log(|\Lambda|) N^{-1/d} \log^{\gamma_d} N.$$

We will do this by combining the oracle inequality from Equation (17) with an upper bound on $R(f_0, f)$. Since we have already proved the oracle inequality in the proof of part (a) of this theorem, we will proceed to prove an upper bound on $R(f_0, f)$. We will do this by slightly modifying a result from Han et al. (2019) on the risk of isotonic functions mapping $[0,1]^d \to [-1,1]$.

Theorem C.2 (Theorem 4 of Han et al. (2019))**.** Consider the setting of Theorem 5.1 where instead of the preference function being in $\mathcal{F}_{\texttt{ISO}}$, it belongs to the set of isotonic functions from $\widetilde{\mathcal{X}} = [0,1]^d$

to $[-1,1]$ (which we denote by the set $\widetilde{\mathcal{F}_{\texttt{ISO}}}$). Let the distribution over $\widetilde{\mathcal{X}}$ be denoted by $\widetilde{P}$. If $\widetilde{P}$ satisfies

$$0 < m_0 \leq \inf_{x \in \mathcal{X}} \widetilde{P}(x) \leq \sup_{x \in \widetilde{\mathcal{X}}} \widetilde{P}(x) \leq M_0 \leq \infty,$$

then there exists an estimator $\widehat{f}$ such that for $d \geq 2$, these exists a C_{d,m_0,M_0} depending only on d, m_0, M_0 such that

$$\sup_{\widetilde{f} \in \widetilde{\mathcal{F}_{\texttt{ISO}}}} R(\widehat{f}, f) \leq C_{d,m_0,M_0} N^{-1/d} \log^{\gamma_d} N.$$

Our goal is to show that, given a different set of preference functions and our algorithm (which is slightly different from theirs), the same upper bound still holds. In Han et al. (2019), they pick estimator $\widehat{f}$ as the function in $\widetilde{\mathcal{F}_{\texttt{ISO}}}$ minimizing squared error on the sample data. With a small amount of algebra, one can confirm that this is equivalent to our approach of finding the minimizing function with outputs in $\mathbb{R}$ and truncating the outputs to the relevant range (in their case $[-1,1]$). For values not seen during training, they interpolate using.

$$\widehat{f}(x) := \min\left(\{\widehat{f}(x^{(j)}) : 1 \leq i \leq d, x_i^{(j)} \geq x_i\} \cup \{\max_{j \in [N]} \widehat{f}(x^{(j)})\}\right).$$

This is equivalent to picking the minimum of the smallest estimated functional value of a point seen during training that dominates the test point, and the maximum of any functional value seen in training.

In order to transfer the results of Theorem C.2 to our setting, we have to ensure that (extensions of) their arguments apply given differences in the domain, co-domain, and interpolation method. We will show later that the upper bounds apply, given most reasonable interpolation methods, including ours. Further, since we do an ex-post truncation to the true output space $[0,1]$ instead of $[-1,1]$, this step will weakly lower the risk of the estimator. Therefore, we will focus on adapting their arguments to apply to distributions on $\mathcal{X}$ satisfying the conditions of Theorem 5.1 (b).

To apply their arguments to distributions on $\mathcal{X}$, we will use the fact that the learning problem of isotonic functions defined on $\mathcal{X}$ is equivalent to the learning problem when the function is defined only on the lattice,

$$\mathbb{L}_{d,m} := \prod_{i \in [d]} \{1/m, 2/m, \ldots, 1\}.$$

To bound the risk $R(\widehat{f}, \widetilde{f})$, the authors define the event that the infinity norm of the function error exceeds some threshold and consider the risk when this event occurs and when it does not. Specifically, for some value of $a > 0$, they define $A = \{\|\widehat{f} - f\|_\infty \leq a\}$ and indicators $\mathbb{I}_A$ and $\mathbb{I}_{A^C}$ to write the following decomposition of risk

$$R(\widehat{f}, \widetilde{f}) = \mathbb{E}[\|\widehat{f} - \widetilde{f}\|_{L_2(\widetilde{P})} \mathbb{I}_A] + \mathbb{E}[\|\widehat{f} - \widetilde{f}\|_{L_2(\widetilde{P})} \mathbb{I}_{A^C}].$$

They prove upper bounds on each of these terms separately. We will show that, in both cases, the upper bounds still apply when the marginal distribution is defined on the lattice for a suitable choice of m_0, M_0. First, consider the risk given bounded error, $\mathbb{E}[\|\widehat{f} - \widetilde{f}\|_{L_2(\widetilde{P})} \mathbb{I}_A]$. The authors use arguments not depending on $\widetilde{P}$ to show that if certain upper bounds on certain complexity measures for subsets of isotonic functions on $\widetilde{\mathcal{X}}$ hold, then when $a = 6 \log^{\frac{1}{2}} N$ one can upper-bound the value of $\mathbb{E}[\|\widehat{f} - \widetilde{f}\|_{L_2(\widetilde{P})} \mathbb{I}_A]$ by $C_{m_0,M_0} N^{-1/d} \log^{\gamma_d} N$. While in the key result bounding the risk conditional on A (Proposition 7), depend on a Rademacher and Gaussian complexity term, they show via further analysis (specifically Lemma 5 and 6, neither of which depending on properties of $\widetilde{P}$) that it suffices to bound the Rademacher complexity of the set of isotonic functions in the $L_2(\widetilde{P})$ ball around the zero function with outputs in $[-1,1]$.

Now we will more specifically discuss the complexity measure that they use. Let $\xi^{(1)}, \ldots, \xi^{(N)}$ be a sequence of random signs. Define $B_2(r, \widetilde{P})$ to be the ball of radius r in the $L_2(\widetilde{P})$ norm. Let $\mathcal{G}_{\widetilde{x}}(0, r, 1) = \{\widetilde{f} \in \widetilde{\mathcal{F}_{\texttt{ISO}}} : \widetilde{f} \in B_2(r, \widetilde{P}) \text{ and } \|\widetilde{f}\|_\infty \leq 1\}$ be the set of isotonic functions over $\widetilde{\mathcal{X}}$

in the $L_2(\widetilde{P})$ ball of radius r around the zero function and without outputs in $[-1, 1]$. The authors bound the normalized Rademacher complexity of this class of functions which can be written as

$$\mathbb{E}\left[\sup_{\widetilde{f}\in\mathcal{G}_{\widetilde{x}}(0,r,1)}\left|\frac{1}{N^{1/2}}\sum_{j=1}^{N}\xi^{(j)}\widetilde{f}(\widetilde{x}^{(j)})\right|\right].$$

We can define the respective quantities for the probability measures restricted to the lattice (or equivalently over $\mathcal{X}$) in terms by $\mathcal{G}_x(0, r, 1) = \{f \in \mathcal{F}_{\texttt{ISO}}(\mathbb{R}) : f \in B_2(r, P) \text{ and } \|f\|_\infty \leq 1\}$ for $B_2(r, P)$ the $L_2(P)$ ball around the zero function. The relevant complexity measure on this restricted domain is

$$\mathbb{E}\left[\sup_{f\in\mathcal{G}_x(0,r,1)}\left|\frac{1}{N^{1/2}}\sum_{j=1}^{N}\xi^{(j)}f(x^{(j)})\right|\right].$$

In the following proposition we show that the Rademacher complexity of $\mathcal{G}_x(0, r, 1)$ given a distribution P satisfying $\min_{x\in\mathcal{X}} P(x)|\mathcal{X}| \geq p_{\texttt{min}}$ can be upper bounded by the complexity of $\mathcal{G}_{\widetilde{x}}(0, r, 1)$ given distributions over $\widetilde{\mathcal{X}}$ satisfying the conditions of Theorem C.2 for $m_0 = p_{\texttt{min}}, M_0 = m^d$.

Proposition C.3. For any distribution P satisfying $\min_{x\in\mathcal{X}} P(x)|\mathcal{X}| \geq p_{\texttt{min}}$, there exists a distribution $\widetilde{P}$ over $\widetilde{\mathcal{X}}$ satisfying the conditions of Proposition 8 with constants $m_0 = p_{\texttt{min}}, M_0 = m^d$ such that

$$\mathbb{E}\left[\sup_{f\in\mathcal{G}_x(0,r,1)}\left|\sum_{j=1}^{N}\xi^{(j)}f(x^{(j)})\right|\right] \leq \mathbb{E}\left[\sup_{\widetilde{f}\in\mathcal{G}_{\widetilde{x}}(0,r,1)}\left|\sum_{j=1}^{N}\xi^{(j)}\widetilde{f}(\widetilde{x}^{(j)})\right|\right].$$

The result of this proposition (which we prove later) is an inequality in the complexities of function classes. This enables us to directly apply the distribution-dependent upper bounds. Let the operator $\lesssim_c$ denote an inequality up to a constant factor depending only on c for any parameter c. By Proposition C.3, $\mathbb{E}[\|\widehat{f} - f\|_{L_2(P)}\mathbb{I}_A] \lesssim_{p_{\texttt{min}}} N^{-1/d}\log^{\gamma_d} N$. Thus, assuming Proposition C.3, we have a bound of the first term.

To bound the second term, $\mathbb{E}[\|\widehat{f} - f\|_{L_2(P)}\mathbb{I}_{A^C}]$, because $f, \widehat{f}$ map to points in $[0, 1]$, we have the following inequality

$$\mathbb{E}[\|\widehat{f} - f\|_{L_2(P)}\mathbb{I}_{A^C}] \leq P(A^C).$$

In order to bound $P(A^C)$, the authors prove that $P(A^C) \lesssim N^{-1}$ (see (Han et al., 2019) Lemma 10) for their specific choice of a. Specifically, they use the fact that $\|\widehat{f} - \widetilde{f}\|_\infty \leq \|\widehat{f}\|_\infty + \|\widetilde{f}\|_\infty \leq 2 + \max_{j\in[N]}|w^{(j)}|$ by the constraint that $\|\widetilde{f}\|_\infty \leq 1$ and $\|\widehat{f}\|_\infty \leq \max_{j\in[N]}\widetilde{f}(x^{(j)}) + w^{(j)}$. This argument is the only one dependent on the interpolation method and relies only on the fact that the value of unseen x is within the range of observed outputs. Our interpolation method is an average of the function values for observed outputs, and thus their argument still applies to attain a $O(N^{-1})$ upper bound on $\mathbb{E}[\|\widehat{f} - f\|_{L_2(P)}\mathbb{I}_{A^C}]$. Because the second term is asymptotically smaller, the risk is dominated by $\mathbb{E}[\|\widehat{f} - f\|_{L_2(P)}\mathbb{I}_A]$. Thus

$$R(f_0, f) \lesssim_{d,p_{\texttt{min}}} N^{-1/d}\log^{\gamma_d} N + N^{-1}$$

Applying our oracle inequality on the risk of $f_{\texttt{CV}}$:

$$\begin{aligned} R(f_{\texttt{CV}}, f) &\leq 2\inf_{\lambda\in\Lambda} R(f_\lambda, f) + C/N\log(1 + |\Lambda|) \\ &\lesssim_{d,p_{\texttt{min}}} N^{-1/d}\log^{\gamma_d} N. \end{aligned}$$

Therefore, all that is left is to prove Proposition C.3.

Proof of Proposition C.3. We will prove this showing that there exists a joint distribution over x and $\widetilde{x}$ such that for all functions $f \in \mathcal{G}_x(0, r, 1)$ there exists a function $\widetilde{f} \in \mathcal{G}_{\widetilde{x}}(0, r, 1)$ such that $f(x) = \widetilde{f}(\widetilde{x})$ with probability 1. If such a construction exists, it implies that given any realized of the $\xi^{(j)}$, any supremum taken over all $\mathcal{G}_x(0, r, 1)$ (including the Rademacher complexity) is at most the supremum taken over $\mathcal{G}_{\widetilde{x}}(0, r, 1)$. As the inequality holds with probability 1 and does not depend on

the joint distribution of ξ and $(x, \widetilde{x})$, it also holds for the expected Rademacher complexities when ξ is independent of the random samples on $\mathcal{X} \times \widetilde{\mathcal{X}}$, proving the Lemma.

Hence, all that is left is to show that such a joint distribution and set of functions that are equal with probability 1 exists. To define this coupling between $\mathcal{X}$ and $\widetilde{\mathcal{X}}$, we first will partition $[0,1]^d$ into hypercubes of equal size. For all $x \in [m]^d$, let $A_x = \prod_{i=1}^d R_{x_i}$ where $R_\ell = (\frac{\ell-1}{m}, \frac{\ell}{m}]$ for $\ell > 1$ and $R_\ell = [\frac{\ell-1}{m}, \frac{\ell}{m}]$ for $\ell = 1$. Let $\mathcal{A} = \{A_x : x \in [m]^d\}$. By construction, all elements of $\mathcal{A}$ have volume $\frac{1}{m^d}$. Let $\widetilde{x} \sim \mathrm{Unif}(A_x)$, that is, $\widetilde{x}$ takes on all values in the hypercube $A_{\widetilde{x}}$ with equal probability. Now we will confirm that the distribution of $\widetilde{x}$, denoted by $\widetilde{P}$, satisfies the condition of Proposition 8, namely that

$$p_{\texttt{min}} \leq \inf_{\widetilde{x} \in \widetilde{\mathcal{X}}} \widetilde{P}(\widetilde{x}) \leq \sup_{\widetilde{x} \in \widetilde{\mathcal{X}}} P_{\widetilde{x}}(\widetilde{x}) \leq m^d.$$

Let $x(\widetilde{x})$ be the element of $\mathcal{X}$ corresponding to the hypercube that $\widetilde{x}$ is contained in. By construction $\widetilde{P}(\widetilde{x}) = P(x(\widetilde{x}))m^d$. Since $\frac{P(x)}{|\mathcal{X}|} \in [p_{\texttt{min}}, 1]$ for all $x \in [m]^d$, it follows that $P(x(\widetilde{x})) \in [m_0, M_0]$ for all $\widetilde{x} \in \widetilde{\mathcal{X}}$ for our specific choice of the constants.

Given this joint distribution, now we need to show that for all functions $f \in \mathcal{G}_x(0, r, 1)$ there exists a function $\widetilde{f} \in \mathcal{G}_{\widetilde{x}}(0, r, 1)$ such that $f(x) = \widetilde{f}(\widetilde{x})$. Pick any $f \in \mathcal{G}_x(0, r, 1)$. Consider the function $\widetilde{f}(\widetilde{x}) = f(x(\widetilde{x}))$. Since $x(\widetilde{x})$ and f are isotonic, and the composition of isotonic functions is isotonic, so is $\widetilde{f}$. Further, by construction $\mathbb{E}_{\widetilde{x} \sim \widetilde{P}} \|\widetilde{f}\|_2 = \mathbb{E}_{x \sim \P} \|f\|_2 \leq r$ and $\|\widetilde{f}\|_\infty = \|f\|_\infty \leq 1$ which implies that $\widetilde{f} \in B_2(r, \widetilde{P}) \cap B_\infty(1)$ and thus $\widetilde{f} \in \mathcal{G}_{\widetilde{x}}(0, r, 1)$.

This concludes our proof that there exists a joint distribution over x and $\widetilde{x}$ such that for all functions $f \in \mathcal{G}_x(0, r, 1)$ there exists a function $\widetilde{f} \in \mathcal{G}_{\widetilde{x}}(0, r, 1)$ such that $f(x) = \widetilde{f}(\widetilde{x})$ with probability 1. Since such a distribution exists and is a sufficient condition for the statement of the Lemma, we are done. □

D Supplementary material for Section 6

Necessary compute resources. All experiments were run locally on a personal computer. The computational requirements are minimal—all experiments complete in under a few hours on standard consumer hardware (we used the "caffeinate" command in terminal) and do not require GPUs or specialized infrastructure.

Code and data availability. Our code is publicly available via the following GitHub link: `https://github.com/madelineceli13/evaluator-preference-algorithm`

The datasets for ICLR and Tripadvisor are too large to store on GitHub. The can be accessed at the following links:

1. The full Tripadvisor dataset is available here: `https://github.com/diegoantognini/HotelRec`
2. The full ICLR dataset is available here: `https://huggingface.co/datasets/IntelLabs/AI-Peer-Review-Detection-Benchmark`

D.1 Synthetic simulations

We use simulation studies to evaluate the performance of our algorithm relative to other algorithms as sample size increases for learning both linear and non-linear preferences.

D.1.1 Comparison of our algorithm with linear regression

In particular, we compare our algorithm against non-negative least squares for three classes of preference or utility functions common in the economics literature. Each of these preference

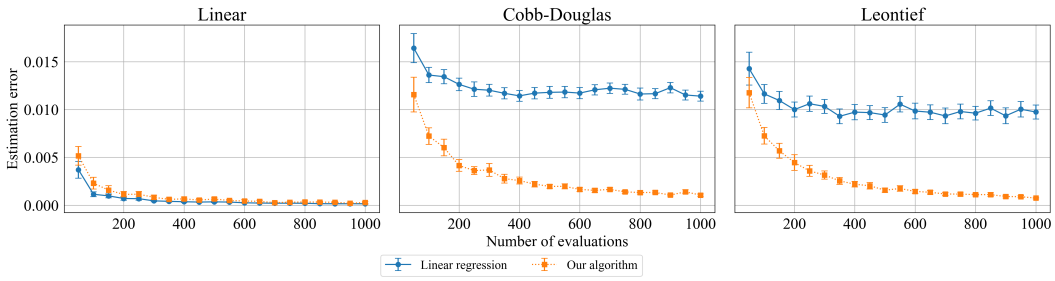


Figure 4: Simulation results for synthetic preferences.

functions are parameterized by a vector $a \in \mathbb{R}^d_{\geq 0}$ and are normalized such that they provide outputs in $[0, 1]$.

- *Linear utility.* Under linear preferences, criteria are perfect substitutes for one another. That means we can exchange an increase in criteria 1 for some increase in criteria 2, that doesn't depend on the level of the criteria scores:
$$f(x) = \frac{1}{m\|a\|_1} \sum_{i \in [d]} a_i x_i.$$
- *Leontief utility.* Leontief utility assumes that criteria 1 and 2 perfectly complement one another. The output of the function is the lowest of the weighted criteria scores. Therefore, an increase in criteria 1 matters only if criteria 2 is sufficiently high:
$$f(x) = \frac{1}{m \min_{i \in [d]} a_i} \min_{i \in d} a_i x_i.$$
- *Cobb-Douglas utility.* Cobb-Douglas utility assumes that the criteria scores (raised to some power) multiply one another. Hence the relative benefit of a criteria score is dependent upon the other scores. In Cobb-Douglas preferences, the complementarity of the criteria lies somewhere between linear and Leontief preferences:
$$f(x) = \frac{1}{m^{\|a\|_1}} \prod_{i \in [d]} x_i^{a_i}.$$

To measure the ability of both algorithms to learn these three classes of preference functions, we run 50 trials for a set of 20 sample sizes between 50 and 1000 with $d = 2$ criteria and $m = 5$ possible scores values per criterion. For each trial, the preference parameter a is drawn uniformly from $[1, 2]^d$ and we observe samples with independent $N(0, \sigma^2)$ noise with $\sigma = 0.2$. We set P to be the uniform distribution over $\mathcal{X}$. After we fit the models on the chosen data, we compute the risk or estimation error for both the linear and cross-validation estimator. Error bars represent a 95% confidence interval for the expected estimation error and are computed using the sample standard deviation.

Our theoretical results, namely Theorem 5.1 part (a), predicts that our algorithm should converge at a similar rate compared with linear regression when the function is truly linear. The purpose of the simulations in part is to determine the degree of the constant factors. As shown in Figure 4, they have very similar risks. At the smallest sample size ($N = 50$), the difference between our algorithm and linear regression is not statistically significant, and by $N = 150 - 200$, they are effectively identical. This evidence suggests that, in practice, our algorithm really does quite similarly to linear regression for linear models in the worst case. We have also tested for higher values of d, m, and noise at $\sigma = 0.5$ rather than $\sigma = 0.2$, and found similar results to hold, though our algorithm becomes comparable at slightly higher sample sizes for the larger d or m cases. For the sake of time, we only test 10 trials. These results are shown in Figure 5.

Additionally, note that for the non-linear functions (Cobb-Douglas and Leontief), linear regression algorithm is unable to decrease error at some point around $N = 400$ samples. This illustrates the bias when linear models are used to learn functions their model class does not contain, aligning with our bias result, Proposition 4.1. In contrast to linear regression, our algorithm is still able to

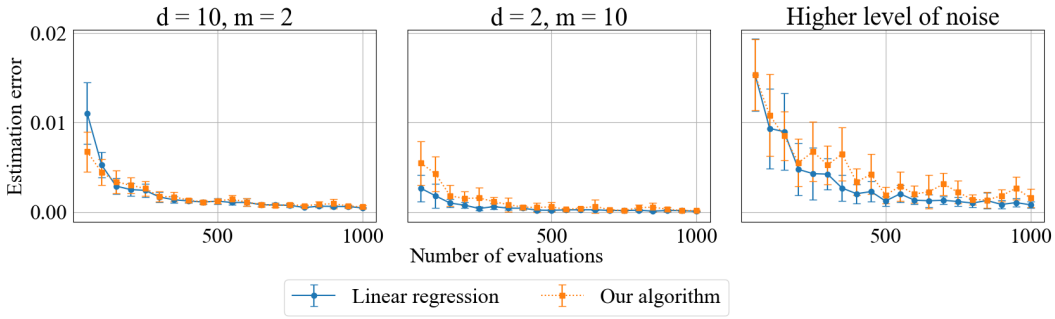


Figure 5: Additional simulations for linear utility.

learn the non-linear utility functions, albeit at a slower rate. The two algorithms are comparable at small sample sizes where little information about the preferences are known, but as the sample size increases, the value-added of using our algorithm increases. For Leontief preferences, at $N = 50$ the average error of linear regression is about 1.2 times that of our algorithm, at $N = 1000$ the average error is over 13 times larger.

D.1.2 Comparison of our algorithm with alternative machine learning methods

In addition to comparing our algorithm with non-negative least squares, we test two additional methods: a monotone-constrained neural network (Runje and Shankaranarayana, 2023) and a monotone-constrained gradient-boosted tree ensemble. A monotone-constrained neural network (henceforth monotone NN) enforces monotonicity in each variable. We use the methods detailed in Runje and Shankaranarayana (2023). A monotone-constrained gradient-boosted tree ensemble (henceforth monotone GBM for gradient-boosting machine) follows the structure of a typical GBM. Starting from a constant prediction, the model fits an estimator, in this case a regression tree, to the residuals (the errors the ensemble makes on each training point) and then adds that tree (scaled by a learning rate) to the current prediction. Then the model recomputes the new residuals and repeats, with each new tree correcting the errors from the previous ones, until a stopping criterion is reached (e.g., a maximum number of trees, minimum leaf size or marginal error reduction). The final estimator is the sum of all fitted trees. In a monotone GBM, the only additional component is that the estimators are constrained to generate isotonic outputs. As such, any weighted combination of the individual estimators (with non-negative weights) will output an isotonic function. Each method—monotone NNs and GBMs—is tuned by cross-validation on the same train/validation/test split and clipping range as our method, so all reported estimation and reducible errors are directly comparable. In particular, we test each algorithm on linear, Cobb-Douglas, and Leontief preference functions using the same setup for the comparison with linear regression. As shown in Figure 6, our algorithm outperforms GBM on all preference types. Moreover, while the monotone NN has a similar estimation error to our algorithm for Linear and Cobb-Douglas preferences, it does significantly worse for Leontief preferences.

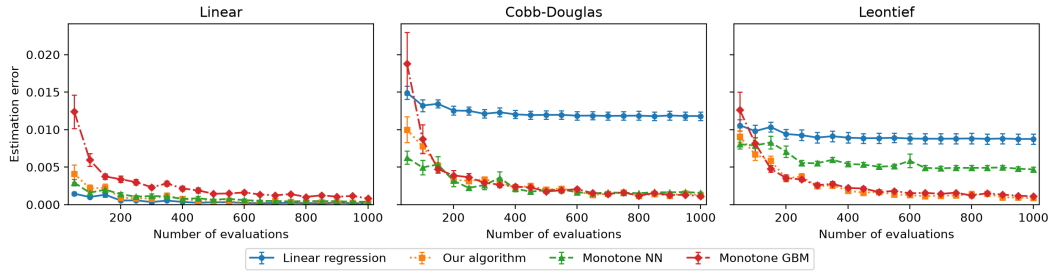


Figure 6: Simulation results for synthetic preferences including monotone NN and GBM baselines. Error bars are computed using the standard error of the mean estimation error.

Dataset	# criteria (d)	# values (m)	score range	# samples (N)
Tripadvisor ratings				
All (2019)	6	5	1–5	332,349
Couples (2014–2019)				302,894
Family (2014–2019)				371,185
Business (2014–2019)				197,031

Table 1: Summary of dataset attributes.

Table 2: Empirical results on Tripadvisor dataset. Standard errors are computed with the standard error of the mean for prediction error. Δ (%) denotes the relative error reduction of our algorithm over linear regression.

	Prediction Error			Reducible Error		
	Lin. Reg.	Our algo	Δ (%)	Lin. Reg.	Our algo	Δ (%)
All - 2019	0.245 (0.004)	0.217 (0.003)	11.6%	0.041 (0.004)	0.013 (0.003)	**69.1%**
Business travel	0.209 (0.004)	0.192 (0.003)	8.0%	0.031 (0.004)	0.014 (0.003)	**54.6%**
Couples travel	0.176 (0.002)	0.161 (0.002)	8.0%	0.026 (0.002)	0.012 (0.002)	**54.2%**
Family travel	0.219 (0.003)	0.200 (0.002)	8.3%	0.03 (0.003)	0.011 (0.002)	**61.3%**

D.2 Supplementary information for Section 6.1

Data information and Methodology We restrict our attention to the set of reviews containing only overall scores and the six most common criteria: value, rooms, location, cleanliness, service, and sleep quality. To ensure we have captured the relevant information in someone's review, we include only ratings in years after 2014 where the number of users ratings other additional criteria (e.g., business service and check in/ front desk) is negligible. Additionally, as trip type is not part of the HotelRec dataset we use a rule-based classification of review comments as a proxy. We only test categories for which the classification was reliable meaning that out of a random sample, at least 8 out of 10 were correctly classified: these were business, family, and couples. For learning preferences of different trip types, we included data from 2014-2019. To avoid long run-times, for the dataset of all users we used evaluations from just 2019. See Table 1 for an overview the sample sizes and attributes of these datasets.

Additional results Firstly, Table 2 presents a full report of the measures displayed in Figure 2. Second, to complement our findings on prediction and estimation error, we study qualitative aspects of the learned preference function.

In particular, we ask: *what insights can our learned functions tell us about people's preferences?* To further illustrate the benefit of learning preferences as well as possible, we provide an example of how we can use learned functions to understand what people care about. One question of interest is the relative impact an increase in the score of one criterion (all else constant) compared with another. For instance, does raising a service or rooms rating matter more and how does this depend on other aspects of the stay? There is no natural analog of weights for multidimensional isotonic functions. Instead, we capture the relative effects of a marginal change in criteria score by graphing the learning function for different values of that criterion while holding the remaining criteria fixed.

To answer question 3, we plot the values of the learned preference function for the 2019 dataset fixing all but one criterion at 3 stars. Specifically we look at the effects of the first three criteria: service, location, and value. The result is shown in Figure 7. Interestingly, we find that the curves generally have a concave shape. This is consistent with the intuition that one truly bad aspect of a stay can have a significant impact on the overall experience, but even really good service or location can only make an otherwise average experience marginally better. Additionally, the relative benefit of an increase in criteria scores (service, location, or value) depend the level of that score. When changing from 1-2 stars, the service criteria has the largest impact, however when moving from 4-5 stars, they all

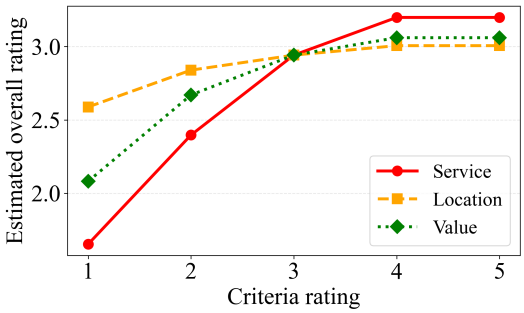


Figure 7: Impact of increasing different criteria scores on the overall rating of the learned preference function for the dataset of Tripadvisor reviews from 2019. Other criteria scores are fixed at 3 out of 5 stars.

Table 3: Level of preference consistency and alignment by evaluator type. Standard errors are in parenthesis for preference misalignment. 95% bootstrap CI are in brackets for irreducible error.

	GPT-4o	Llama 3.1 70b	Human evaluators
Irreducible error	0.321 [0.311,0.330]	1.139 [1.122,1.153]	1.251 [1.228,1.267]
Preference misalignment	0.971 (0.014)	5.157 (0.024)	—

Dataset	# criteria (d)	# values (m)	score range	# samples (N)
ICLR peer reviews				
Human evaluators				27,857
GPT-4o	3	4	1–10	27,857
Llama 3.1 70b				27,834

Table 4: Summary of ICLR dataset attributes.

essentially have no effect. **A linear model, which assumes constant marginal changes, cannot capture either of these two insights.** This example illustrates that linear regression will always be constrained in its ability to tell us about evaluator preferences. Hence, if the goal is to learn what evaluators' care about, it may be worth running our algorithm even if the error reduction is modest.

D.3 Supplementary material for Section 6.2

The reviews from AI and human evaluators are for the same papers from ICLR 2024 and each dataset contains 27k reviews (see Table 4). Figure 1 (b) illustrates an example review. Reviewers provide criteria assessments for presentation, contribution, and soundness. These each take on values integer values between 1 (poor) and 4 (excellent). Additionally, the reviewers provide an overall rating or recommendation, which can take on values between 1 (strong reject) and 10 (strong accept). The AI reviews are generated using two widely used LLMs: GPT-4o and Llama 3.1 70b.

To estimate uncertainty in our noise alignment measure, we cannot simply report the standard deviation of the mean. This is because we use that same data to fit the function, which would create an issue of "double-dipping." Instead, we use a bootstrap procedure with 1000 resamples. For each resample, we sample with replacement from the dataset and recompute irreducible error; we report the mean across resamples alongside 95% confidence intervals computed via the percentile method, taking the 2.5th and 97.5th percentiles of the bootstrap distribution.

D.4 Results on additional human preference datasets

In addition to Tripadvisor, we test our algorithm relative to linear regression on two other datasets of human preferences. However, their sample size was insufficient to generate statistical significant findings.

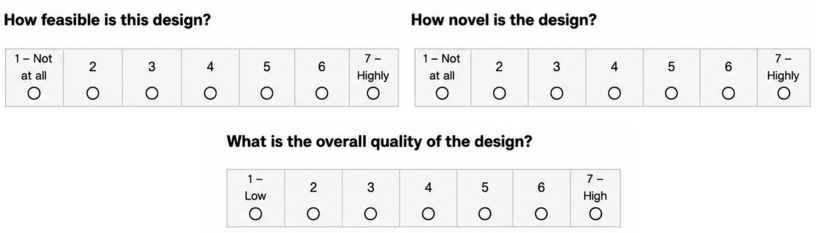


Figure 8: Review template for Astrobee evaluations.

Dataset	# criteria (d)	# values (m)	score range	# samples (N)
ICLR peer review data				
2023			1–10	13,212
2024	3	4	1–8	28,028
2025			1–10	46,748
Astrobee Challenges Series evaluations	2	7	1–7	3,850

Table 5: Summary of dataset attributes for ICLR and Astrobee evaluations.

ICLR Peer Review data . The first dataset is of criteria scores and overall scores or recommendations provided by peer reviewers for the International Conference on Learning Representations (ICLR) from 2023-2025. Since all of the reviews are publicly available, we obtained them via web scraping. Additionally, each year used different criteria and/or rating scales. Every year has 3 criteria which are rated on a scale of 1-4, and an overall recommendation or rating. In 2023, the criteria were correctness, technical novelty and significance, and empirical novelty and significance. In 2024 and 2025 the criteria were soundness, presentation, and contribution. For all years, overall ratings took on values in $\{1, 3, 5, 6, 8, 10\}$. In this context, the non-equal intervals are meant to show the relative difference in each level of recommendation. For instance, in 2024 3 meant "reject, not good enough" and 6 meant "marginally above the acceptance threshold." Since the criteria and/or rating scale varied based on the year, we analyzed each separately. The data was cleaned to contain only reviews for which all criteria scores were filled in and obtained via web-scraping. An example of an evaluation from 2024 is presented in Figure 1 (b).

Astrobee Challenges Series evaluations The last dataset we tested contains evaluations of engineering robotics solutions in the aerospace domain. The solutions being evaluated came from a competition called the Astrobee Challenges Series run by NASA and other researchers. They asked people to propose solutions to 17 different challenges relating to the creation of an attachment and orientation arm for Astrobee, a free-flying robotic system used by NASA. The winning solution got money and were incorporated into new designs. After the competition ended, Lane and co-authors did a study to see how the type of evaluator impacted their specific preferences (Lane et al., 2023). Evaluators were asked to provide criteria scores for the feasibility and novelty of the solution as well as an overall quality score. Each score took on integer values between 1 and 7. We use data from that study which contains $3,850$ evaluations from 374 evaluators corresponding to 101 solutions of 9 challenges. An example of what the response questions look like is given in Figure 8.

Results In all cases, our algorithm does weakly better. For ICLR some of the improvements in estimation error are very large (as much as 81.6% less error), however none of the results are statistically significant. Therefore we do not place much emphasis on them. We suspect that the lack of significance is a function of insufficient sample size. The largest of the datasets is ICLR 2025 with a sample size of 47k. For comparison, the smallest Tripadvisor dataset is over four times this size, 197k. The smallest dataset is the Astrobee evaluations with 3.8k samples which is nearly one hundredth the size of the largest Tripadvisor dataset with 371k samples.

Table 6: Results for linear regression and our algorithm on ICLR and Astrobee datasets. Standard errors are computed with the standard error of the mean prediction error. Δ (%) denotes the relative error reduction of our algorithm over linear regression.

	Prediction Error			Reducible Error		
	Lin. Reg.	Our algo	$\Delta(\%)$.	Lin. Reg.	Our algo	$\Delta(\%)$
ICLR 2023	1.518 (0.043)	1.491 (0.043)	1.8%	0.053 (0.043)	0.027 (0.043)	50.2%
ICLR 2024	1.531 (0.051)	1.462 (0.042)	4.5%	0.084 (0.051)	0.015 (0.042)	81.6%
ICLR 2025	1.280 (0.019)	1.257 (0.019)	1.8%	0.03 (0.019)	0.007 (0.019)	75.3%
Astrobee	0.893 (0.081)	0.893 (0.081)	0.0%	0.079 (0.081)	0.079 (0.081)	0.5%

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: Our paper concerns the problem of learning evaluator preferences and we issues with and methods for learning preferences reliably given model mismatch, as explained in the introduction and abstract.
   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: Our final discussion section discusses limitations alongside directions for future work.
   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: We outline our model in section 3, use formal statements of our results, and include proofs in the appendix.
   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: We include a github link to all of our code. Alternatively, others can reproduce our main results by referring to our data information and methodology in appendices D and our algorithm in the main body of the paper.
   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: We provide instructions for reproducibility, in addition to our code as supplementary material. We provide links to the Tripadvisor and ICLR dataset.
   Guidelines:
   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.
6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: Yes. Our experiments rely only on our algorithm (explained fully in the paper), and we release code to reproduce those results.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: For reducible error and prediction error we compute standard errors using the squared prediction error samples. For irreducible error, we use bootstrapping. For preference misalignment, we compute usual standard errors.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: Our experiments do not require any specialized compute resources. We explain this in the Appendix for Section 6.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).
9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: We have reviewed the code of ethics and confirmed that our work conforms to it.
   Guidelines:
   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [Yes]
    Justification: Yes, we discuss the societal context in our introduction, and as we introduce our experimental results, and touch on negative consequences in our discussion.
    Guidelines:
    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate

Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we cite all of the owners of the two LLMs and four datasets we use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not include research with human subjects. All of our preference data comes from existing work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Since we do not do researcher with human subjects, no approval process is relevant.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used for an important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.