Preference learning along multiple criteria: A game-theoretic perspective

Kush Bhatia
EECS, UC Berkeley
kush@cs.berkeley.edu

Ashwin Pananjady
Simons Institute, UC Berkeley
ashwinpm@berkeley.edu

Peter L. Bartlett
EECS and Statistics, UC Berkeley
peter@berkeley.edu

Anca D. Dragan
EECS, UC Berkeley
anca@berkeley.edu

Martin J. Wainwright
EECS and Statistics, UC Berkeley
wainwrig@berkeley.edu

Abstract

The literature on ranking from ordinal data is vast, and there are several ways to aggregate overall preferences from pairwise comparisons between objects. In particular, it is well-known that any Nash equilibrium of the zero-sum game induced by the preference matrix defines a natural solution concept (winning distribution over objects) known as a von Neumann winner. Many real-world problems, however, are inevitably multi-criteria, with different pairwise preferences governing the different criteria. In this work, we generalize the notion of a von Neumann winner to the multi-criteria setting by taking inspiration from Blackwell's approachability. Our framework allows for non-linear aggregation of preferences across criteria, and generalizes the linearization-based approach from multi-objective optimization.

From a theoretical standpoint, we show that the Blackwell winner of a multi-criteria problem instance can be computed as the solution to a convex optimization problem. Furthermore, given random samples of pairwise comparisons, we show that a simple, "plug-in" estimator achieves (near-)optimal minimax sample complexity. Finally, we showcase the practical utility of our framework in a user study on autonomous driving, where we find that the Blackwell winner outperforms the von Neumann winner for the overall preferences.

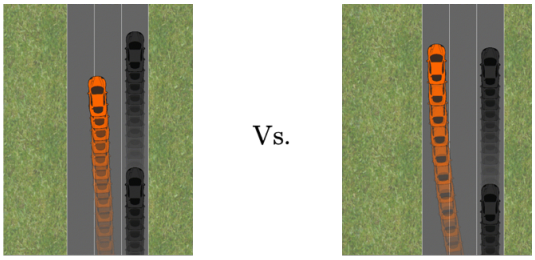
1 Introduction

Economists, social scientists, engineers, and computer scientists have long studied models for human preferences, under the broad umbrella of social choice theory [10, 7]. Learning from human preferences has found applications in interactive robotics for learning reward functions [45, 39], in medical domains for personalizing assistive devices [59, 9], and in recommender systems for optimizing search engines [15, 28]. The recent focus on safety in AI has popularized human-in-the-loop learning methods that use human preferences in order to promote value alignment [16, 46, 6].

The most popular form of preference elicitation is to make pairwise comparisons [51, 13, 33]. Eliciting such feedback involves showing users a pair of objects and asking them a query: Do you prefer object A or object B? Depending on the application, an object could correspond to a product in a search query, or a policy or reward function in reinforcement learning. A vast body of classical work dating back to Condorcet and Borda [17, 12] has focused on defining and producing a "winning" object from the result of a set of pairwise comparisons.

In relatively recent work, Dudik et al. [22] proposed the concept of a von Neumann winner, corresponding to a distribution over objects that beats or ties every other object in the collection. They showed that under an expected utility assumption, such a randomized winner always exists and

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a)

	A	B
A	0.50	0.60
B	0.40	0.50
R	0.55	0.63

$\mathbf{P}_{\text{ex}}^{\text{Overall}}$

	A	B
A	0.50	0.80
B	0.20	0.50
R	0.35	0.65

$\mathbf{P}_{\text{ex}}^{\text{Comfort}}$

	A	B
A	0.50	0.30
B	0.70	0.50
R	0.60	0.40

$\mathbf{P}_{\text{ex}}^{\text{Speed}}$

(b)

Figure 1. (a) Policy A focuses on optimizing comfort and policy B on speed, and these are compared pairwise in different environments. (b) Preference matrices, where entry (i, j) of the matrix contains the proportion of comparisons between the pair (i, j) that are won by object i. (The diagonals are set to half by convention). The overall pairwise comparisons are given by the matrix $\mathbf{P}_{\text{ex}}^{\text{Overall}}$, and preferences along each of the criteria by matrices $\mathbf{P}_{\text{ex}}^{\text{Comfort}}$ and $\mathbf{P}_{\text{ex}}^{\text{Speed}}$ (the numbers here are illustrative of our user-study in Section 4). Policy R is a randomized policy $1/2$ A $+1/2$ B. While the preference matrices satisfy the linearity assumption individually along speed and comfort, the assumption is violated overall, wherein R is preferred over both A and B.

overcomes limitations of existing winning concepts—the Condorcet winner does not always exist, while the Borda winner fails an independence of clones test [47]. However, the assumption of expected utility relies on a strong hypothesis about how humans evaluate distributions over objects: it posits that the probability with which any distribution over objects π beats an object is linear in π.

Consequences of assuming linearity: In order to better appreciate these consequences, consider as an example the task of deciding between two policies (say A and B) to deploy in an autonomous vehicle. Suppose that these policies have been obtained by optimizing two different objectives, with policy A optimized for comfort and policy B optimized for speed. Figure 1(a) shows a snapshot of these two policies. When compared overall, 60% of the people preferred Policy A over B – making it the von Neumann winner. The linearity assumption then posits that a randomized policy that mixes between A and B can *never* be better than both A and B; but we see that the Policy R = $1/2$ A + $1/2$ B is actually preferred by a majority over both A and B! Why is the linearity assumption violated here?

One possible explanation for such a violation is that the comparison problem is actually *multi-criteria* in nature. If we look at the preferences for the criterion speed and comfort individually in Figure 1(b), we see that Policy A does quite poorly on the speed axis while B lags behind in comfort. In contrast, Policy R does acceptably well along both the criteria and hence is preferred overall to both Policies A and B. It is indeed impossible to come to this conclusion by only observing the overall comparisons. This observation forms the basis of our main proposal: decompose the single overall comparison and ask humans to provide preferences along *simpler* criteria. This decomposition of the comparison task allows us to place structural assumptions on comparisons along each criterion. For instance, we may now posit the linearity assumption along each criterion separately rather than on the overall comparison task. In addition to allowing for simplified assumptions, breaking up the task into such simpler comparisons allows us to obtain richer and more accurate feedback as compared to the single overall comparison. Indeed, such a motivation for eliciting simpler feedback from humans finds its roots in the the study of cognitive biases in decision making, which suggests that the human mind resorts to simple heuristics when faced with a complicated questions [53].

Contributions: In this paper, we formalize these insights and propose a new framework for preference learning when pairwise comparisons are available along multiple, possibly conflicting, criteria. As shown by our example in Figure 1, a single distribution which is the von Neumann winner along every criteria might not exist. To counter this, we formulate the problem of finding the "best" randomized policy by drawing on tools from the literature on vector valued pay-offs in game theory. Specifically, we take inspiration from Blackwell's approachability [11] and introduce the notion of a Blackwell winner. This solution concept strictly generalizes the concept of a von Neumann winner, and recovers the latter when there is only a single criterion present. Section 2 describes this framework in detail, and Section 3 collects our statistical and computational guarantees for learning the Blackwell winner from data. Section 4 describes a user study with an autonomous driving environment, in which we ask human subjects to compare self-driving policies along multiple

criteria such as safety, aggressiveness, and conservativeness. Our experiment demonstrates that the Blackwell winner is able to better trade off utility along these criteria and produces randomized policies that outperform the von Neumann winner for the overall preferences.

Related work. Most closely related to our work is the field of computational social choice, which has focused on defining notions of winners from overall pairwise comparisons (see the survey [37] for a review). Amongst them, three deterministic notions of a winner—the Condorcet [17], Borda [12], and Copeland [18] winners—have been widely studied. In addition, Dudik et al. [22] recently introduced the notion of a (randomized) von Neumann winner. Starting with the work of Yue et al. [57], there have been several research papers studying an online version of preference learning, called the Dueling Bandits problem. Algorithms have been proposed to compete with Condorcet [60, 62, 4], Copeland [61, 56], Borda [30] and von Neumann [22] winners.

The theoretical foundations of decision making based on multiple criteria have been widely studied within the operations research community . This sub-field—called multiple-criteria decision analysis—has focused largely on scoring, classification, and sorting based on multiple-criteria feedback. See the surveys [44, 63] for thorough overviews of existing methods and their associated guarantees. The problem of eliciting the user's relative weighting of the various criteria has also been considered [20]. However, relatively less attention has been paid to the study of randomized decisions and statistical inference, both of which form the focus of our work. From an applied perspective, the combination of multi-criteria assessments has received attention in disparate fields such as psychometrics [40, 35], healthcare [50], and recidivism prediction [55]. In many of these cases, a variety of approaches—both linear and non-linear—have been empirically evaluated [19]. Justification for non-linear aggregation of scores along the criteria has a long history in psychology and the behavioral sciences [27, 24, 54].

In the game theory literature, Blackwell [11] introduced the notion of approachability as a generalization of a zero-sum game with vector-valued payoffs (for a detailed discussion see Appendix A). Blackwell's approachability and its connections with no-regret learning and calibrated forecasting have been extensively studied [1, 42, 34]. These connections have enabled applications of Blackwell's results to problems ranging from constrained reinforcement learning [36] to uncertainty estimation for question-answering tasks [31]. In contrast, our framework for preference learning along multiple criteria deals with a single shot game and uses the idea of the target set to define the concept of a Blackwell winner. Another body of literature related to our work studies Nash equilibria in games with perturbed payoffs, under both robust [3, 32] and uncertain (or Bayesian) [25] formulations (see the recent survey by Perchet [43]). Perturbation theory for Nash equilibria has been derived in these contexts, and it is well-known that the Nash equilibrium is not (at least in general) stable to perturbations of the payoff matrix. On the other hand, the results of [22] consider Nash equilibria of perturbed, symmetric, zero-sum games, but show that the *payoff* of the perturbed Nash equilibrium is indeed stable. Our work provides a similar characterization for the multi-criteria setting.

2 Framework for preference learning along multiple criteria

We now set up our framework for preference learning along multiple criteria. We consider a collection of d objects over which comparisons can be elicited along k different criteria. We index the objects by the set $[d] := \{1, \ldots, d\}$ and the criteria by the set $[k]$.

2.1 Probabilistic model for comparisons

Since human responses to comparison queries are typically noisy, we model the pairwise preferences as random variables drawn from an underlying population distribution. In particular, the result of a comparison between a pair of objects (i_1, i_2) along criterion j is modeled as a draw from a Bernoulli distribution, with $p(i_1, i_2; j) = \mathbb{P}(i_1 \succeq i_2 \text{ along criterion } j)$. By symmetry, we must have

$$p(i_2, i_1; j) = 1 - p(i_1, i_2; j) \text{ for each triple } i_1 \in [d],\ i_2 \in [d], \text{ and } j \in [k]. \tag{1}$$

We let $\pi_1, \pi_2 \in \Delta_d$ represent[1] two distributions over the d objects. With a slight abuse of notation, let $p(\pi_1, \pi_2; j)$ denote the probability with which an object drawn from distribution π_1 beats an object

[1] We let Δ_d denote the d-dimensional simplex.

drawn from distribution π_2 along criterion j. We assume for each individual criterion j that the probability $p(\pi_1, \pi_2; j)$ is linear in the distributions π_1 and π_2, i.e. that it satisfies the relation

$$p(\pi_1, \pi_2; j) := \mathbb{E}_{i_1 \sim \pi_1, i_2 \sim \pi_2} \left[p(i_1, i_2; j) \right]. \tag{2}$$

Equation (2) encodes the per-criterion linearity assumption highlighted in Section 1. We collect the probabilities $\{p(i_1, i_2; j)\}$ into a *preference tensor* $\mathbf{P} \in [0,1]^{d \times d \times k}$ and denote by $\mathcal{P}_{d,k}$ the set of all preference tensors that satisfy the symmetry condition (1). Specifically, we have

$$\mathcal{P}_{d,k} = \{\mathbf{P} \in [0,1]^{d \times d \times k} \mid \mathbf{P}(i_1, i_2; j) = 1 - \mathbf{P}(i_2, i_1; j) \text{ for all } (i_1, i_2, j)\} . \tag{3}$$

Let $\mathbf{P}^j$ denote the $d \times d$ matrix corresponding to the comparisons along criterion j, so that $p(\pi_1, \pi_2; j) = \pi_1^\top \mathbf{P}^j \pi_2$. Also note that a comparison between a pair of objects (i_1, i_2) induces a *score vector* containing k such probabilities. Denote this vector by $\mathbf{P}(i_1, i_2) \in [0,1]^k$, whose j-th entry is given by $p(i_1, i_2; j)$. Denote by $\mathbf{P}(\pi_1, \pi_2)$ the score vector for a pair of distribution (π_1, π_2).

In the single criterion case when $k = 1$, each comparison between a pair of objects is along an *overall* criterion. We let $\mathbf{P}_{\mathsf{ov}} \in [0,1]^{d \times d}$ represent such an overall comparison matrix. As mentioned in Section 1, most preference learning problems are multi-objective in nature, and the overall preference matrix $\mathbf{P}_{\mathsf{ov}}$ is derived as a non-linear combination of per-criterion preference matrices $\{\mathbf{P}^j\}_{j=1}^k$. Therefore, even when the linearity assumption (2) holds across each criterion, it might not hold for the *overall* preference $\mathbf{P}_{\mathsf{ov}}$. In contrast, when the matrices $\mathbf{P}^j$ are aggregated linearly to obtain the overall matrix $\mathbf{P}_{\mathsf{ov}}$, we recover the assumptions of Dudik et al. [22].

2.2 Blackwell winner

Given our probabilistic model for pairwise comparisons, we now describe our notion of a Blackwell winner. When defining a winning distribution for the multi-criteria case, it would be ideal to find a distribution π^* that is a von Neumman winner along *each* of the criteria separately. However, as shown in our example from Figure 1, such a distribution need not exist. We thus need a generalization of the von Neumann winner that explicitly accounts for conflicts between the criteria.

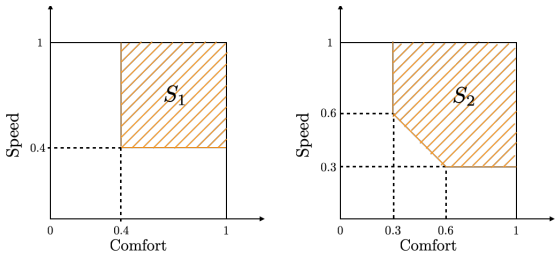


Figure 2. Two target sets S_1 and S_2 for our example from Figure 1 that capture trade-offs between comfort and speed. Set S_1 requires feasible score vectors to satisfy 40% of the population along both comfort and speed. Set S_2 requires both scores to be greater than 0.3 but with a linear trade-off: the combined score must be at least 0.9.

Blackwell [11] asked a related question for the theory of zero-sum games: how can one generalize von Neumann's minimax theorem to vector-valued games? He proposed the notion of a *target set*: a set of acceptable payoff vectors that the first player in a zero-sum game seeks to attain. Within this context, Blackwell proposed the notion of approachability, i.e. how the player might obtain payoffs in a repeated game that are close to the target set on average. We take inspiration from these ideas to define a solution concept for the multi-criteria preference problem.Our notion of a winner also relies on a *target set*, which we denote by $S \subset [0,1]^k$, and which in our setting contains *score vectors*. This set provides a way to combine different criteria by specifying combinations of preference scores that are acceptable. Figure 2 provides an example of two such sets.

Observe that for our preference learning problem, the target set S is by definition monotonic with respect to the orthant ordering, that is, if $z_1 \geq z_2$ coordinate-wise, then $z_2 \in S$ implies $z_1 \in S$. Our goal is to then produce a distribution π^* that can achieve a target score vector for any distribution with which it is compared—that is $\mathbf{P}(\pi^*, \pi) \in S$ for all $\pi \in \Delta_d$. When such a distribution π^* exists, we say that the problem instance $(\mathbf{P}, S)$ is *achievable*. On the other hand, it is clear that there are problem instances $(\mathbf{P}, S)$ that are not achievable. While Blackwell's workaround was to move to the setting of repeated games, preference aggregation is usually a one-shot problem. Consequently, our relaxation instead introduces the notion of a *worst-case distance* to the target set. In particular, we measure the distance between any pair of score vectors $u, v \in [0,1]^k$ as $\rho(u, v) = \|u - v\|$ for some norm $\|\cdot\|$. Using the shorthand $\rho(u, S) := \inf_{v \in S} \|u - v\|$, the *Blackwell winner* π^* for an instance $(\mathbf{P}, S, \|\cdot\|)$ is now defined as the one which minimizes the maximum distance to the set S, i.e.,

$$\pi(\mathbf{P}, S, \|\cdot\|) \in \operatorname*{argmin}_{\pi \in \Delta_d} [v(\pi; \mathbf{P}, S, \|\cdot\|)], \quad \text{where} \quad v(\pi; \mathbf{P}, S, \|\cdot\|) := \max_{\pi' \in \Delta_d} \rho(\mathbf{P}(\pi, \pi'), S) \ . \tag{4}$$

Observe that equation (4) has an interpretation as a zero-sum game, where the objective of the minimizing player is to make the score vector $\mathbf{P}(\pi, \pi')$ as close as possible to the target set S.

We now look at commonly studied frameworks for single criterion preference aggregation and multi-objective optimization and show how these can be naturally derived from our framework.

Example: Preference learning along a single criterion. A particular special case of our framework is when we have a single criterion ($k = 1$) and the preferences are given by a matrix $\mathbf{P}_{\mathsf{ov}}$. The score $\mathbf{P}_{\mathsf{ov}}(i_1, i_2)$ is a scalar representing the probability with which object i_1 beats object i_2 in an overall comparison. As a consequence of the von Neumann minimax theorem, we have

$$\max_{\pi_1 \in \Delta_d} \min_{\pi_2 \in \Delta_d} \mathbf{P}_{\mathsf{ov}}(\pi_1, \pi_2) = \min_{\pi_2 \in \Delta_d} \max_{\pi_1 \in \Delta_d} \mathbf{P}_{\mathsf{ov}}(\pi_1, \pi_2) = \frac{1}{2}, \tag{5}$$

with any maximizer above called the von Neumann winner [22]. Thus, for *any* preference matrix $\mathbf{P}_{\mathsf{ov}}$, a von Neumann winner is preferred to any other object with probability at least $\frac{1}{2}$.

Let us show how this uni-criterion formulation can be derived as a special case of our framework. Consider the target set $S = [\frac{1}{2}, 1]$ and choose the distance function $\rho(a, b) = |a - b|$. By equation (5), the target set $S = [\frac{1}{2}, 1]$ is achievable *for all* preference matrices $\mathbf{P}_{\mathsf{ov}}$, and so the von Neumann winner and the Blackwell winner $\pi(\mathbf{P}_{\mathsf{ov}}, [\frac{1}{2}, 1], |\cdot|)$ coincide. ♣

Example: Weighted combinations of a multi-criterion problem. One of the common approaches used in multi-objective optimization to reduce a multi-dimensional problem to a uni-dimensional counterpart is by introducing a weighted combinations of objectives. Formally, consider a weight vector $w \in \Delta_k$ and the corresponding preference matrix $\mathbf{P}(w) := \sum_{j \in [k]} w_j \mathbf{P}^j$ obtained by combining the preference matrices along the different criteria. A winning distribution can then be obtained by solving for the von Neumann winner of $\mathbf{P}(w)$ given by $\pi(\mathbf{P}(w), [\frac{1}{2}, 1], |\cdot|)$. The following proposition establishes that such an approach is a particular special case of our framework.

Proposition 1. *(a) For every weight vector $w \in \Delta_k$, there exists a target set $S_w \in [0, 1]^k$ such that for any norm $\|\cdot\|$, we have*

$$\pi(\mathbf{P}, S_w, \|\cdot\|) = \pi(\mathbf{P}(w), [1/2, 1], |\cdot|) \quad \textit{for all} \quad \mathbf{P} \in \mathcal{P}_{d,k}.$$

(b) Conversely, there exists a set S and a preference tensor $\mathbf{P}$ with a unique *Blackwell winner π^* such that for all $w \in \Delta_k$, exactly one of the following is true:*

$$\pi(\mathbf{P}(w), [1/2, 1], |\cdot|) \neq \pi^* \quad \textit{or} \quad \operatorname*{argmax}_{\pi \in \Delta_d} \min_{i \in [d]} \mathbf{P}(\pi, i) = \Delta_d \ .$$

Thus, while the Blackwell winner is always able to recover any linear combination of criteria, the converse is not true. Specifically, part (b) of the proposition shows that for a choice of preference tensor $\mathbf{P}$ and target set S, either the von Neumann winner for $\mathbf{P}(w)$ is not equal to the Blackwell winner, or it degenerates to the entire simplex Δ_d and is thus uninformative. Consequently, our framework is strictly more general that weighting the individual criteria. ♣

3 Statistical guarantees and computational approaches

In this section, we provide theoretical results on computing the Blackwell winner from samples of pairwise comparisons along the various criteria.

Observation model and evaluation metrics. We operate in the natural passive observation model, where a sample consists of a comparison between two randomly chosen objects along a randomly chosen criterion. Specifically, we assume access to an oracle that when queried with a tuple $\eta = (i_1, i_2, j)$ comprising a pair of objects (i_1, i_2) and a criterion j, returns a comparison $y(\eta) \sim \mathsf{Ber}(p(i_1, i_2; j))$. Each query to the oracle constitutes one sample. In the passive sampling model, the tuple of objects and criterion is sampled uniformly, with replacement, that is $(i_1, i_2) \sim \mathsf{Unif}\{\binom{[d]}{2}\}$ and $j \sim \mathsf{Unif}\{[k]\}$ where $\mathsf{Unif}\{A\}$ denotes the uniform distribution over the elements of a set A. Given access to samples $\{y_1(\eta_1), \ldots, y_n(\eta_n)\}$ from this observation model, we define the empirical preference tensor (specifically the upper triangular part)

$$\widehat{\mathbf{P}}_n(i_1, i_2, j) := \frac{\sum_{\ell=1}^n y_\ell(\eta_\ell) \mathbb{I}[\eta_\ell = (i_1, i_2, j)]}{1 \vee \sum_\ell \mathbb{I}[\eta_\ell = (i_1, i_2, j)]} \quad \text{for } i_1 < i_2 \ , \tag{6}$$

where each entry of the upper-triangular tensor is estimated using a sample average and the remaining entries are calculated to ensure the symmetry relations implied by the inclusion $\widehat{\mathbf{P}}_n \in \mathcal{P}_{d,k}$.

As mentioned before, we are interested in computing the solution $\pi^* := \pi(\mathbf{P}, S, \|\cdot\|)$ to the optimization problem (4), but with access only to samples from the passive observation model. For any estimator $\widehat{\pi} \in \Delta_d$ obtained from these samples, we evaluate its error based on its value with respect to the tensor $\mathbf{P}$, i.e.,

$$\Delta_{\mathbf{P}}(\widehat{\pi}, \pi) := v(\widehat{\pi}; S, \mathbf{P}, \|\cdot\|) - v(\pi^*; S, \mathbf{P}, \|\cdot\|). \tag{7}$$

Note that the error $\Delta_{\mathbf{P}}$ implicitly also depends on the set S and the norm $\|\cdot\|$, but we have chosen our notation to be explicit only in the preference tensor $\mathbf{P}$. For the rest of this section, we restrict our attention to convex target sets S and refer them to as *valid sets*. Having established the background, we are now ready to provide sample complexity bounds on the estimation error $\Delta_{\mathbf{P}}(\widehat{\pi}, \pi^*)$.

3.1 Upper bounds on the error of the plug-in estimator

While, our focus in this section is to provide upper bounds on the error of the plug-in estimator $\widehat{\pi}_{\mathsf{plug}} = \pi(\widehat{\mathbf{P}}, S, \|\cdot\|)$, we first state a general perturbation bound which relates the error of the optimizer $\pi(\widetilde{\mathbf{P}}, S, \|\cdot\|)$ to the deviation of the tensor $\widetilde{\mathbf{P}}$ from the true tensor $\mathbf{P}$. We use $\mathbf{P}(\cdot, i) \in [0,1]^{d\times k}$ to denote a matrix formed by viewing the i-th slice of $\mathbf{P}$ along its second dimension.

Theorem 1. *Suppose the distance ρ is induced by the norm $\|\cdot\|_q$ for some $q \geq 1$. Then for each valid target set S and preference tensor $\widetilde{\mathbf{P}}$, we have*

$$\Delta_{\mathbf{P}}(\pi(\widetilde{\mathbf{P}}), \pi^*) \leq 2 \max_{i\in[d]} \|\widetilde{\mathbf{P}}(\cdot, i) - \mathbf{P}(\cdot, i))\|_{\infty,q}. \tag{8}$$

Note that this theorem is entirely deterministic: it bounds the deviation in the optimal solution to the problem (4) as a function of perturbations to the tensor $\mathbf{P}$. It also applies *uniformly* to all valid target sets S. In particular, this result generalizes the perturbation result of Dudik et al. [22, Lemma 3] which obtained such a deviation bound for the single criterion problem with π^* as the von Neumann winner. Indeed, one can observe that by setting the distance $\rho(u, v) = |u - v|$ in Theorem 1 for the uni-criterion setup, we have the error $\Delta_{\mathbf{P}}(\pi(\widetilde{\mathbf{P}}), \pi^*) \leq 2\|\widetilde{\mathbf{P}} - \mathbf{P}\|_{\infty,\infty}$, matching the bound of [22].

Let us now illustrate a consequence of this theorem by specializing it to the plug-in estimator, and with the distances given by the ℓ_∞ norm.

Corollary 1. *Suppose that the distance ρ is induced by the ℓ_∞-norm $\|\cdot\|_\infty$. Then there exists a universal constant $c > 0$ such that given a sample size $n > cd^2k\log(\frac{cdk}{\delta})$, we have for each valid target set S*

$$\mathbb{E}\left[\Delta_{\mathbf{P}}(\widehat{\pi}_{\mathsf{plug}}, \pi^*)\right] \leq c\sqrt{\frac{d^2k}{n}\log\left(\frac{cdk}{\delta}\right)}, \tag{9}$$

with probability greater than $1 - \delta$.

The bound (9) implies that the plug-in estimator $\widehat{\pi}_{\mathsf{plug}}$ is an ϵ-approximate solution whenever the number of samples scales as $n = \widetilde{O}(\frac{d^2k}{\epsilon^2})$. Observe that this sample complexity scales quadratically in the number of objects d and linearly in the number of criteria k. This scaling represents the effective dimensionality of the problem instance, since the underlying preference tensor $\mathbf{P}$ has $O(d^2k)$ unknown parameters. Notice that the corollary holds for sample size $n = \widetilde{O}(d^2k)$; this should not be thought of as restrictive, since otherwise, the bound (9) is vacuous.

3.2 Information-theoretic lower bounds

While Corollary 1 provides an upper bound on the error of the plug-in estimator that holds *for all* valid target sets S, it is natural to ask if this bounds is sharp, i.e., whether there is indeed a target set S for which one can do no better than the plug-in estimator. In this section, we address this question by providing lower bounds on the minimax risk $\mathfrak{M}_{n,d,k}(S, \|\cdot\|_\infty) := \inf_{\widehat{\pi}} \sup_{\mathbf{P}\in\mathcal{P}} \mathbb{E}\left[\Delta_{\mathbf{P}}(\widehat{\pi}, \pi^*)\right]$, where the infimum is taken over *all* estimators that can be computed from n samples from our observation model. It is important to note that the error $\Delta_{\mathbf{P}}$ is computed using the ℓ_∞ norm and for the set S. Our lower bound will apply to the particular choice of target set $S_0 = [1/2, 1]^k$.

Theorem 2. *There is a universal constant c such that for all $d \geq 4$, $k \geq 2$, and $n \geq cd^4k$, we have*

$$\mathfrak{M}_{n,d,k}(S_0, \|\cdot\|_\infty) \geq c\sqrt{\frac{d^2k}{n}}. \tag{10}$$

Comparing equations and (9) and (10), we see that for the ℓ_∞-norm and the set S_0, we have provided upper and lower bounds that match up to a logarithmic factor in the dimension. Thus, the plug-in estimator is indeed optimal for this pair $(\|\cdot\|_\infty, S_0)$.Further, observe that the above lower bound is non-asymptotic, and holds for all values of $n \gtrsim d^4k$. This condition on the sample size arises as a consequence of the specific packing set used for establishing the lower bound, and improving it is an interesting open problem.

However, this raises the question of whether the set S_0 is special, or alternatively, whether one can obtain an S-dependent lower bound. The following proposition shows that at least *asymptotically*, the sample complexity for *any* polyhedral set S obeys a similar lower bound.

Proposition 2 (Informal)**.** *Suppose that we have a valid polyhedral target set S, and that $d \geq 4$. There exists a positive integer $n_0(d, k, S)$ such that for all $n \geq n_0(d, k, S)$ we have*

$$\mathfrak{M}_{n,d,k}(S, \|\cdot\|_\infty) \gtrsim \sqrt{\frac{d^2k}{n}}\ . \tag{11}$$

We defer the formal statement and proof of this proposition to Appendix B. This proposition establishes that the plugin estimator $\widehat{\pi}_{\mathsf{plug}}$ is indeed asymptotically optimal in the ℓ_∞ norm for broad class of sets S.

3.3 Computing the plug-in estimator

In the last few sections, we discussed the statistical properties of the plug-in estimator, and showed that its sample complexity was optimal in a minimax sense. We now turn to the algorithmic question: how can the plug-in estimator $\widehat{\pi}_{\mathsf{plug}}$ be computed? Our main result in this direction is the following theorem that characterizes properties of the objective function $v(\pi; \mathbf{P}, S, \|\cdot\|)$.

Theorem 3. *Suppose that the distance function is given by an ℓ_q norm $\|\cdot\|_q$ for some $q \geq 1$. Then for each valid target set S, the objective function $v(\pi; \mathbf{P}, S, \|\cdot\|_q)$ is convex in π, and Lipschitz in the ℓ_1 norm, i.e.,*

$$|v(\pi_1; \mathbf{P}, S, \|\cdot\|_q) - v(\pi_2; \mathbf{P}, S, \|\cdot\|_q)| \leq k^{\frac{1}{q}} \cdot \|\pi_1 - \pi_2\|_1 \text{ for each } \pi_1, \pi_2 \in \Delta_d.$$

Theorem 3 establishes that the plug-in estimator can indeed be computed as the solution to a (constrained) convex optimization problem. In Appendix C, we discuss a few specific algorithms based on zeroth-order and first-order methods for obtaining such a solution and an analysis of the corresponding iteration complexity for these methods; see Propositions 5 and 6 in the appendix.

4 Autonomous driving user study

In order to evaluate the proposed framework, we applied it to an autonomous driving environment. The objective is to study properties of the randomized policies obtained by our multi-criteria framework—the Blackwell winner for specific choices of the target set—and compare them with the alternative approaches of linear combinations of criteria and the single-criterion (overall) von Neumann winner. We briefly describe the components of the experiment here; see Appendix D for more details.

Self-driving Environment. Figure 1(a) shows a snapshot of one of the worlds in this environment with the autonomous car shown in orange. We construct three different worlds in this environment:

- W1: The first world comprises an empty stretch of road with no obstacles (20 steps).
- W2: The second world consists of a sequence of cones placed in certain sequences (80 steps).
- W3: The third world has additional cars driving at varying speeds in their fixed lanes (80 steps).

Policies. For our *base policies*, we design five different reward functions encoding different self-driving behaviors. These polices, named Policy A-E, are then set to be the model predictive control based policies based on these reward functions wherein we fix the planning horizon to 6. We defer the details of these reward functions to Appendix D. A *randomized policy* $\pi \in \Delta_5$ is given by a distribution over the base policies A-E. Such a randomized policy is implemented in our environment by randomly sampling a base policy from the mixture distribution after every $H = 18$ time steps and executing this selected policy for that duration. To account for the randomization, we execute each such policy for 5 independent runs in each of the worlds and record these behaviors.

Subjective Criteria. We selected five subjective criteria to compare the policies, with questions asking which of the two policies was C1: Less aggressive, C2: More predictable, C3: More quick, C4: More conservative, and had C5: Less collision risk. Such a framing of question ensures that higher score value along any of C1-C5 is preferred; thus a higher score along C1 would imply less aggressive while along C2 would mean more predictable. In addition to the these base criteria, we also consider an *Overall Preference* which compares any pair of policies in an aggregate manner. Additionally, we also asked the users to rate the importance of each criterion in their overall preference.

Main Hypotheses. The central focus of the main hypotheses is on comparing the randomized policies given by the Blackwell winner, the overall von Neumann winner, and those given by weighing the criteria linearly.

- MH1 There exists a set S such that the Blackwell winner with respect to S and ℓ_∞-norm produced by our framework outperforms the overall von Neumann winner.
- MH2 The Blackwell winner for oblivious score sets S outperforms both oblivious[2] and data-driven weights for linear combination of criteria.

Independent Variables. The independent variable of our experiment is the choice of algorithms for producing the different randomized winners. These comprise the von Neumann winner based on overall comparisons, Blackwell winners based on two oblivious target sets, and 9 different linear combinations weights (3 data-driven and 6 oblivious).

We begin with the two target sets S_1 and S_2 for our evaluation of the Blackwell winner which were selected in a data-oblivious manner. Set S_1 is an axis-aligned set promoting the use of safer policies with score vector constrained to have a larger value along the collision risk axis. Similar to Figure 2(b), the set S_2 adds a linear constraint along aggressiveness and collision risk. This target set thus favors policies which are less aggressive and have lower collision risk. For evaluating hypothesis MH2, we considered several weight vectors, both oblivious and data-dependent, comprising average of the users' self-reported weights, that obtained by regressing the overall criterion on C1-C5, and a set of oblivious weights. See Appendix D for details of the sets S_1 and S_2, and the weights $w_{1:9}$.

Data collection. The experiment was conducted in two phases, both of which involved human subjects on Amazon Mechanical Turk (Mturk) (see Appendix D for an illustration of the questionnaire). The first phase of the experiment involved preference elicitation for the five base policies A-E. Each user was asked to provide comparison data for all ten combinations of policies. The cumulative comparison data is given in Appendix D, and the average weight vector elicited from the users was found to be $w_1 = [0.21, 0.19, 0.20, 0.18, 0.22]$. We ran this study with 50 subjects.

In the overall preference elicitation, we saw an approximate ordering amongst the base policies: $\mathrm{C} \succ \mathrm{E} \succsim \mathrm{D} \succsim \mathrm{B} \succ \mathrm{A}$. Thus, Policy C was the von Neumann winner along the overall criterion. For each of the linear combination weights w_1 through w_9, Policy C was the weighted winner. The Blackwell winners R1 and R2 for the sets S_1 and S_2 with the ℓ_∞ distance were found to be $\mathrm{R1} = [0.09, 0.15, 0.30, 0.15, 0.31]$ and $\mathrm{R2} = [0.01, 0.01, 0.31, 0.02, 0.65]$.

In the second phase, we obtained preferences from a set of 41 subjects comparing the randomized polices R1 and R2 with the baseline policies A-E. The results are aggregated in Table 1 in Appendix D.

Analysis for main hypotheses. Given that the overall von Neumann winner and those corresponding to weights $w_{1:9}$ were all Policy C, hypotheses MH1 and MH2 reduced whether users prefer at least one of $\{\mathrm{R1}, \mathrm{R2}\}$ to the deterministic policy C, that is whether $\mathbf{P}_{\mathsf{ov}}(\mathrm{C}, \mathrm{R1}) < 0.5$ or $\mathbf{P}_{\mathsf{ov}}(\mathrm{C}, \mathrm{R2}) < 0.5$.

[2]We use the term oblivious to denote variables that were *fixed* before the data collection phase and data-driven to denote those which are based on collected data.

Policies C and E were preferred to R1 by 0.71 and 0.61 fraction of the respondents, respectively. On the other hand, R2 was preferred to the von Neumann winner C by 0.66 fraction of the subjects. Using the data, we conducted a hypothesis test with the null and alternative hypotheses given by

$$H_0 : \mathbf{P}_{\mathsf{ov}}(\mathrm{C}, \mathrm{R2}) \geq 0.5, \quad \text{and} \quad H_1 : \mathbf{P}_{\mathsf{ov}}(\mathrm{C}, \mathrm{R2}) < 0.5.$$

Among the hypotheses that make up the (composite) null, our samples have the highest likelihood for the distribution $\mathsf{Ber}(0.5)$. We therefore perform a one-sided hypothesis test with the Binomial distribution with number of samples $n = 41$, success probability $p = 0.5$ and number of successes $x = 14$ (indicating number of subjects which preferred Policy C to R2). The p-value for this test was obtained to be 0.0298. This supports both our claimed hypotheses MH1 and MH2.

5 Discussion and future work

In this paper, we considered the problem of eliciting and learning from preferences along multiple criteria, as a way to obtain rich feedback under weaker assumptions. We introduced the notion of a Blackwell winner, which generalizes many known winning solution concepts. We showed that the Blackwell winner was efficiently computable from samples with a simple and optimal procedure, and also that it outperformed the von Neumann winner in a user study on autonomous driving. Our work raises many interesting follow-up questions: How does the sample complexity vary as a function of the preference tensor $\mathbf{P}$? Can the process of choosing a good target set be automated? What are the analogs of our results in the setting where pairwise comparisons can be elicited *actively*?

Broader impact

An important step towards deploying AI systems in the real world involves aligning their objectives with human values. Examples of such objectives include safety for autonomous vehicles, fairness for recommender systems, and effectiveness of assistive medical devices. Our paper takes a step towards accomplishing this goal by providing a framework to aggregate human preferences along such subjective criteria, which are often hard to encode mathematically. While our framework is quite expressive and allows for non-linear aggregation across criteria, it leaves the choice of the target set in the hands of the designer. As a possible negative consequence, getting this choice wrong could lead to incorrect inferences and unexpected behavior in the real world.

Acknowledgments and Disclosure of Funding

We would like to thank Niladri Chatterji, Robert Kleinberg and Karthik Sridharan for helpful discussions, and Andreea Bobu, Micah Carroll, Lawrence Chan and Gokul Swamy for helping with the user study setup.

AP is supported by a Swiss Re research fellowship at the Simons Institute for the Theory of Computing and KB is supported by a JP Morgan AI Fellowship. This work was partially supported by Office of Naval Research Young Investigator Award and a AFOSR grant to ADD, and by Office of Naval Research Grant DOD ONR-N00014-18-1-2640 to MJW.

Additional revenue: ADD is employed as a consultant at Waymo, LLC and PLB is employed as a consultant at Google.

References

[1] J. Abernethy, P. L. Bartlett, and E. Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 27–46, 2011.

[2] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, 2010.

[3] M. Aghassi and D. Bertsimas. Robust game theory. *Mathematical Programming*, 107(1-2): 231–273, 2006.

[4] N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864, 2014.

[5] A. R. Alimov and I. Tsar'kov. Connectedness and other geometric properties of suns and chebyshev sets. *Fundamentalnaya i Prikladnaya Matematika*, 19(4):21–91, 2014.

[6] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

[7] K. J. Arrow et al. Social choice and individual values. 1951.

[8] V. Balestro, H. Martini, and R. Teixeira. Convex analysis in normed spaces and metric projections onto convex bodies. *arXiv preprint arXiv:1908.08742*, 2019.

[9] E. Bıyık, N. Huynh, M. J. Kochenderfer, and D. Sadigh. Active preference-based gaussian process regression for reward learning. *arXiv preprint arXiv:2005.02575*, 2020.

[10] D. Black. On the rationale of group decision-making. *Journal of political economy*, 56(1): 23–34, 1948.

[11] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

[12] J. d. Borda. Mémoire sur les élections au scrutin. *Histoire de l'Academie Royale des Sciences pour 1781 (Paris, 1784)*, 1784.

[13] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[14] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8, 2015.

[15] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)*, 30(1):1–41, 2012.

[16] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[17] M. d. Condorcet. Essai sur l'application de l'analyse a la probabilité des decisions rendues a la pluralité des voix. 1785.

[18] A. H. Copeland. A reasonable social welfare function. Technical report, mimeo, 1951. University of Michigan, 1951.

[19] K. M. Douglas and R. J. Mislevy. Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics*, 35(3):280–306, 2010.

[20] M. Doumpos and C. Zopounidis. Regularized estimation for preference disaggregation in multiple criteria decision making. *Computational Optimization and Applications*, 38(1):61–80, 2007.

[21] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5), 2015.

[22] M. Dudík, K. Hofmann, R. E. Schapire, A. Slivkins, and M. Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, 2015.

[23] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 2005.

[24] D. Frisch and R. T. Clemen. Beyond expected utility: Rethinking behavioral decision research. *Psychological bulletin*, 116(1):46, 1994.

[25] D. Fudenberg and D. K. Levine. Self-confirming equilibrium. *Econometrica: Journal of the Econometric Society*, pages 523–545, 1993.

[26] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4), 2013.

[27] W. M. Goldstein and J. Beattie. Judgments of relative importance in decision making: The importance of interpretation and the interpretation of importance. In *Frontiers of mathematical psychology*, pages 110–137. Springer, 1991.

[28] K. Hofmann, S. Whiteson, and M. De Rijke. A probabilistic method for inferring preferences from clicks. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 249–258, 2011.

[29] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17), 2008.

[30] K. G. Jamieson, S. Katariya, A. Deshpande, and R. D. Nowak. Sparse dueling bandits. 2015.

[31] V. Kuleshov and S. Ermon. Estimating uncertainty online against an adversary. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[32] E. Lehrer. Partially specified probabilities: Decisions and games. *American Economic Journal: Microeconomics*, 4(1):70–100, 2012.

[33] R. D. Luce. Individual choice behavior. 1959.

[34] S. Mannor, V. Perchet, and G. Stoltz. Approachability in unknown games: Online learning meets multi-objective optimization. In *Conference on Learning Theory*, pages 339–355, 2014.

[35] M. T. McBee, S. J. Peters, and C. Waterman. Combining scores in multiple-criteria assessment systems: The impact of combination rule. *Gifted Child Quarterly*, 58(1):69–89, 2014.

[36] S. Miryoosefi, K. Brantley, H. Daume III, M. Dudik, and R. E. Schapire. Reinforcement learning with convex constraints. In *Advances in Neural Information Processing Systems*, pages 14070–14079, 2019.

[37] H. Moulin. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[38] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2), 2017.

[39] M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928*, 2019.

[40] J. P. Papay. Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48(1):163–193, 2011.

[41] J.-P. Penot and R. Ratsimahalo. Characterizations of metric projections in Banach spaces and applications. In *Abstract and Applied Analysis*, volume 3, 1970.

[42] V. Perchet. Approachability, regret and calibration; implications and equivalences. *arXiv preprint arXiv:1301.2663*, 2013.

[43] V. Perchet. A note on robust nash equilibria with uncertainties. *RAIRO-Operations Research*, 48(3):365–371, 2014.

[44] J.-C. Pomerol and S. Barba-Romero. *Multicriterion decision in management: principles and practice*, volume 25. Springer Science & Business Media, 2012.

[45] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems*, 2017.

[46] W. Saunders, G. Sastry, A. Stuhlmueller, and O. Evans. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2067–2069, 2018.

[47] M. Schulze. A new monotonic, clone-independent, reversal symmetric, and Condorcet-consistent single-winner election method. *Social Choice and Welfare*, 36(2):267–303, 2011.

[48] O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, 2013.

[49] O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1), 2017.

[50] A. Teixeira-Pinto and S.-L. T. Normand. Statistical methodology for classifying units on the basis of multiple-related measures. *Statistics in medicine*, 27(9):1329–1350, 2008.

[51] L. L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[52] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[53] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185 (4157):1124–1131, 1974.

[54] A. Tversky and D. Kahneman. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[55] G. D. Walters. Taking the next step: Combining incrementally valid indicators to improve recidivism prediction. *Assessment*, 18(2):227–233, 2011.

[56] H. Wu and X. Liu. Double thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, pages 649–657, 2016.

[57] Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[58] L. Zajíček. On the fréchet differentiability of distance functions. *Proceedings of the 12th Winter School on Abstract Analysis*, pages 161–165, 1984.

[59] J. Zhang, P. Fiers, K. A. Witte, R. W. Jackson, K. L. Poggensee, C. G. Atkeson, and S. H. Collins. Human-in-the-loop optimization of exoskeleton assistance during walking. *Science*, 356(6344):1280–1284, 2017.

[60] M. Zoghi, S. Whiteson, R. Munos, and M. De Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. *arXiv preprint arXiv:1312.3393*, 2013.

[61] M. Zoghi, Z. S. Karnin, S. Whiteson, and M. De Rijke. Copeland dueling bandits. In *Advances in Neural Information Processing Systems*, pages 307–315, 2015.

[62] M. Zoghi, S. Whiteson, and M. de Rijke. Mergerucb: A method for large-scale online ranker evaluation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 17–26, 2015.

[63] C. Zopounidis and M. Doumpos. Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, 138(2):229–246, 2002.

A Blackwell's approachability

Blackwell [11] introduced the concept of approachability as a generalization of the minimax theorem to vector-valued payoffs. Formally, a Blackwell game is an extension of two-player zero-sum games with vector-valued reward functions.

Let $\mathcal{X}, \mathcal{Y}$ denote the action spaces for the two players and $r : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^k$ be the corresponding vector-valued reward function. Further, let $S \subseteq \mathbb{R}^k$ denote a target set. The objective of player 1 is to ensure that the reward vector r lies in the set S while that of player 2 is ensure that the reward r lies outside this set S. Following [1], we introduce the notion of satisfiability and response-satisfiability.

Definition 1 (Satisfiability)**.** *For a Blackwell game parameterized by $(\mathcal{X}, \mathcal{Y}, r, S)$, we say that,*

- *S is* satisfiable *if there exists $x \in \mathcal{X}$ such that for all $y \in \mathcal{Y}$, we have that $r(x, y) \in S$.*
- *S is* response-satisfiable *if for every $y \in \mathcal{Y}$, there exists $x \in \mathcal{X}$ such that $r(x, y) \in S$.*

In the case of scalar rewards, Von Neumann's minimax theorem indicates that any set which is satisfiable is also response-satsifiable. In other words, there exists a strategy for Player 1, oblivious of Player 2's strategy which ensures that the reward belongs to the set S if the set S is response-satisfiable. The existence of such a relation was crucial in obtaining the concept of the Von Neumann winner described in Section 2 for the uni-criterion problem.

However, such a statement fails to hold in the general vector-valued case (see [1] for a counterexample). In order to overcome this limitation, Blackwell [11] defined the notion of approachability as follows.

Definition 2 (Blackwell's Approachability)**.** *Given a Blackwell game $(\mathcal{X}, \mathcal{Y}, r, S)$, we say that a set S is approachable if there exists an algorithm $\mathcal{A}$ which selects points in $\mathcal{X}$ such that for any sequence $y_1, \dots y_t \in \mathcal{Y}$,*

$$\lim_{T \to \infty} \rho \left(\frac{1}{T} \sum_{t=1}^{T} r(x_t, y_t), S \right) \to 0 \ ,$$

where $x_t = \mathcal{A}(y_1, \dots, y_{t-1})$ is the algorithm's play at time t for some distance function ρ.

The celebrated Blackwell's theorem then claims that any set S is approachable iff it is response-satisfiable. This means that while no single choice of action in the set $\mathcal{X}$ can guarantee a response in the set S, there exists an algorithm which ensures that in the repeated game, the average rewards approach the set S, for any choice of opponent play.

Note that our definition of *achievability* is a stronger requirement than Blackwell's approachability. While approachability requires the time-averaged payoff in a repeated game to belong to the pre-specified set S, achievability requires the same to be true in a single-shot play of the game. Indeed, as the following lemma shows, one can construct examples of multi-criteria preference problems which are approachable but not achievable.

Proposition 3 (Approachability does not imply achievability)**.** *There exists a preference tensor $\mathbf{P} \in \mathcal{P}_{d,k}$ and a target set $S \subset [0, 1]^k$ such that*
a) For the Blackwell game given by $(\Delta_d, \Delta_d, \mathbf{P}, S)$, the set S is approachable, and
b) The set S is not achievable with respect to $\mathbf{P}$.

Proof. We will consider an example in a 2-dimensional action space with 2 criteria. Consider the preference matrix given by:

$$\mathbf{P}^1 = \begin{bmatrix} \frac{1}{2} & 1 \\ 0 & \frac{1}{2} \end{bmatrix} \qquad \text{and} \qquad \mathbf{P}^2 = \begin{bmatrix} \frac{1}{2} & 0 \\ 1 & \frac{1}{2} \end{bmatrix} \ , \tag{12}$$

along with the convex set $S = [\frac{1}{2}, 1]^2$. The tensor $\mathbf{P}$ represents the strongest possible trade-off between the two objects: Object 1 is preferred over 2 along the first criterion while the reverse is true for the second criterion.

The Blackwell game given by $(\Delta_d, \Delta_d, \mathbf{P}, S)$ can indeed be shown to be approachable. The set S is response-satisfiable since for every strategy $y \in \Delta_d$ chosen by the column player, the choice of

$x = y$ would yield a reward vector $\mathbf{P}(x, y) = [\frac{1}{2}, \frac{1}{2}] \in S$. Then, by Blackwell's theorem [11], the set S is approachable.

In contrast, consider any choice of distribution $\pi_1 = [p, 1-p]$ for the multi-criteria preference problem. The corresponding score vectors for responses $i_2 = 1, 2$ are given by:

$$r_1 = \mathbf{P}(\pi_1, i_2 = 1) = \left[\frac{p}{2}, 1 - \frac{p}{2}\right] \qquad \text{and} \qquad r_2 = \mathbf{P}(\pi_1, i_2 = 2) = \left[\frac{1}{2} + \frac{p}{2}, \frac{1}{2} - \frac{p}{2}\right] .$$

For any choice of the parameter $p \in [0, 1]$, one cannot have both r_1 and r_2 simultaneously belong to the set S. Hence, we have that the set S is not achievable with respect to $\mathbf{P}$.

This example can be extended to any arbitrary dimension k by extending the tensor to have $\mathbf{P}^j$ equal to the all-half matrix for any criterion $j > 2$ and the target set to be $S = [\frac{1}{2}, 1]^k$. Similarly, in order to extend the example to any dimension, consider the preference tensor (for $k = 2$)

$$\mathbf{P}_d^1 = \begin{bmatrix} \mathbf{P}^1 & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{1/2} \\ \mathbf{P}_{1/2} & \mathbf{P}^1 & \cdots & \mathbf{P}_{1/2} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{P}_{1/2} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}^1 \end{bmatrix} \quad \text{and} \quad \mathbf{P}_d^2 = \begin{bmatrix} \mathbf{P}^2 & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{1/2} \\ \mathbf{P}_{1/2} & \mathbf{P}^2 & \cdots & \mathbf{P}_{1/2} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{P}_{1/2} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}^2 \end{bmatrix},$$

with the smaller matrices $\mathbf{P}^1$ and $\mathbf{P}^2$ from equation (12) at the diagonal and $\mathbf{P}_{1/2}$ denoting the all-half tensor of the appropriate dimension. A similar calculation as for the $d = 2$ case yields that the set S is not achievable. This establishes the required claim. □

B Proof of main results

In this section, we provide formal proofs of all the results stated in the main paper. Appendix C to follow collects some additional results and their proofs.

B.1 Proof of Proposition 1

We establish both parts of the proposition separately.

B.1.1 Proof of part (a)

For any weight vector $w \in \Delta_k$, consider the set

$$S_w = \left\{ r \in [0, 1]^k \mid \langle w, r \rangle \geq 1/2 \right\}.$$

The set S_w is clearly convex. Indeed, for any two vectors $r_1, r_2 \in S_w$ and any scalar $\alpha \in [0, 1]$, we have

$$\langle w, \alpha r_1 + (1 - \alpha) r_2 \rangle = \alpha \langle w, r_1 \rangle + (1 - \alpha) \langle w, r_2 \rangle \in \left[\frac{1}{2}, 1\right].$$

It is straightforward to verify that the set S_w is also monotonic with respect to the orthant ordering.

We now show that a von Neumann winner π^* of the (single-criterion) preference matrix $\mathbf{P}_w := \mathbf{P}(w)$ can be written as $\pi(\mathbf{P}, S_w, \|\cdot\|)$ for an arbitrary choice of norm $\|\cdot\|$. For each $\widetilde{\pi} \in \Delta_d$, we have

$$\langle w, \mathbf{P}(\pi^*, \widetilde{\pi}) \rangle = \sum_{j \in [k]} w_j \mathbf{P}^j(\pi^*, \widetilde{\pi}) = \mathbf{P}_w(\pi^*, \widetilde{\pi})) \overset{\text{(i)}}{\geq} \frac{1}{2},$$

where the inequality (i) follows since π^* is a von Neumann winner for the matrix $\mathbf{P}_w$. Thus, we have the inclusion $\mathbf{P}(\pi^*, \widetilde{\pi}) \in S_w$ for all $\widetilde{\pi} \in \Delta_d$, so that $\max_{\widetilde{\pi} \in \Delta_d} \rho(\mathbf{P}(\pi^*, \widetilde{\pi}), S_w) = 0$ for any distance metric ρ. Consequently, we have

$$\pi^* \in \operatorname*{argmin}_{\pi \in \Delta_k} \max_{\widetilde{\pi} \in \Delta_d} \rho(\mathbf{P}(\pi, \widetilde{\pi}), S_w),$$

which establishes the claim for part (a). □

B.1.2 Proof of part (b)

Consider the multi-criteria preference instance given by target set $S = [\frac{1}{2}, 1]^k$, the ℓ_∞ distance function and the preference tensor $\mathbf{P}$

$$\mathbf{P}^1 = \begin{bmatrix} \frac{1}{2} & 1 \\ 0 & \frac{1}{2} \end{bmatrix}, \quad \mathbf{P}^2 = \begin{bmatrix} \frac{1}{2} & 0 \\ 1 & \frac{1}{2} \end{bmatrix}, \quad \text{and} \quad \mathbf{P}^j = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

The *unique* Blackwell winner for this instance $(\mathbf{P}, S, \|\cdot\|_\infty)$ is given by

$$\underbrace{\pi(\mathbf{P}, S, \|\cdot\|_\infty)}_{\pi^*} = [1/2, 1/2]\,. \tag{13}$$

For any weight $w \in [0,1]^k$, consider the von Neumann winners corresponding to the weighted matrices $\mathbf{P}_w$

$$\pi(\mathbf{P}_w, [1/2, 1], |\cdot|) = \begin{cases} [1,0] & \text{for } w \text{ s.t. } \mathbf{P}_w(1,2) > 0.5 \\ [0,1] & \text{for } w \text{ s.t. } \mathbf{P}_w(1,2) < 0.5 \\ \pi \in \Delta_2 & \text{otherwise} \end{cases}. \tag{14}$$

Comparing equations (13) and (14) establishes the required claim. □

B.2 Proof of Theorem 1

Let us use the shorthand $\widetilde{\pi} := \pi(\widetilde{\mathbf{P}})$. We begin by decomposing the desired error term as

$$\begin{aligned}
&\Delta_{\mathbf{P}}(\widetilde{\pi}, \pi^*) \\
&= \underbrace{v(\widetilde{\pi}; S, \mathbf{P}, \|\cdot\|) - v(\widetilde{\pi}; S, \widetilde{\mathbf{P}}, \|\cdot\|)}_{\text{Perturbation error at } \widetilde{\pi}} + \underbrace{v(\widetilde{\pi}; S, \widetilde{\mathbf{P}}, \|\cdot\|) - v(\pi^*; S, \widetilde{\mathbf{P}}, \|\cdot\|)}_{\leq 0} + \underbrace{v(\pi^*; S, \widetilde{\mathbf{P}}, \|\cdot\|) - v(\pi^*; S, \mathbf{P}, \|\cdot\|)}_{\text{Perturbation error at } \pi^*}
\end{aligned}$$

In order to obtain a bound on the perturbation errors, note that for any distribution π, we have

$$\begin{aligned}
v(\pi; S, \mathbf{P}, \|\cdot\|) - v(\pi; S, \widetilde{\mathbf{P}}, \|\cdot\|) &= \max_{i_1}[\rho(\mathbf{P}(\pi, i_1), S)] - \max_{i_2}[\rho(\widetilde{\mathbf{P}}(\pi, i_2), S)] \\
&\overset{\text{(i)}}{\leq} \max_{i}[\rho(\mathbf{P}(\pi, i), S) - \rho(\widetilde{\mathbf{P}}(\pi, i), S)],
\end{aligned} \tag{15}$$

where step (i) follows by setting the i_2 equal to i_1. Noting that the distance is given by the ℓ_q norm, we have

$$\begin{aligned}
v(\pi; S, \mathbf{P}, \|\cdot\|) - v(\pi; S, \widetilde{\mathbf{P}}, \|\cdot\|) &\leq \max_i[\min_{z_1 \in S} \|\mathbf{P}(\pi, i) - z_1\|_q - \min_{z_2 \in S} \|\widetilde{\mathbf{P}}(\pi, i) - z_2\|_q] \\
&\overset{\text{(i)}}{\leq} \max_i[\|\mathbf{P}(\pi, i) - \widetilde{\mathbf{P}}(\pi, i)\|_q],
\end{aligned}$$

where the inequality (i) follows by setting z_2 equal to z_1. Taking a supremum over all distributions π completes the proof. □

B.3 Proof of Corollary 1

By Theorem 1, it suffices to provide a bound on the quantity $\max_i \|\mathbf{P}(\cdot, i) - \widehat{\mathbf{P}}(\cdot, i))\|_{\infty,\infty}$ for the plug-in preference tensor $\widehat{\mathbf{P}}$. Now by definition, we have

$$\max_i \|\mathbf{P}(\cdot, i) - \widehat{\mathbf{P}}(\cdot, i))\|_{\infty,\infty} = \max_{i_1, i_2, j} |\mathbf{P}^j(i_1, i_2) - \widehat{\mathbf{P}}^j(i_1, i_2)|\,.$$

For each $i = (i_1, i_2, j)$ representing some index of the tensor, let $N_i := \#\{\ell \mid \eta_\ell = i\}$ denote the number of samples observed at that index. Since N_i can be written as a sum of i.i.d. Bernoulli random variables, applying the Hoeffding bound yields

$$\Pr\left\{\left|N_i - \frac{n}{d^2 k}\right| \geq c\sqrt{\frac{n \log(c/\delta)}{d^2 k}}\right\} \leq \delta \text{ for each } \delta \in (0,1).$$

Note that we also have $n \geq c_0 d^2 k \log(c_1 d/\delta)$ by assumption. For a large enough choice of the constants (c_0, c_1), applying the union bound yields the sequence of sandwich relations

$$\frac{n}{2d^2k} \leq N_i \leq \frac{3n}{2d^2k} \quad \text{for all indices } i \text{ with probability greater than } 1-\delta. \tag{16}$$

Furthermore, conditioned on N_i (for $i = (i_1, i_2, j)$), the Hoeffding bound yields the relation

$$\Pr\left\{|\mathbf{P}^j(i_1,i_2) - \widehat{\mathbf{P}}^j(i_1,i_2)| \geq c\sqrt{\frac{\log(c/\delta)}{N_i}}\right\} \leq \delta \text{ for each } \delta \in (0,1).$$

Putting this together with a union bound, we have

$$\Pr\left\{\max_{i_1,i_2,j} |\mathbf{P}^j(i_1,i_2) - \widehat{\mathbf{P}}^j(i_1,i_2)| \geq c\sqrt{\frac{\log(cd^2k/\delta)}{\min_i N_i}}\right\} \leq \delta. \tag{17}$$

Combining inequalities (16) and (17) with a final union bound completes the proof. □

B.4 Proof of Theorem 2

Suppose throughout that $k \geq 2$, and recall the axis-aligned convex target set $S_0 = [\frac{1}{2}, 1]^k$. We split our proof into two cases depending on whether d is even or odd.

Case 1: d even. We use Le Cam's method and construct two problem instances with preference tensors given by $\mathbf{P}_0$ and $\mathbf{P}_1$. Two key elements in the construction are the following $2 \times 2 \times 2$ tensors, which we denote by $\mathbf{P}_{\mathsf{cr}}$ and $\widetilde{\mathbf{P}_{\mathsf{cr}}}$, respectively. Their entries are given by

$$\mathbf{P}_{\mathsf{cr}}^1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}+\gamma \\ \frac{1}{2}-\gamma & \frac{1}{2} \end{bmatrix} \quad , \quad \mathbf{P}_{\mathsf{cr}}^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}-\gamma \\ \frac{1}{2}+\gamma & \frac{1}{2} \end{bmatrix},$$

$$\widetilde{\mathbf{P}_{\mathsf{cr}}}^1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}+\frac{\gamma}{d} \\ \frac{1}{2}-\frac{\gamma}{d} & \frac{1}{2} \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{P}_{\mathsf{cr}}}^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}-\frac{\gamma}{d} \\ \frac{1}{2}+\frac{\gamma}{d} & \frac{1}{2} \end{bmatrix}.$$

Note that these tensors are parameterized by a scalar $\gamma \in [0, 1/2]$, whose exact value we specify shortly. Also denote by $\mathbf{P}_{1/2}$ the $2 \times 2 \times 2$ all-half tensor. We are now ready to construct the pair of $d \times d \times k$ preference tensors $(\mathbf{P}_0, \mathbf{P}_1)$.

In order to construct tensor $\mathbf{P}_0$, we specify its entries on the first two criteria according to

$$\mathbf{P}_0^{1:2} = \begin{bmatrix} \mathbf{P}_{1/2} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{1/2} \\ \mathbf{P}_{1/2} & \mathbf{P}_{\mathsf{cr}} & \cdots & \mathbf{P}_{1/2} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{P}_{1/2} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{\mathsf{cr}} \end{bmatrix}, \tag{18}$$

and set the entries on the remaining $k-2$ criteria to $1/2$.

On the other hand, the first two criteria of the tensor $\mathbf{P}_1$ are given by

$$\mathbf{P}_1^{1:2} = \begin{bmatrix} \widetilde{\mathbf{P}_{\mathsf{cr}}} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{1/2} \\ \mathbf{P}_{1/2} & \mathbf{P}_{\mathsf{cr}} & \cdots & \mathbf{P}_{1/2} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{P}_{1/2} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{\mathsf{cr}} \end{bmatrix}, \tag{19}$$

with the entries on the remaining $k-2$ criteria once again set identically to $1/2$.

Note that the tensors $\mathbf{P}_0$ and $\mathbf{P}_1$ only differ on the first $2 \times 2 \times 2$ block. Furthermore, the following lemma provides an exact calculation of the values $\min_\pi v(\pi; \mathbf{P}_0, S_0, \|\cdot\|_\infty)$ and $\min_\pi v(\pi; \mathbf{P}_1, S_0, \|\cdot\|_\infty)$.

Lemma 1. *We have*

$$\mathcal{V}_0 := \min_{\pi} v(\pi; \mathbf{P}_0, S_0, \|\cdot\|_\infty) = 0 \quad \textit{and} \quad \mathcal{V}_1 := \min_{\pi} v(\pi; \mathbf{P}_0, S_0, \|\cdot\|_\infty) = \frac{\gamma}{3d-2}.$$

Given samples from these two instances, we now use Le Cam's lemma [see 52, Chap 2] to lower bound the minimax risk as

$$\mathfrak{M}_{n,d,k}(S_0, \|\cdot\|_\infty) \geq \frac{|\mathcal{V}_0 - \mathcal{V}_1|}{2}\left(1 - \|\mathbb{P}_0^n - \mathbb{P}_1^n\|_{\text{TV}}\right) = \frac{\gamma}{2(3d-2)}\left(1 - \|\mathbb{P}_0^n - \mathbb{P}_1^n\|_{\text{TV}}\right), \quad (20)$$

where $\mathbb{P}_0^n$ and $\mathbb{P}_1^n$ are the probability distributions induced on sample space by the passive sampling strategy applied to the tensor $\mathbf{P}_0$ and $\mathbf{P}_1$, respectively.

Using Pinsker's inequality, the decoupling property for KL divergence and the fact that that $\text{KL}(P\|Q) \leq \chi^2(P\|Q)$, we have

$$\|\mathbb{P}_0^n - \mathbb{P}_1^n\|_{\text{TV}} \leq \sqrt{\frac{n}{2}\text{KL}(\mathbb{P}_1\|\mathbb{P}_0)} \leq \sqrt{\frac{n}{2}\chi^2(\mathbb{P}_1\|\mathbb{P}_0)}\,. \quad (21)$$

The chi-squared distance between the two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ is given by

$$\chi^2(\mathbb{P}_1\|\mathbb{P}_0) = \frac{1}{d^2k}\sum_{(i_1,i_2,j)}\left(\frac{\mathbf{P}_1^j(i_1,i_2)}{\mathbf{P}_2^j(i_1,i_2)} - 1\right)^2 \overset{\text{(i)}}{=} \frac{2}{d^2k}\left(\left(\frac{2\gamma}{d}\right)^2 + \left(-\frac{2\gamma}{d}\right)^2\right) = \frac{16\gamma^2}{d^4k}\,,$$

where step (i) follows from the fact that $\mathbf{P}_1$ and $\mathbf{P}_2$ differ only in 4 entries and that the passive sampling strategy samples each index uniformly at random. Putting together the pieces, we have:

$$\mathfrak{M}_{n,d,k}(S_0, \|\cdot\|_\infty) \geq \frac{\gamma}{2(3d-2)}\left(1 - \sqrt{\frac{n}{2}\,\frac{16\gamma^2}{d^4k}}\right) \overset{\text{(ii)}}{=} \frac{1}{48\sqrt{2}}\sqrt{\frac{d^2k}{n}}.$$

where step (ii) follows by setting $\gamma^2 = \frac{d^4k}{32n}$ and using the fact that $3d-2 \leq 3d$. Note that since we require $\gamma^2 \leq \frac{1}{4}$, the above bound is valid only for $n \gtrsim d^4k$. This concludes the proof for even d.

Case 2: d odd. By assumption, we have $d \geq 5$. In this case, we construct $\mathbf{P}_0$ and $\mathbf{P}_1$ exactly as before, but replace $\mathbf{P}_{\text{cr}}$ in the last two rows of both $\mathbf{P}_0$ and $\mathbf{P}_1$ with the following modified $3 \times 3 \times 2$ tensor:

$$\mathbf{P}^1_{\text{cr},3} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}+\gamma & \frac{1}{2}-\gamma \\ \frac{1}{2}-\gamma & \frac{1}{2} & \frac{1}{2}-\gamma \\ \frac{1}{2}+\gamma & \frac{1}{2}+\gamma & \frac{1}{2} \end{bmatrix} \quad \text{and} \quad \mathbf{P}^2_{\text{cr},3} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}-\gamma & \frac{1}{2}+\gamma \\ \frac{1}{2}+\gamma & \frac{1}{2} & \frac{1}{2}+\gamma \\ \frac{1}{2}-\gamma & \frac{1}{2}-\gamma & \frac{1}{2} \end{bmatrix}.$$

By mimicking its proof, it can be verified that this modification ensures that the corresponding values $\mathcal{V}_0$ and $\mathcal{V}_1$ still satisfy Lemma 1. Thus, the lower bound remains unchanged up to constant factors. □

B.4.1 Proof of Lemma 1

Let us compute the two values separately.

Computing $\mathcal{V}_0$. The choice of distribution $\pi^* = [1, 0, \ldots, 0]$ yields the score vector $[1/2, 1/2, \ldots, 1/2]$, which is in the set S_0. Thus, we have $\mathcal{V}_0 = 0$.

Computing $\mathcal{V}_1$. Note that the optimal distribution π^* achieving the value $\mathcal{V}_1$ will be of the form

$$\pi^* = [p/2, p/2, (1-p)/(d-2), \ldots, (1-p)/(d-2)] \text{ for some } p \in [0,1].$$

This follows from the symmetry in the preference tensor for row objects ranging from 3 to d. Given such a distribution π^*, the distance of the reward vector from the set S_0 is given by

$$\inf_{z \in S}\|\mathbf{P}(\pi^*, i_2) - z\|_\infty = \begin{cases} \frac{\gamma p}{2d} & i_2 = 1, 2 \\ \frac{\gamma(1-p)}{d-2} & \text{o.w.} \end{cases}.$$

Thus, for any value of $p > 2d/(3d-2)$, the distance is maximized for $i_2 \in \{1, 2\}$, and yields a value $\gamma p/(2d)$. On the other hand, for $p < 2d/(3d-2)$, the maximizing index is $i_2 \geq 3$, and the maximizing value is $\gamma(1-p)/(d-2)$. Optimizing these values for p yields the claim. □

B.5 Instance dependent lower bounds

In this section, we give a formal statement of Proposition 2 along with its proof.

We begin by defining some notation. For any $\alpha, \beta \in [-\frac{1}{2}, \frac{1}{2}]$ and choice of criteria $j_1, j_2 \in [k]$, we define the tensor $\mathbf{P}_{\alpha,\beta}^{(j_1,j_2)} \in [0,1]^{2\times 2\times k}$ as

$$\mathbf{P}_{\alpha,\beta}^{j_1} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}+\alpha \\ \frac{1}{2}-\alpha & \frac{1}{2} \end{bmatrix}, \quad \mathbf{P}_{\alpha,\beta}^{j_2} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2}+\beta \\ \frac{1}{2}-\beta & \frac{1}{2} \end{bmatrix} \quad \text{and} \quad \mathbf{P}_{\alpha,\beta}^{j} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \text{ for } j \neq \{j_1, j_2\} .$$

Further, we denote by $\mathbf{P}_{1/2}$ the all-half tensor whose dimensions may vary depending on the context. Any distribution π over the two objects can be parameterized by a value $q \in [0,1]$ with q being the probability placed on the first object and $1-q$ the probability on the second object. We will consider the distance function given by the ℓ_∞ norm. Given this distance function, we overload our notation for the value

$$v(q; \mathbf{P}_{\alpha,\beta}^{(j_1,j_2)}, S) = \max_i [\rho(\mathbf{P}_{\alpha,\beta}^{(j_1,j_2)}(q,i), S)] \quad \text{and} \quad \mathcal{V}(\mathbf{P}_{\alpha,\beta}^{(j_1,j_2)}; S) = \min_q v(q; \mathbf{P}_{\alpha,\beta}^{(j_1,j_2)}; S) . \tag{22}$$

We now state our main assumption for the score set S which allows us to formulate our lower bound.

Assumption 1. *There exists a pair of criteria (j_1, j_2), values $\alpha_0 \in (0, \frac{1}{2}]$ and $\beta_0 \in [-\frac{1}{2}, 0]$, and a gap parameter $\gamma > 0$ such that*

$$\mathcal{V}(\mathbf{P}_{1/2}; S) + \gamma \leq \mathcal{V}(\mathbf{P}_{\alpha_0,\beta_0}^{(j_1,j_2)}; S)$$

for the all-half tensor $\mathbf{P}_{1/2} \in [0,1]^{2\times 2\times k}$.

The assumption above indicates that there exists a pair of criteria along which one can observe some sort of trade-off when they interact with the underlying score set S. The preference tensor $\mathbf{P}_{\alpha_0,\beta_0}^{(j_1,j_2)}$ captures this trade-off and the gap parameter γ quantifies it. Going forward, we assume without loss of generality that $(j_1, j_2) = (1, 2)$ and drop the dependence of the tensor on these indices, writing $\mathbf{P}_{\alpha_0,\beta_0} \equiv \mathbf{P}_{\alpha_0,\beta_0}^{(1,2)}$. The following lemma indicates the importance of the special values of $(\alpha, \beta) = (0, 0)$ for which $\mathbf{P}_{0,0} = \mathbf{P}_{1/2}$.

Lemma 2. *For any $\alpha, \beta \in [-\frac{1}{2}, \frac{1}{2}]$, we have $\mathcal{V}(\mathbf{P}_{0,0}; S) \leq \mathcal{V}(\mathbf{P}_{\alpha,\beta}; S)$.*

The above lemma establishes that for any set, the value attained by setting $(\alpha_0, \beta) = (0, 0)$ will be lower than any other setting of the same parameters. For any parameter $\delta \in [0,1]$, denote by $\mathbf{P}_{\mathsf{wt},\delta}$ the weighted tensor

$$\mathbf{P}_{\mathsf{wt},\delta} := (1-\delta)\mathbf{P}_{0,0} + \delta \mathbf{P}_{\alpha_0,\beta_0}.$$

In order to understand the value $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta}; S)$, we establish the following structural lemma which gives us insight into how this value varies as a function of the parameter $\delta \in [0,1]$.

Lemma 3. *Consider a target set S that is given by an intersection of h half-spaces. Then, the value function $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta}; S)$ is a piece-wise linear and continuous function of $\delta \in [0,1]$ with at most $4h$ pieces.*

The above lemma states that the value $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta}; S)$ is a piece-wise linear function of δ. Consider the first such piece which has a non-zero slope. Such a line has to exist since $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta})$ is continuous in δ and we have $\mathcal{V}(\mathbf{P}_{\mathsf{wt},0}) < \mathcal{V}(\mathbf{P}_{\mathsf{wt},1})$. Also, this slope has to be positive since we know from Lemma 2 that $\mathcal{V}(\mathbf{P}_{\mathsf{wt},0}) \leq \mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta})$ for any $\delta \in [0,1]$. Denote the starting point of this line by δ_0 and the corresponding slope by m_0, and observe that the value $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta_0}) = \mathcal{V}(\mathbf{P}_{\mathsf{wt},0})$. With this notation, we now proceed to prove the lower bound on sample complexity for any polyhedral target score set S.

Proposition 4 (Formal)**.** *Suppose that we have a valid polyhedral target set S satisfying Assumption 1 with parameters (α_0, β_0). Then, there exists a universal constant c such that for all $d \geq 4$, $k \geq 2$, and $n \geq \frac{d^2 k}{\delta^2} \frac{(1/2-\delta_0\alpha_0)^2}{\alpha_0^2+\beta_0^2}$, we have*

$$\mathfrak{M}_{n,d,k}(S, \|\cdot\|_\infty) \geq c \frac{m_0(\frac{1}{2}-\delta_0\alpha_0)}{\sqrt{\alpha_0^2+\beta_0^2}} \sqrt{\frac{d^2 k}{n}} . \tag{23}$$

Proof. For this proof, we focus on the case when the number of criteria k is even. The proof for the case when k is odd can be obtained similar to the proof of Theorem 2.

We use Le Cam's method for obtaining a lower bound on the minimax value and construct the lower bound instances using the tensor given by $\mathbf{P}_{\mathsf{wt},\delta}$. For some $\delta \in [0,1]$ (to be fixed later), consider the parameter $\delta_1 = \delta_0 + \delta$. Using these values of δ_0 and δ_1, we create the following two instances $\mathbf{P}_0$ and $\mathbf{P}_1$:

$$\mathbf{P}_0 = \begin{bmatrix} \mathbf{P}_{\mathsf{wt},\delta_0} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{1/2} \\ \mathbf{P}_{1/2} & \mathbf{P}_{\alpha_0,\beta_0} & \cdots & \mathbf{P}_{1/2} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{P}_{1/2} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{\alpha_0,\beta_0} \end{bmatrix} \quad \text{and} \quad \mathbf{P}_1 = \begin{bmatrix} \mathbf{P}_{\mathsf{wt},\delta_1} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{1/2} \\ \mathbf{P}_{1/2} & \mathbf{P}_{\alpha_0,\beta_0} & \cdots & \mathbf{P}_{1/2} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{P}_{1/2} & \mathbf{P}_{1/2} & \cdots & \mathbf{P}_{\alpha_0,\beta_0} \end{bmatrix},$$

where $\mathbf{P}_{\alpha_0,\beta_0}$ is as given by Assumption 1. The following lemma now shows that that there exists a small enough $\bar{\delta}$ such that the value function $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta}; S)$ is linear in the range $\delta \in [\delta_0, \delta_1]$.

Lemma 4. *There exists a $\bar{\delta} \in (0,1)$ such that for all $\delta \in [0, \bar{\delta}]$ and $\delta_1 = \delta_0 + \delta$, we have*

a. The value $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta_1}; S) = \mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta_0}; S) + \delta m_0$.

b. The minimizer π_1^* for $\mathbf{P}_1^*$ is given by $\pi_1^* = [q_0, 1-q_0, 0 \ldots, 0]$.

We defer the proof of this lemma to the end of the section. Thus, for a small enough value of $\delta \in [0, \bar{\delta}]$, we have $|\mathcal{V}(\mathbf{P}_0) - \mathcal{V}(\mathbf{P}_1)| = \delta m_0$. As was shown in the proof of Theorem 2, the minimax rate is lower bounded as

$$\mathfrak{M}_{n,d,k}(S, \|\cdot\|_\infty) \geq \frac{|\mathcal{V}(\mathbf{P}_0) - \mathcal{V}(\mathbf{P}_1)|}{2} \left(1 - \|\mathbb{P}_0^n - \mathbb{P}_1^n\|_{\text{TV}}\right) \geq \frac{\delta m_0}{2} \left(1 - \sqrt{\frac{n}{2} \chi^2(\mathbb{P}_1 \| \mathbb{P}_0)}\right), \tag{24}$$

where $\mathbb{P}_0^n$ and $\mathbb{P}_1^n$ are the probability distributions induced on sample space by the passive sampling strategy and the preference tensor $\mathbf{P}_0$ and $\mathbf{P}_1$ respectively. In order to obtain the requisite lower bound, we proceed to compute an upper bound on the chi-squared distance between the two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ as

$$\begin{aligned} \chi^2(\mathbb{P}_1 \| \mathbb{P}_0) &= \frac{1}{d^2 k} \sum_{(i_1, i_2, j)} \left(\frac{\mathbf{P}_1^j(i_1, i_2)}{\mathbf{P}_0^j(i_1, i_2)} - 1 \right)^2 \\ &\overset{\text{(i)}}{\leq} \frac{2}{d^2 k} \left(\left(\frac{\alpha_0^2 \delta^2}{(\frac{1}{2} - \delta_0 \alpha_0)^2} \right) + \left(\frac{\beta_0^2 \delta^2}{(\frac{1}{2} + \delta_0 \beta_0)^2} \right) \right) \\ &\overset{\text{(ii)}}{\leq} \frac{2\delta^2}{d^2 k} \left(\frac{\alpha_0^2 + \beta_0^2}{(\frac{1}{2} - \delta_0 \alpha_0)^2} \right), \end{aligned}$$

where (i) follows from the fact that the instances $\mathbf{P}_0$ and $\mathbf{P}_1$ differ only in 4 entries and (ii) follows from the assumption that $|\alpha_0| \geq |\beta_0|$. Now, substituting the value of $\delta^2 = \frac{d^2 k}{4n} \cdot \frac{(\frac{1}{2} - \delta_0 \alpha_0)^2}{\alpha_0^2 + \beta_0^2}$ and using the above bound with equation (24), we have

$$\mathfrak{M}_{n,d,k}(S, \|\cdot\|_\infty) \geq \frac{m_0 (\frac{1}{2} - \delta_0 \alpha_0)}{8\sqrt{\alpha_0^2 + \beta_0^2}} \sqrt{\frac{d^2 k}{n}},$$

which holds whenever we have $\delta \in [0, \bar{\delta}]$ or equivalently $n \geq \frac{d^2 k}{4\delta^2} \frac{(\frac{1}{2} - \delta_0 \alpha_0)^2}{\alpha_0^2 + \beta_0^2}$. This establishes the desired claim. □

B.5.1 Proof of Lemma 2

For any $\alpha, \beta \in [-\frac{1}{2}, \frac{1}{2}]$, consider the value

$$\begin{aligned}\mathcal{V}(\mathbf{P}_{\alpha,\beta}; S) &= \min_{q\in[0,1]} \max_{i} [\rho(\mathbf{P}_{\alpha,\beta}(q, i), S)] \\ &= \min_{q\in[0,1]} \max_{\tau\in[0,1]} [\rho(\mathbf{P}_{\alpha,\beta}(q, \tau), S)] \\ &\overset{\text{(i)}}{\geq} \rho\left(\left[\frac{1}{2}\right]^k, S\right) = \mathcal{V}(\mathbf{P}_{1/2}; S)\ ,\end{aligned}$$

where (i) follows by setting $\tau = q$ and $\left[\frac{1}{2}\right]^k$ denotes the vector with each entry set to half. This establishes the claim. □

B.5.2 Proof of Lemma 3

Let us denote by q_0 any minimizer of the value $v(q; \mathbf{P}_{\alpha_0,\beta_0}, S)$ and the two score vectors corresponding to the choices for i in equation (22) by $z_{1,i} := \mathbf{P}_{\alpha_0,\beta_0}(q_0, i)$. Observe that for $\mathbf{P}_{\mathsf{wt},\delta}$, the distribution given by q_0 is still a minimizer of its value. Further, the score vectors for the two column choices are given by:

$$z_{\delta,i} = (1-\delta)\left[\frac{1}{2}\right]^k + \delta z_{1,i} \quad \text{for } i = \{1, 2\}.$$

Recall that the distance function is given by $\rho(z_{\delta,i}, S) = \min_{z\in S} \|z_{\delta,i} - z\|_\infty$. Now, the minimizer z will lie on the closest hyperplane(s) to the point $z_{\delta,i}$. In order to establish the claim, it suffices to show that for any fixed hyperplane[3] H, the distance function given by $\rho(z_{\delta,i}, H)$ is a piece-wise linear function for $\delta \in [0, 1]$.

Let us consider a point $z_{\delta,i}$ which does not belong to the half-space given by H, since otherwise, the distance to the halfspace is 0. If we have $\rho(z_{\delta,i}, H) = \zeta$, then the vector $z_{\delta,i} + \zeta 1_k$ must lie on the hyperplane H. This follows from the monotonicity property of the hyperplane H.

For any $\delta = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ such that $z_{\delta_1,i}$ and $z_{\delta_2,i}$ do not belong to the half-space given by H, we have

$$\rho(z_{\delta,i}) = \frac{1}{2}\underbrace{\rho(z_{\delta_1,i})}_{\zeta_1} + \frac{1}{2}\underbrace{\rho(z_{\delta_2,i})}_{\zeta_2}\ ,$$

where the above equality follows since $z_{\delta_1,i} + \zeta_1 1_k$ and $z_{\delta_2,i} + \zeta_2 1_k$ both lie on the hyperplane H and therefore $z_{\delta,i} + \frac{\zeta_1+\zeta_2}{2} 1_k$ also lies on the hyperplane. Combined with the fact that for any point $z_{\delta,i}$ which lies in the half-space given by H, the distance $\rho(z_{\delta,i}, H) = 0$, we have that the function $\rho(z_{\delta,i}, H)$ is a piece-wise linear function with at most 2 linear pieces for $\delta \in [0, 1]$.

Since $\rho(z_{\delta,i}, S)$ is a minimum over h hyperplanes, this function is itself a piece-wise linear function with at most $2h$ pieces. The desired claim now follows from noting that the value function $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta}; S)$is a maximum over two piece-wise linear functions each with at most $2h$ pieces. □

B.5.3 Proof of Lemma 4

Consider $\delta_1 = \delta_0 + \delta$ such that δ_0 and δ_1 share the same linear piece. This can be guaranteed to hold true for all $\delta \leq \bar{\delta}_1$ by the piecewise linear nature of the value $\mathcal{V}(\mathbf{P}_{\mathsf{wt},\delta})$.

For part (b) of the claim, let us consider the tensor $\tilde{\mathbf{P}} = \mathbf{P}_1(3:, 3:)$ formed by removing the first two rows and columns from the tensor $\mathbf{P}_1$. Then, from Assumption 1, we have that $\mathcal{V}(\tilde{\mathbf{P}}; S) \geq \mathcal{V}(\mathbf{P}_{1/2}; S) + \tilde{\gamma}$ for some $\tilde{\gamma} > 0$. Selecting a value of $\bar{\delta}_2$ such that $\bar{\delta}_2 m_0 \leq \tilde{\gamma}$, we can ensure that condition (b.) is satisfied.

Finally, setting $\bar{\delta} = \min(\bar{\delta}_1, \bar{\delta}_2)$ completes the proof. □

B.6 Proof of Theorem 3

Let us prove the two claims of the theorem separately. We use the shorthand $v(\pi) := v(\pi; \mathbf{P}, S, \|\cdot\|)$ for convenience.

[3] we use the hyperplane H and the half-space induced by it interchangeably.

Establishing convexity. Consider any two distributions $\pi_1, \pi_2 \in \Delta_k$ and a scalar $\alpha \in [0,1]$. Since the set S is closed and convex, we have

$$\begin{aligned}
v(\alpha\pi_1 + (1-\alpha)\pi_2) &= \max_{i\in[d]} \min_{z\in S} [\rho(\mathbf{P}(\alpha\pi_1 + (1-\alpha)\pi_2, i), z)] \\
&\overset{\text{(i)}}{=} \max_{i\in[d]} \min_{z_1,z_2\in S} [\rho(\alpha\mathbf{P}(\pi_1, i) + (1-\alpha)\mathbf{P}(\pi_2, i), \alpha z_1 + (1-\alpha) z_2)] \\
&\overset{\text{(ii)}}{\leq} \max_{i\in[d]} \left(\alpha \cdot \min_{z_1\in S} [\rho(\mathbf{P}(\pi_1, i), z_1)] + (1-\alpha) \cdot \min_{z_2\in S} [\rho(\mathbf{P}(\pi_2, i), z_2)] \right) \\
&\overset{\text{(iii)}}{\leq} \alpha v(\pi_1) + (1-\alpha) v(\pi_2) \ ,
\end{aligned}$$

where (i) follows from the convexity of S and linearity of the preference evaluation (Eq. (2)), (ii) follows from the convexity of the distance function given by ℓ_q norm and (iii) follows from distributing the max over the two terms. This establishes the first part of the theorem.

Establishing the Lipschitz bound. Consider any two distributions $\pi_1, \pi_2 \in \Delta_d$. The difference in their value function can then be upper bounded as

$$\begin{aligned}
|v(\pi_1) - v(\pi_2)| &= |\max_{i_1\in[d]} [\rho(\mathbf{P}(\pi_1, i_1), S)] - \max_{i_2\in[d]} [\rho(\mathbf{P}(\pi_2, i_2), S)]| \\
&\overset{\text{(i)}}{\leq} \max_{i\in[d]} |\rho(\mathbf{P}(\pi_1, i), S) - \rho(\mathbf{P}(\pi_2, i), S)| \\
&= \max_{i\in[d]} |\min_{z_1\in S} \rho(\mathbf{P}(\pi_1, i), z_1) - \min_{z_2\in S} \rho(\mathbf{P}(\pi_2, i), z_2)| \\
&\overset{\text{(ii)}}{\leq} \max_{i\in[d]} \max_{z\in S} |\rho(\mathbf{P}(\pi_1, i), z) - \rho(\mathbf{P}(\pi_2, i), z)| \ ,
\end{aligned}$$

where (i) follows from using the inequality $|\max_x f(x) - \max_y g(y)| \leq \max_x |f(x) - g(x)|$ and (ii) follows through a similar inequality $|\min_x f(x) - \min_y g(y)| \leq \max_x |f(x) - g(x)|$. Since the distance function ρ is specified by the ℓ_q norm $\|\cdot\|_q$, we have

$$\begin{aligned}
|v(\pi_1) - v(\pi_2)| &\leq \max_{i\in[d]} \|\mathbf{P}(\pi_1, i) - \mathbf{P}(\pi_2, i)\|_q \\
&= \left[\sum_{j=1}^{k} \left(\langle \pi_1 - \pi_2, \mathbf{P}^j(\cdot, i) \rangle \right)^q \right]^{\frac{1}{q}} \\
&\overset{\text{(i)}}{\leq} k^{\frac{1}{q}} \cdot \|\pi_1 - \pi_2\|_1 \ ,
\end{aligned}$$

where (i) follows from an application of Hölder's inequality ($\ell_1 - \ell_\infty$) to the inner product $\langle \pi_1 - \pi_2, \mathbf{P}^j(\cdot, i) \rangle$ and the fact that $\mathbf{P}^j(i_1, i_2) \in [0,1]$ for any (i_1, i_2, j). This establishes the Lipschitz bound and concludes the proof of the theorem. □

C Additional results and their proofs

This section covers additional sample complexity results as well as optimization algorithms for finding the Blackwell winner of a multi-criteria preference learning instance.

C.1 Sample complexity bounds for ℓ_1 norm

Corollary 2. *Suppose that the distance ρ is induced by the ℓ_1 norm $\|\cdot\|_1$. Then there exists a universal constant $c > 0$ such that given a sample size $n > cd^2k \log(\frac{cdk}{\delta})$, we have for each valid target set S*

$$\Delta_{\mathbf{P}}(\widehat{\pi}_{\mathsf{plug}}, \pi^*) \leq ck \sqrt{\frac{d^2 k}{n} \log\left(\frac{cdk}{\delta}\right)} \tag{25}$$

with probability exceeding $1 - \delta$.

Algorithm 1: Zeroth-order method for multi-criteria preference learning

Input: Time steps T, step size η, smoothing radius δ
Initialize: $\theta_1 = 0$
for $t = 1, \ldots, T$ **do**
  $\pi_t = \text{argmax}_{\pi \in \Delta_d} \langle \theta_t, \pi \rangle - r(\pi)$ where $r(\pi) = \sum_i \pi_i \log(\pi_i)$
  Sample u_t uniformly from the Euclidean unit sphere $\{u \mid \|u\|_2 = 1\}$
  For every $i \in [d]$, query points $z_{1,i} = \mathbf{P}(\pi_t + \delta u_t, i)$ and $z_{2,i} = \mathbf{P}(\pi_t + \delta u_t, i)$
  Set $v(\pi_t + \delta u_t; \mathbf{P}, S, \rho) = \max_i \rho(z_{1,i}, S)$ and $v(\pi_t - \delta u_t; \mathbf{P}, S, \rho) = \max_i \rho(z_{2,i}, S)$
  Set sub-gradient estimate $\hat{g}_t = \frac{d}{2\delta} \left(v(\pi_t + \delta u_t; \mathbf{P}, S, \rho) - v(\pi_t - \delta u_t; \mathbf{P}, S, \rho) \right) u_t$
  Update $\theta_{t+1} = \theta_t - \eta \hat{g}_t$
Output: $\bar{\pi}_T = \frac{1}{T} \sum_{t=1}^{T} \pi_t$

Proof. Being somewhat more explicit with our notation, let $N_{(i_1,i_2,j)}$ denote the number of samples observed under the passive sampling model at index (i_1, i_2, j) of the tensor. Proceeding as in equation (17), we have

$$\Pr \left\{ \|\mathbf{P}^j(\cdot, i_2) - \widehat{\mathbf{P}}^j(\cdot, i_2)\|_\infty \geq c \sqrt{\frac{\log(cd/\delta)}{\min_{i_1 \in [d]} N_{(i_1,i_2,j)}}} \right\} \leq \delta.$$

Summing over all criteria $j \in [k]$ along with a union bound, we obtain

$$\Pr \left\{ \|\mathbf{P}(\cdot, i_2) - \widehat{\mathbf{P}}(\cdot, i_2))\|_{\infty,1} \geq ck \sqrt{\frac{\log(cdk/\delta)}{\min_{i_1,j} N_{(i_1,i_2,j)}}} \right\} \leq \delta.$$

Finally, in order to obtain a bound on the maximum deviation in the $(\infty, 1)$-norm, we take a union bound over all d choices of the index i_2, and apply inequality (16) to obtain

$$\max_{i_2} \|\mathbf{P}(\cdot, i_2) - \widehat{\mathbf{P}}(\cdot, i_2))\|_{\infty,1} \leq ck \sqrt{\frac{d^2 k}{n} \log \left(c \frac{dk}{\delta} \right)}$$

with probability exceeding $1 - \delta$. □

A few comments regarding the corollary are in order. The above corollary suggests that the sample complexity required for obtaining an ϵ-accurate solution with respect to the ℓ_1 norm is $n = \widetilde{O}(\frac{d^2 k^3}{\epsilon^2})$. Observe that this bound is a factor of k^2 worse than the corresponding one for ℓ_∞ norm established in Corollary 1. This additional sample complexity occurs since for any vector $v \in \mathbb{R}^k$, we have $\|v\|_1 \leq k\|v\|_\infty$. This implies that the error when measured with respect to ℓ_1 can be upto k times larger; since the sample complexity scales as $\frac{1}{\epsilon^2}$, the corresponding increase with respect to the number of criteria k is quadratic.

C.2 Optimization algorithms

Recall that Theorem 3 established that the objective function $v(\pi; \mathbf{P}, S, \|\cdot\|_q)$ is convex in π and Lipschitz with respect to the ℓ_1 norm. This implies that one could compute the plug-in solution $\widehat{\pi}_{\text{plug}}$ as a solution to a constrained optimization problem. In this section, we discuss a few specific algorithms based on zeroth-order and first-order methods for obtaining such a solution.

C.2.1 Zeroth-order optimization

Zeroth-order methods for minimizing a function $f(x)$ over $x \in \mathcal{X}$ work with a function query oracle. That is, at each time step, the algorithm has access to an oracle which returns the value $f(x)$ for any point $x \in \mathcal{X}$. In our setup, since we are interested in minimizing the value function $v(\pi; \mathbf{P}, S, \rho)$ over $\pi \in \Delta_d$, such a function query requires access to the target set S via an oracle $\mathcal{O}_S^0$ such that

$$\mathcal{O}_S^0(z) \to \min_{z_1 \in S} \rho(z, z_1) \ ,$$

for the underlying distance function $\rho(\cdot)$. The oracle $\mathcal{O}_S^0$ essentially takes as input a score vector $z \in [0,1]^k$ and outputs the distance of this point to the target set S. Given this oracle, it is easy to see that for any π, one can compute the corresponding value function $v(\pi; \mathbf{P}, S, \rho)$.

There have been several algorithms proposed for optimization with such oracles when the underlying function f is convex [23, 2, 48, 21, 38, 49] or non-convex, smooth [26]. The key idea in the proposed algorithms is to utilize the zeroth-order oracle to constuct estimates of the (sub-)gradient of the function f using a class of techniques called *randomized smoothing*. The algorithms then differ in the construction of these estimates depending on the underlying randomness as well as on the number of oracle calls during each time step.

Given the results of Theorem 3, we can restrict our focus on algorithms for the class of convex Lipschitz function f. To this end, Shamir [49] proposed an algorithm for optimizing such functions which required *two* function evaluations at each time. The algorithm, adapted to the multi-criteria preference learning problem, is detailed in Algorithm 1. For our setup, we select the negative entropy regularization, $r(\pi) = \sum_i \pi_i \log(\pi_i)$ to suit the geometry of our domain $\mathcal{X} = \Delta_d$.

The proposed algorithm, maintains an estimate of the distribution, π_t, and at each time step t, queries the function value $v(\cdot; \mathbf{P}, S, \rho)$ at the following two points: $\pi_t + \delta u_t$ and $\pi_t - \delta u_t$, where u is sampled uniformly from the Euclidean unit sphere and $\delta > 0$ represents the smoothing radius. Given these queries, the sub-gradient estimate, $\hat{g}_t$ is then obtained as:

$$\hat{g}_t := \frac{d}{2\delta} \left(v(\pi_t + \delta u_t; \mathbf{P}, S, \rho) - v(\pi_t - \delta u_t; \mathbf{P}, S, \rho) \right) u_t \ .$$

The sub-gradient estimate is then used to update the parameter estimate π_{t+1} using the mirror descent algorithm with the specified regularization function. The zeroth-order method in Algorithm 1 does not require the underlying function to be smooth and hence works for our problem setup with arbitrary non-differentiable distance functions. We can now obtain the following convergence result, based on Theorem 1 from the work of Shamir [49].

Proposition 5. *Suppose the conditions of Theorem 3 hold, and that Algorithm 1 is run for T iterations with step-size $\eta_t = \frac{c}{k^{1/q}\sqrt{dT}}$ and smoothing radius $\delta = \frac{c \log d}{\sqrt{T}}$, and produces a sequence $\pi_1, \pi_2, \ldots, \pi_T$. Then we have*

$$v\left(\bar{\pi}_T; \mathbf{P}, S, \|\cdot\|_q\right) \leq \min_{\pi \in \Delta_d} v(\pi; \mathbf{P}, S, \|\cdot\|_q) + c k^{\frac{1}{q}} \cdot \sqrt{\frac{d \log^2 d}{T}}$$

where $\bar{\pi}_T = \frac{1}{T} \sum_{t=1}^{T} \pi_t$.

Proof. By Theorem 3, the value function $v(\pi; \mathbf{P}, S, \|\cdot\|_q)$ is convex and $L_{\mathsf{v}} = k^{\frac{1}{q}}$-Lipschitz with respect to $\|\cdot\|_1$. Also, the choice of the regularizer $r(\pi) = \sum_i \pi_i \log(\pi_i)$ is 1-strongly convex with respect to the $\|\cdot\|_1$. Plugging in the above values in Theorem 1 from [49] establishes the above convergence rate. □

Thus, in order to obtain a distribution $\hat{\pi}$ that is ϵ-close to π^* in function value, we need to run Algorithm 1 for $T = O\left(\frac{k^{\frac{2}{q}} d \log^2 d}{\epsilon^2}\right)$ iterations. Also, note that each iteration of the algorithm requires d calls to the oracle $\mathcal{O}_S^0$. Therefore the total oracle complexity of the procedure is $O\left(\frac{k^{\frac{2}{q}} d^2 \log^2 d}{\epsilon^2}\right)$.

C.3 First-order optimization

In this section, we look at first-order methods to compute the plug-in estimator. Let us denote by $\partial v(\pi)$ the set of sub-differentials of the function $v(\cdot; \mathbf{P}, S, \|\cdot\|)$ evaluated at π. Further, let the set $\Gamma(\pi)$ denote the set of maximizers for a policy π, that is,

$$\Gamma(\pi) = \left\{ \tilde{\pi} \in \Delta_d \mid \tilde{\pi} \in \operatorname*{argmax}_{\pi_2 \in \Delta_d} \min_{z \in S} \left[\|\mathbf{P}(\pi, \pi_2) - z\| \right] \right\} \ . \tag{26}$$

Note that both of these quantities depend implicitly on the tuple $(S, \mathbf{P}, \|\cdot\|)$, but we have dropped this dependence in the notation. Given the setup above, Lemma 5 below characterizes this set $\partial v(\pi)$ for any smooth ℓ_q norm (with $1 < q < \infty$).

Algorithm 2: First-order method for multi-criteria preference learning

Input: Time steps T, step size η
Initialize: $\theta_1 = 1_k$
for $t = 1, \ldots, T$ **do**
  Set the distribution $\pi_t = \frac{\theta_t}{\|\theta_t\|_1}$
  Obtain $g_t \in \text{conv}\left\{ \frac{\mathbf{P}(\cdot,\pi_2)[\mathbf{P}(\pi_t,\pi_2) - \Pi_S(\mathbf{P}(\pi_t,\pi_2))]}{\|\mathbf{P}(\pi_t,\pi_2) - \Pi_S(\mathbf{P}(\pi_t,\pi_2))\|_q} \mid \pi_2 \in \Gamma(\pi_t) \right\}$ [See eq.(26) for $\Gamma(\pi_t)$]
  Update $\theta_{t+1,i} = \pi_{t,i} \exp(-\eta g_{t,i})$
Output: $\bar{\pi}_T = \frac{1}{T}\sum_{t=1}^{T} \pi_t$

Lemma 5. *Suppose that the distance is induced by a smooth ℓ_q norm for $1 < q < \infty$. Then the set of sub-differentials of v at π is given by:*

$$\partial v(\pi) = \textit{conv}\left\{ \frac{\mathbf{P}(\cdot,\pi_2)\left[\mathbf{P}(\pi,\pi_2) - \Pi_S(\mathbf{P}(\pi,\pi_2))\right]}{\|\mathbf{P}(\pi,\pi_2) - \Pi_S(\mathbf{P}(\pi,\pi_2))\|_q} \mid \pi_2 \in \Gamma(\pi) \right\},$$

where $\Pi_S(z)$ denotes the unique projection of the point z onto set S along $\|\cdot\|_q$.

We defer the proof of the above lemma to later in the section. Note that in order to access such a sub-gradient, we need access to an oracle $\mathcal{O}_S^1$ that provides projection queries of the form

$$\mathcal{O}_S^1(z) \to \operatorname*{argmin}_{z_1 \in S} \rho(z, z_1).$$

The oracle $\mathcal{O}_S^1$ takes in a point z and outputs the closest point in the set S to this point. Given such an oracle, we can compute the sub-gradient of the function $v(\pi; \mathbf{P}, S, \rho)$ using Lemma 5 by evaluating it at the point given by $\mathbf{P}(\pi, \pi_2)$ for some $\pi_2 \in \Gamma(\pi)$.

Given access to such a projection oracle $\mathcal{O}_S^1$, we detail out a procedure based on a standard implementation of mirror descent with entropic regularization (or Exponentiated gradient method) in Algorithm 2 to minimize the objective $v(\pi; G)$. Note that we select the negative entropy function, $r(\pi) = \sum_i \pi_i \log(\pi_i)$, as the regularization function for the mirror descent procedure since our parameter space is given by the simplex Δ_k and the negative entropy function is known to be 1-strongly convex with respect to $\|\cdot\|_1$ over this space.

The algorithm works by maintaining at each time instance a distribution π_t over the set of objects and updates it via an exponentiated gradient update. That is, the sub-gradient g_t is evaluated at the current point π_t using access to both $\mathcal{O}_S^1$ and $\mathcal{O}_S^0$, and is used to update each coordinate of the variable θ_t. The updated distribution π_{t+1} is obtained via a KL-projection of θ_t onto the simplex Δ_k, which can be shown to be equivalent to the normalization $\theta/\|\theta\|_1$. We now proceed to prove a convergence result for this gradient-based Algorithm 2, based on a standard analysis of the mirror descent procedure (for example, see [14, Theorem 4.2]).

Proposition 6. *Suppose the conditions of Theorem 3 hold and consider any ℓ_q-norm for $1 < q < \infty$. Suppose that running Algorithm 1 for T iterations with step-size $\eta_t = \frac{1}{k^{1/q}}\sqrt{\frac{2\log d}{T}}$ produces a sequence $\pi_1, \pi_2, \ldots, \pi_T$. Then we have*

$$v(\bar{\pi}_T; \mathbf{P}, S, \|\cdot\|_q) \leq \min_{\pi \in \Delta_d} v(\pi; \mathbf{P}, S, \|\cdot\|_q) + k^{\frac{1}{q}} \cdot \sqrt{\frac{2\log d}{T}}$$

where $\bar{\pi}_T = \frac{1}{T}\sum_{t=1}^{T} \pi_t$.

Proof. Note that the function $v(\pi; \mathbf{P}, S, \|\cdot\|_q)$ is convex and $k^{\frac{1}{q}}$-Lipschitz with respect to the ℓ_1 norm from Theorem 3. Further, the mirror map given by negative entropy function is 1-strongly convex with respect to $\|\cdot\|_1$. Plugging in these values in Theorem 4.2 from [14] establishes the required convergence rate. □

In order to obtain an ϵ-accurate solution in function value, it suffices to run the above algorithm for $T = O\left(\frac{k^{\frac{2}{q}}\log d}{\epsilon^2}\right)$ iterations, with each iteration using 1 call to the oracle $\mathcal{O}_S^1$ and d calls

to the oracle $\mathcal{O}_S^0$ (to obtain the set Γ). Thus, we see that the total oracle complexity changes as $\mathcal{O}_S^1 : O\left(\frac{k^{\frac{2}{q}} \log d}{\epsilon^2}\right)$ calls and $\mathcal{O}_S^0 : O\left(\frac{k^{\frac{2}{q}} d \log d}{\epsilon^2}\right)$ calls – effectively, an $O(d \log d)$ decrease in the calls to $\mathcal{O}_S^0$ is compensated by a corresponding increase of $O(\frac{\log d}{\epsilon^2})$ calls to the stronger oracle $\mathcal{O}_S^1$.

Proof of Lemma 5. Consider the function $\phi(\pi_1, \pi_2) = \max_{z \in S} \|\mathbf{P}(\pi_1, \pi_2) - z\|$ over the domain $\pi_2 \in \Delta_d$. For any fixed π_2, we have that the function $\phi(\pi_1, \pi_2)$ is convex in π_1. Thus, by Danskin's theorem, we have that the subdifferential set is given by:

$$\partial v(\pi) = \text{conv}\left\{ \frac{\partial \phi(\pi, \pi_2)}{\partial \pi} \mid \pi_2 \in \Gamma(\pi) \right\} , \tag{27}$$

where conv represents the convex hull of the set. Let us now focus on the partial derivative $\frac{\partial \phi(\pi, \pi_2)}{\partial \pi}$ for any π_2 which is a maximizer. This partial derivative involves differentiation of a metric projection onto a convex set, which has been studied extensively in the literature of convex analysis [41, 58, 5]. Recently, Balestro et al. [8] established that for distance functions given by smooth norms, the derivative of metric projection for any $z \notin S$ is given by:

$$\nabla \rho(z, S) = \nabla \min_{z_2 \in S} \|z - z_2\| = \frac{z - \Pi_S(z)}{\|z - \Pi_S(z)\|} ,$$

where $\Pi_S(z)$ denotes the unique projection of the point z onto set S. Combining this with the chain rule of differentiation, we have that:

$$\frac{\partial \phi(\pi, \pi_2)}{\partial \pi} = \frac{\mathbf{P}(\cdot, \pi_2) \left[\mathbf{P}(\pi, \pi_2) - \Pi_S(\mathbf{P}(\pi, \pi_2))\right]}{\|\mathbf{P}(\pi, \pi_2) - \Pi_S(\mathbf{P}(\pi, \pi_2))\|_q} .$$

The above, in conjunction with equation (27) establishes the desired claim. □

D Details of user study

In this section, we provide the deferred details of the user study from Section 4.

Self-driving environment. The self-driving environment consists of an autonomous car which can be controlled by providing real-valued inputs acceleration and angular acceleration at every time step. We allow the policies to have access to the dynamics of this environment. Observe that there is no explicit reward function in the environment and each policy differs in the way it optimizes a chosen reward function to drive the car forward in a safe manner.

Policies. The MPC based Policies A-E were constructed by optimizing linear rewards comprising features F1-F9 as

- F1 Distance from the starting point along y-axis.
- F2 Velocity of the autonomous car.
- F3 Distance from the center of each lane.
- F4 Gaussian collision detector for nearby objects.
- F5 Collision detector which works at smaller radii than F4.
- F6 Over-speeding feature which penalizes higher speeds.
- F7 Reward for over-taking vehicles in the front.
- F8 Gaussian off-road detector.
- F9 Reward to promote speeding up near obstacles.

For each of the base policy, we set the weights of the features to encode different driving behaviors.

- Pol A programmed to prefer the right-most lane and progress forward at a slow speed.
- Pol B programmed to prefer the left-most lane and move forward as fast as possible.
- Pol C programmed to be conservative, avoids collision and proceeds forward.
- Pol D programmed to get attracted towards other cars and obstacles.
- Pol E programmed to prefer center lane and exhibit opportunistic behavior by moving ahead of other cars.

Details of target set and linear weights. We selected the two data-oblivious sets to trade-off between the criteria C1-C5 as

$$\begin{aligned} S_1 &= \{z \mid z \in [0,1]^5, z_1 \geq 0.3, z_2 \geq 0.3, z_3 \geq 0.2, z_4 \geq 0.3, z_5 \geq 0.4\}, \\ S_2 &= \{z \mid z \in [0,1]^5, z_1 \geq 0.25, z_2 \geq 0.25, z_3 \geq 0.25, z_4 \geq 0.25, z_5 \geq 0.25, z_1 + z_5 \geq 0.9\}. \end{aligned} \tag{28}$$

In addition, we selected 9 set of weights $w_{1:9}$ for linearly combining the different criteria.

w_1: Average of the users' self-reported weights.

w_2: Weight vector obtained by regressing the overall criterion on C1-C5 with squared loss as

$$w_2 \in \operatorname*{argmin}_{w \in \Delta_5} \sum_{i_1, i_2} (\mathbf{P}_{\mathsf{ov}}(i_1, i_2) - \sum_j w(j) \mathbf{P}^j(i_1, i_2))^2.$$

w_3: Weight obtained by regressing Bradley-Terry-Luce (BTL) scores. The BTL parametric model assumes a real-valued score v_i for each policy and posits that $\Pr(\text{Pol } i \succeq \text{Pol } j) = \exp(v_i)/\exp(v_i) + \exp(v_j)$. Denoting the scores obtained from the overall preferences by v^{ov} and those obtained from the individual criteria by v^j for $j \in [5]$, the weight

$$w_2 \in \operatorname*{argmin}_{w \in \Delta_5} \sum_i (v_i^{\mathsf{ov}} - \sum_j w(j) v_i^j)^2.$$

w_4: Data-oblivious weight $w_4 = [0.2, 0.2, 0.2, 0.2, 0.2]$.

w_5: Data-oblivious weight $w_5 = [0.25, 0.5/3, 0.5/3, 0.5/3, 0.25]$.

w_6: Data-oblivious weight $w_6 = [0.30, 0.4/3, 0.4/3, 0.4/3, 0.30]$.

w_7: Data-oblivious weight $w_7 = [0.5/3, 0.5/3, 0.25, 0.5/3, 0.25]$.

w_8: Data-oblivious weight $w_8 = [0.4/3, 0.4/3, 0.3, 0.4/3, 0.30]$.

w_9: Data-oblivious weight $w_9 = [0.3, 0.1/2, 0.3, 0.1/2, 0.3]$.

The set of data oblivious weights were chosen to account for different trade-offs along the criteria C1-C5 including the uniform weight w_4.

Data Collection. Table 1 shows the comparison data collected from the Mturk users in both the phases of the experiment. The entry i, j of the comparison matrices represents the fraction of users which preferred Policy i over Policy j. The top 5 rows and columns of each matrix correspond to the baseline policies while the bottom rows correspond to the two randomized policies R1 and R2 obtained as the Blackwell winner corresponding to sets S_1 and S_2 respectively.

In addition, we would like to highlight some details from an experiment design perspective. Since the experiment was run in two phases, we could not guarantee the same set of subjects to participate in both parts of the experiment. In order to limit distribution shifts, we restricted the nationality of the subjects to United States and began both the phases on the same time and day of the week. Also, in order to prevent biased evaluations, the ordering of the policy pairs as well as the ordering policies within a comparison was randomized across the users.

Figures 3, 4 and 5 shows the experiment setup we used for obtaining comparison data from Amazon Mechanical Turk users consisting of the instructions, the policy comparison page and the questionnaire that the users were asked to fill out.

Implementation Details. The computation of the Blackwell winner for the different target set S was done using the CVX package in Matlab. For the MPC policies, the horizon length H was set to be 18, three times the planning horizon $= 6$.

	A	B	C	D	E
A	0.50	0.64	0.45	0.41	0.39
B	0.36	0.50	0.30	0.30	0.25
C	0.55	0.70	0.50	0.55	0.57
D	0.59	0.70	0.45	0.50	0.52
E	0.61	0.75	0.43	0.48	0.50
R1	0.49	0.80	0.22	0.46	0.29
R2	0.49	0.88	0.66	0.61	0.41

(a) C1: Aggressiveness

	A	B	C	D	E
A	0.50	0.57	0.50	0.50	0.41
B	0.43	0.50	0.30	0.39	0.45
C	0.50	0.70	0.50	0.43	0.59
D	0.50	0.61	0.57	0.50	0.57
E	0.59	0.55	0.41	0.43	0.50
R1	0.46	0.71	0.32	0.51	0.39
R2	0.51	0.71	0.61	0.59	0.51

(b) C2: Predictability

	A	B	C	D	E
A	0.50	0.16	0.25	0.32	0.30
B	0.84	0.50	0.89	0.82	0.68
C	0.75	0.11	0.50	0.73	0.61
D	0.68	0.18	0.27	0.50	0.41
E	0.70	0.32	0.39	0.59	0.50
R1	0.73	0.22	0.76	0.78	0.76
R2	0.90	0.24	0.44	0.66	0.66

(c) C3: Quickness

	A	B	C	D	E
A	0.50	0.59	0.45	0.57	0.39
B	0.41	0.50	0.32	0.34	0.32
C	0.55	0.68	0.50	0.48	0.59
D	0.43	0.66	0.52	0.50	0.50
E	0.61	0.68	0.41	0.50	0.50
R1	0.44	0.80	0.20	0.39	0.24
R2	0.41	0.80	0.71	0.59	0.39

(d) C4: Conservativeness

	A	B	C	D	E
A	0.50	0.52	0.41	0.50	0.43
B	0.48	0.50	0.32	0.55	0.55
C	0.59	0.68	0.50	0.55	0.57
D	0.50	0.45	0.45	0.50	0.50
E	0.57	0.45	0.43	0.50	0.50
R1	0.54	0.68	0.32	0.49	0.41
R2	0.63	0.73	0.59	0.61	0.54

(e) C5: Collision Risk

	A	B	C	D	E
A	0.50	0.39	0.25	0.43	0.34
B	0.61	0.50	0.30	0.50	0.50
C	0.75	0.70	0.50	0.57	0.61
D	0.57	0.50	0.43	0.50	0.48
E	0.66	0.50	0.39	0.52	0.50
R1	0.66	0.76	0.29	0.59	0.39
R2	0.66	0.73	**0.66**	0.56	0.51

(f) Overall Preferences

Table 1. Each matrix consists of pairwise comparisons between policies elicited from a user study with around 50 participants on Mturk. An entry i, j of the comparison matrices represents the fraction of users which preferred Policy i over Policy j. Policies A-E comprise the base set of policies while Policies R1-R2 are the randomized Blackwell winners obtained from the sets in equation (28). While Policy C is the overall von Neumann winner, Policy R2 is preferred over it by 66% of the users.

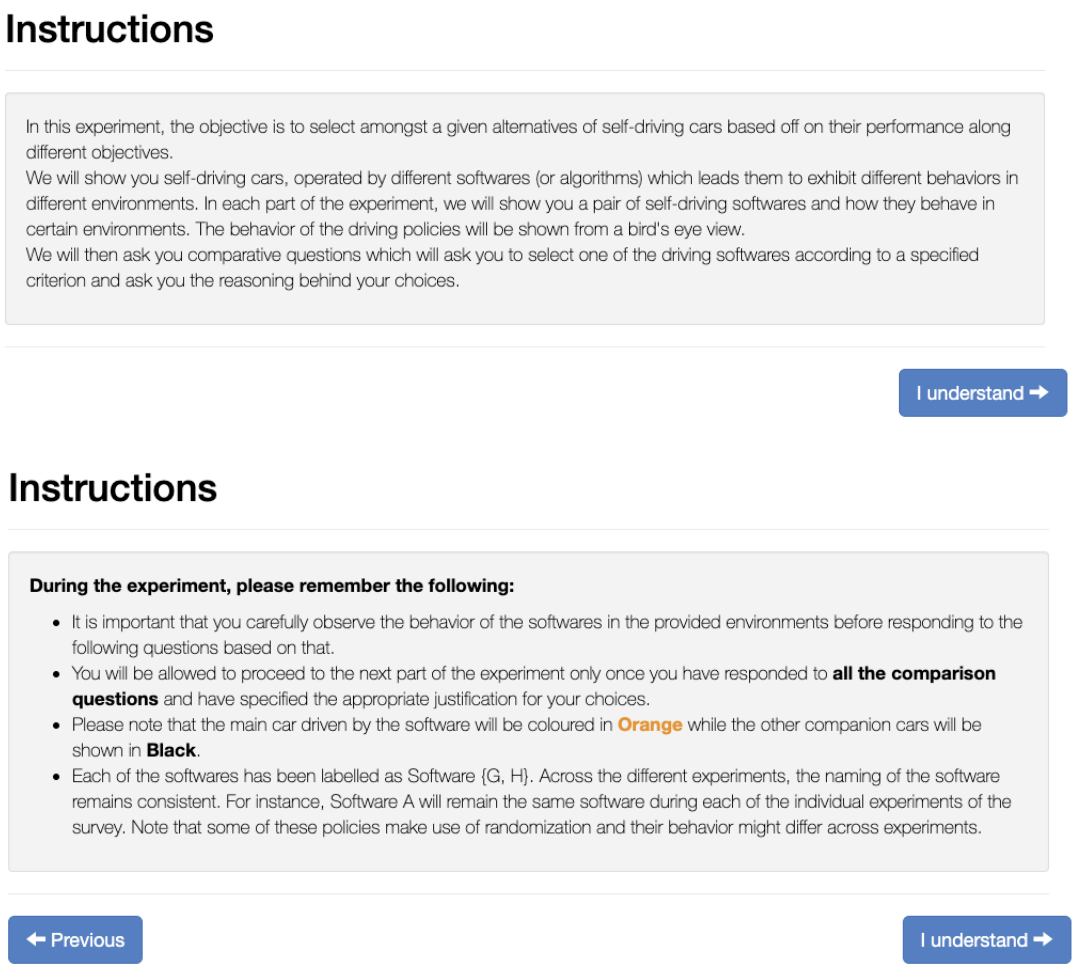


Figure 3. Instructions provided to the users before the experiment began. The users were asked to compare behavior of policies and were told to expect some policies to exhibit a randomized behavior.

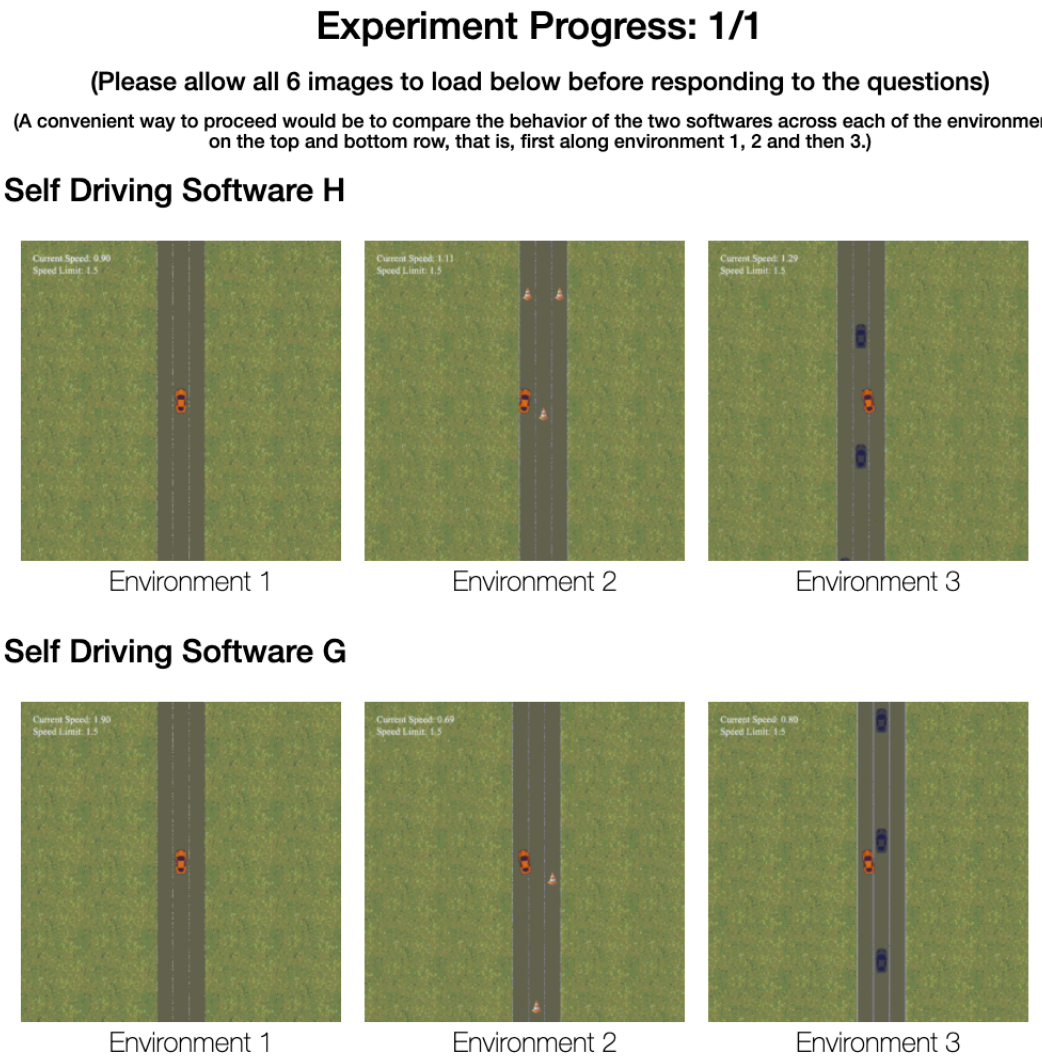


Figure 4. Layout of the experiment where each panel shows a GIF exhibiting a Policy controlling the autonomous vehicle in one of the worlds of the environment. The users were instructed to compare behaviors across each of the columns before proceeding to answer the questions.

Questions

Q1*. Which of the two softwares exhibits a less aggressive behavior? :

○ **Software H** ○ **Software G**

Q2*. Which of the two softwares is more predictable in their behavior? That is, for which of the two softwares do you think you will be able to anticipate its performance in a new environment. :

○ **Software H** ○ **Software G**

Q3*. Which of the two softwares will get you to your destination the quickest?:

○ **Software H** ○ **Software G**

Q4*. Which of the two softwares is more conservative in its driving approach?:

○ **Software H** ○ **Software G**

Q5*. Which of the two softwares if has a lower risk of collision with another car or an obstacle?:

○ **Software H** ○ **Software G**

Q6*. [Overall Preference] Imagine you were to select one of the two softwares to get you to your destination. Which of the two softwares would you prefer?:

○ **Software H** ○ **Software G**

* Please provide a brief sentence about how you made your selections: (Press 'Enter' after typing the sentence)

* For each of the following characteristic, please indicate their relevance in determining the overall preference betweem the softwares. Please take into account all the expriments that you completed in this study. *(5 = extremely important, 1 = had little importance)*

Aggresiveness of the software:
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Predictability of the software:
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Speed or quickness of the software:
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Conservativeness of the software:
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Collision Risk of the software:
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 5. Layout of the questions panel comprising the 6 comparison questions and the form for reporting the relevance of each criterion in the overall evaluation.