Eliciting Truthful Feedback for Preference-Based Learning via the VCG Mechanism

Leo Landolt[1,2], Anna Maddux*[,1], Andreas Schlaginhaufen*[,1], Saurabh Vaishampayan[1], Maryam Kamgarpour[1]

[1]EPFL, Switzerland [2]ETH Zürich, Switzerland

Abstract

We study resource allocation problems in which a central planner allocates resources among strategic agents with private cost functions in order to minimize a social cost, defined as an aggregate of the agents' costs. This setting poses two main challenges: (i) the agents' cost functions may be unknown to them or difficult to specify explicitly, and (ii) agents may misreport their costs strategically. To address these challenges, we propose an algorithm that combines preference-based learning with Vickrey–Clarke–Groves (VCG) payments to incentivize truthful reporting. Our algorithm selects informative preference queries via D-optimal design, estimates cost parameters through maximum likelihood, and computes VCG allocations and payments based on these estimates. In a one-shot setting, we prove that the mechanism is approximately truthful, individually rational, and efficient up to an error of $\tilde{\mathcal{O}}(K^{-1/2})$ for K preference queries per agent. In an online setting, these guarantees hold asymptotically with sublinear regret at a rate of $\tilde{\mathcal{O}}(T^{2/3})$ after T rounds. Finally, we validate our approach through a numerical case study on demand response in local electricity markets.

1 INTRODUCTION

Allocating resources among self-interested agents is a fundamental problem in many real-world systems, including electricity markets, communication networks, and transportation systems (Chremos and Malikopoulos, 2024). In these problems, a central planner aims to determine a resource allocation that minimizes the social cost, defined as an aggregate function of individual agents' costs. Agents' cost functions are typically private, thus, the central planner relies on the agents to communicate their cost functions. However, this is often problematic, either because agents cannot explicitly specify their cost functions or because doing so is difficult.

A prominent example arises in local electricity markets, where a grid operator procures energy flexibility from domestic consumers through deferrable appliance usage, such as shifting or adjusting heating or laundry schedules (Li et al., 2020; Tsaousoglou et al., 2022). Reporting the exact economic costs for these adjustments can be difficult and impractical for consumers. Instead, they can typically more easily express their preferences over different flexibility options.

This motivates us to consider a setting where agents provide preference feedback to the central planner rather than reporting their cost functions directly. The central planner can then leverage this feedback to infer agents' underlying cost structures and compute an allocation that minimizes the resulting social cost.

A central challenge, however, is that agents may behave strategically, misreporting their preferences to manipulate the resource allocation outcome for their benefit. This can lead to socially suboptimal outcomes. In electricity markets, for instance, consumers may misreport their preferences to secure more favorable energy tariffs, thereby undermining efficient allocation (Yazdani-Damavandi et al., 2017).

In this paper, we investigate how the central planner can elicit truthful preference feedback from agents to solve the resource allocation problem, minimizing the social cost.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. *The authors contributed equally to this work.

1.1 Related work

Learning from preference feedback was first studied in the bandit literature, where the usual numerical rewards were replaced by pairwise preferences (Yue and Joachims, 2009; Ailon et al., 2014). More recently, preference-based learning has received considerable attention across a variety of applications, most prominently in fine-tuning large language models (Ziegler et al., 2019; Rafailov et al., 2023), reinforcement learning (Christiano et al., 2017), robotics and human-robot interaction, and personalized healthcare decision support. Preference feedback is especially valuable in settings involving human interaction, as human users are often better at expressing relative judgment in the form of preferences between outcomes than providing numerical evaluations quantifying the value of an outcome (Pereira et al., 2019; Lee et al., 2023).

A common modeling assumption in this literature is that agents' preferences can be represented by cardinal models, such as the Bradley–Terry model (Bradley and Terry, 1952; Luce et al., 1959). Prior works have largely focused on the case where agents truthfully report preferences according to such models (Azar et al., 2024; Chowdhury et al., 2024). An important open challenge is to understand how preference learning methods perform in settings where agents may strategically misreport their preferences in pursuit of personal gain.

Strategic behavior in multi-agent resource allocation problems has been extensively studied in mechanism design. In settings with transferable utility – where agents' utilities can be expressed in monetary terms – mechanisms commonly incorporate payments to incentivize socially desirable and efficient outcomes (Fallah et al., 2024; Góis et al., 2025). In such cases, an agent's utility is defined as the payment received minus the incurred costs. A classical result establishes that under the Vickrey–Clarke–Groves (VCG) mechanism, agents are incentivized to truthfully report their cost functions (Vickrey, 1961; Clarke, 1971; Groves, 1973).

However, this approach requires that agents disclose their entire cost functions explicitly. A more realistic assumption is that agents instead provide numerical feedback, i.e., reporting their realized cost for a given allocation (Babaioff et al., 2009; Devanur and Kakade, 2009). Closely related to our setting, Kandasamy et al. (2023) study numerical feedback over a finite allocation space, and show that the VCG mechanism incentivizes truthful reporting asymptotically. This is restrictive – for example in electricity markets, energy flexibility is naturally continuous, and it is unrealistic to expect consumers to report precise economic cost values. Recent work has also explored auction-based mechanisms for eliciting preference feedback in the context of fine-tuning large language models, but focuses on second-price auctions rather than general resource allocation (Zhang and Duan, 2024). To address these limitations, we study resource allocation problems with a compact allocation space in which agents provide preference feedback rather than exact numerical reports.

1.2 Contributions

To the best of our knowledge, this is the first work to study preference learning in resource allocation problems with strategic agents under transferable utilities, where agents' preferences are modeled using the Bradley–Terry framework. Our contributions are threefold:

- We propose a resource allocation algorithm that learns agents' private costs, modeled as a linear function of features, from strategic pairwise preferences. The algorithm uses maximum likelihood estimation for parameter estimation, D-optimal design for query selection, and VCG payments to ensure truthful feedback.
- We establish guarantees analogous to the classical VCG mechanism for truthfulness, individual rationality, and efficiency. Our algorithm satisfies these properties approximately in the one-shot setting (Theorem 1) and asymptotically in the multi-round setting (Theorem 2).
- We demonstrate the applicability of our setup to local electricity markets and validate our guarantees through a numerical case study in this domain.

1.3 Notation

We write $\|\cdot\|$ for the Euclidean norm and $\langle\cdot,\cdot\rangle$ for the standard inner product in $\mathbb{R}^d$. We use the standard notation $\mathcal{O}(T)$ for asymptotic upper bounds, as well as $\tilde{\mathcal{O}}(T) = O(T\,\mathrm{polylog}(T))$ for suppressing polylogarithmic terms. Finally, we denote $[N] := \{1, \dots, N\}$.

2 BACKGROUND AND PROBLEM FORMULATION

We consider a resource allocation problem with N strategic agents, each with a compact allocation set $\mathcal{A}_i$. A central planner chooses an allocation $a = (a_1, \dots, a_N) \in \mathcal{A}$, where $\mathcal{A} = \mathcal{A}_1 \times \dots \times \mathcal{A}_N$. In our electricity market application, N is the number of consumers and a_i is the energy supplied (kWh). The allocation results in a continuous cost $c_i : \mathcal{A} \to \mathbb{R}_+$ for

each agent. The goal of the central planner is to find a feasible allocation that minimizes the social cost

$$J(a) := \sum_{i=1}^{N} c_i(a), \quad \text{s.t. } a \in \mathcal{F}, \tag{1}$$

where $\mathcal{F} \subseteq \mathcal{A}$ is a compact feasible set. In the electricity market application, for example, $\mathcal{F} = \{a \in \mathcal{A} \mid \sum_{i=1}^{N} a_i = P\}$, where P is the total energy flexibility required by the grid (kWh). This leads to a classical problem in mechanism design: a central planner must elicit private cost functions from strategic agents in order to find a socially efficient allocation.

2.1 Auctions

A common approach to address the above problem are auctions, where each agent submits a bid function $b_i : \mathcal{A} \to \mathbb{R}_+$ intended to reflect their cost. A mechanism consists of an allocation rule $a^* : b \to \mathcal{A}$ and a payment rule $p : b \to \mathbb{R}^N$ based on the agents' bid functions $b = (b_1, \dots, b_N)$. Then, the central planner finds an allocation $a^*(b)$ such that:

$$a^*(b) \in \arg\min_{a \in \mathcal{F}} \hat{J}(a; b),$$

where $\hat{J}(a; b) = \sum_{i=1}^{N} b_i(a)$. The central planner assigns this allocation $a^*(b) = (a_1^*(b), \dots, a_N^*(b))$ to the agents who receive payments $p(a^*(b)) = (p_1(a^*(b)), \dots, p_N(a^*(b)))$. As a result, agent i experiences utility:

$$u_i(a^*(b)) = p_i(a^*(b)) - c_i(a^*(b)). \tag{2}$$

An agent is said to bid truthfully if $\bar{b}_i(a) = c_i(a)$ for all $a \in \mathcal{A}$. However, agents may misreport their costs and submit an arbitrary bid b_i to increase their utility. For example, in a pay-as-bid mechanism, also known as first-price auction, since payments to agents are equal to their bids, a rational agent would overbid to ensure profit (Karlin and Peres, 2017, Chapter 14.2). Consequently, $a^*(b)$ may not minimize the true social cost (1). Thus, to find a socially efficient allocation, the central planner must carefully design the payment rule in order to elicit truthful bids.

In particular, any desirable mechanism should satisfy the following three properties:

1) **Truthfulness:** Truthful bidding with $\bar{b}_i(a) = c_i(a)$ for all $a \in \mathcal{A}$, is a weakly dominant strategy Nash equilibrium, that is:

$$u_i(a^*(\bar{b}_i, b_{-i})) \geq u_i(a^*(b_i, b_{-i})),$$

for all $b_i : \mathcal{A} \to \mathbb{R}_+$ and $i \in [N]$. We use b_{-i} for the (potentially untruthful) bids of other agents.

2) **Individual rationality:** A truthful agent has non-negative utility, that is:

$$u_i(a^*(\bar{b}_i, b_{-i})) \geq 0,$$

for all bids b_{-i} of other agents.

3) **Efficiency:** When all agents bid truthfully, the resulting allocation minimizes the social cost, that is:

$$a^*(\bar{b}) \in \operatorname*{argmin}_{a \in \mathcal{F}} J(a).$$

A well-known mechanism that satisfies these properties is the VCG mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973), which we discuss in Section 3.1.

2.2 Preference model

In our setting, agents cannot directly report their cost functions, but instead express their preferences between two allocations. We model these preferences using the stochastic Bradley–Terry model, which reflects bounded rationality in human decision-making (Bradley and Terry, 1952; Luce et al., 1959).

Definition 1 (Bradley–Terry model)**.** *The probability that agent i prefers allocation $a \in \mathcal{A}$ over allocation $a' \in \mathcal{A}$, denoted as $a \succ a'$, is:*

$$\Pr(a \succ a') = \sigma\left(u_i(a) - u_i(a')\right), \tag{3}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function.[1]

Applying the Bradley–Terry model to an agent's utility given in Equation (2), we obtain:

$$\Pr(a \succ a') = \sigma\left(p_i(a) - c_i(a) - p_i(a') + c_i(a')\right).$$

The payments $p_i(a)$ and $p_i(a')$ are known, therefore, they only induce an affine shift in the argument of the sigmoid function. We say that preference feedback is truthful if agent i expresses their preference for a over a' according to a binary label $\bar{y}_i = \mathbb{1}(a \succ a')$ where

$$\bar{y}_i \sim \text{Bernoulli}\big(\sigma(p_i(a) - c_i(a) - p_i(a') + c_i(a'))\big),$$

is a Bernoulli-distributed random variable sampled according to the Bradley–Terry model. While $\bar{y}_i$ represents the agent's truthful preference, they may act strategically to increase their utility. That is, for given allocations a and a', agent i may report a preference

[1] The Bradley–Terry model can be extended with a rationality parameter $\beta_i > 0$. Then, the preference model for agent i is $\Pr(a \succ a') = \sigma\left(\beta_i[u_i(a) - u_i(a')]\right)$. Large β_i implies more deterministic decisions, whereas small β_i introduces more randomness. We assume β_i is known and, without loss of generality, set $\beta_i = 1$ for all i. We discuss this further in Appendix D.

$y_i \in \{0,1\}$ sampled from an arbitrary distribution. Note that we do not impose any structural assumptions on the agents' strategic feedback. In particular, agents can follow dynamic strategies when generating their untruthful preferences.

2.3 Problem formulation

The central planner's goal is to find an allocation that minimizes the social cost (1). Unlike standard auctions, where agents submit bid functions, the planner actively queries each agent with a pair of allocations $a, a' \in \mathcal{A}$. Agents then report their preference in form of a binary label $y_i = \mathbb{1}(a \succ a')$, and the planner learns their costs through repeated interaction. Importantly, to find a socially efficient allocation, the central planner must learn agents' true costs, which is possible only if agents report their preferences truthfully.

Problem: Can we design preference queries and payment rules to ensure *truthfulness, individual rationality, and efficiency*?

Motivated by the success of the VCG mechanism in auctions, we adopt VCG as our payment rule. We then study the above problem in two regimes. As a warm-up, we analyze a one-shot setting: the planner collects a fixed batch of pairwise preferences (e.g., via a questionnaire), estimates costs, and then computes allocations and payments from that estimate. In Section 4, we turn to a multi-round setting in which the planner alternates between exploration and exploitation, improving allocations and payments over time as the cost estimates become more accurate. The one-shot game captures one-time allocations (e.g., crowdsourcing a single task), whereas the multi-round game captures repeated allocations, such as our electricity market application in Section 5.

3 ONE-SHOT GAME: ALGORITHM AND ANALYSIS

In the one-shot game, the planner first collects a set of K pairwise preferences to learn the agents' costs, and then computes the allocation and VCG payments from this estimate. To efficiently explore the possibly infinite set of preference queries, we select queries using optimal design. We now detail each component.

3.1 Algorithm components

Cost estimation. To learn the agents' cost functions, we make the following modeling assumption.

Assumption 1. *Each agent i has a linear cost*

$$c_i(a) := \langle \theta_i^*, \phi_i(a) \rangle$$

with $\|\theta_i^*\| \le B$. The feature maps $\phi_i : \mathcal{A} \to \mathbb{R}^d$ are continuous, the differences $\{\phi_i(a) - \phi_i(a') \mid a, a' \in \mathcal{A}\}$ span $\mathbb{R}^d$, and $\max_{a \in \mathcal{A}} \|\phi_i(a)\| \le L$.

After K preference queries, the planner holds for each agent i the dataset $\mathcal{D}_i = \{(x_{i,k}, y_{i,k})\}_{k=1}^K$, where $x_{i,k} := \phi_i(a_{i,k}) - \phi_i(a'_{i,k})$ and $y_{i,k} = \mathbb{1}(a_{i,k} \succ a'_{i,k})$. Under the Bradley–Terry model, the planner can estimate the agents' costs by minimizing the negative log-likelihood

$$\hat{\theta}_i = \operatorname*{argmin}_{\|\theta_i\| \le B} \mathcal{L}_{\mathcal{D}_i}(\theta_i), \tag{4}$$

where

$$\mathcal{L}_{\mathcal{D}_i}(\theta_i) := -\sum_{(x_{i,k}, y_{i,k}) \in \mathcal{D}_i} \big[y_{ik} \log \sigma(\langle \theta_i, x_{i,k} \rangle) + (1 - y_{i,k}) \log(\sigma(-\langle \theta_i, x_{i,k} \rangle)) \big].$$

While the Bradley–Terry model depends on the agents' utilities, in our algorithm preference queries are collected during exploration rounds. In these rounds, payments are either zero (one-shot game) or constant across alternatives (multi-round game), so they do not affect the reported preferences. The resulting estimated cost function of agent i is then given by

$$\hat{c}_i(a; \mathcal{D}_i) := \langle \hat{\theta}_i, \phi_i(a) \rangle. \tag{5}$$

Importantly, each agent's cost function is learned separately and depends only on her own feedback.

Optimal design. The quality of the estimated cost functions $\hat{c}_i$ depends on how informative the individual preference queries $a_{i,k}$ and $a'_{i,k}$ are. Since the set of all queries $\mathcal{A} \times \mathcal{A}$ could be large or infinite – we only assumed $\mathcal{A}_i$ to be compact – exhaustive exploration (as by Kandasamy et al. (2023) for numerical feedback) is not possible. Instead, we leverage optimal experimental design (Pukelsheim, 2006) to select a set of queries that optimally explore the allocation space. Specifically, choosing comparison pairs via a D-optimal design subroutine (Lattimore and Szepesvári, 2020), denoted as D-Optimal-Design(K) in Algorithm 1, yields the following confidence guarantee for the estimated costs.

Lemma 1. *Under Assumption 1, supppose for each agent i, the planner selects $K > d(d+1)/2$ queries by* D-Optimal-Design(K). *If the agents provide truthful preference feedback $\bar{\mathcal{D}}_i = \{(x_{i,k}, \bar{y}_{i,k})\}_{k=1}^K$, then with probability at least $1 - \delta$,*

$$\big| (c_i(a) - c_i(a')) - \big(\hat{c}_i(a; \bar{\mathcal{D}}_i)) - \hat{c}_i(a'; \bar{\mathcal{D}}_i)\big) \big| \le \epsilon_K(\delta),$$

with $\epsilon_K(\delta) \in \tilde{\mathcal{O}}\left(d\sqrt{\log(1/\delta)/K}\right)$.

The above result follows by combining an MLE confidence guarantee (Schlaginhaufen et al., 2025) with the Kiefer–Wolfowitz theorem for optimal design (Kiefer and Wolfowitz, 1960). Remarkably, the sample complexity for estimating c_i up to a given error ϵ is independent of the size of the allocation space $\mathcal{A}$ and depends only on the dimension d and ϵ. As shown in Appendix A.2, a support of $d(d+1)/2$ distinct comparisons pairs is enough, which can then be queried repeatedly. For the precise definition of the subroutine D-Optimal-Design(K) and implementation details, see Appendix A.2. The full proof of Lemma 1 is provided in A.3.

VCG mechanism. Based on each agent's learned cost (5), we compute allocations and payments using the VCG mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973).

Definition 2 (VCG mechanism)**.** *The allocation and payment rules are defined as*

$$\begin{aligned} \hat{a}(\mathcal{D}) &\in \operatorname*{argmin}_{a\in\mathcal{F}} \hat{J}(a;\mathcal{D}), \\ p_i(\hat{a}(\mathcal{D})) &= \min_{a\in\mathcal{F}} \hat{J}_{-i}(a;\mathcal{D}_{-i}) - \hat{J}_{-i}(\hat{a}(\mathcal{D});\mathcal{D}_{-i}), \end{aligned} \tag{6}$$

where $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_N)$, $\hat{J}(a;\mathcal{D}) = \sum_{i=1}^{N} \hat{c}_i(a;\mathcal{D}_i)$ *and* $\hat{J}_{-i}(a;\mathcal{D}_{-i}) = \sum_{j\neq i} \hat{c}_j(a;\mathcal{D}_j)$.

While the VCG mechanism is well understood in the classical auction setting, where agents provide bid functions, our mechanism (6) operates on cost functions estimated from preference feedback. This raises the question of how the desirable properties (i.e., truthfulness, individual rationality, and efficiency) are affected by the learning errors, which we analyze next.

3.2 Algorithm

We now present the one-shot algorithm, summarized in Algorithm 1. The central planner first chooses K queries according to a D-optimal design subroutine, D-Optimal-Design(K), which ensures sufficient exploration. The planner then collects each agent's preferences between two candidate allocations. These preferences are used to estimate the agents' cost parameter. Finally, the central planner determines VCG allocations and payments based on the agents' learned costs.

3.3 Theoretical guarantees

In this section, we provide bounds for truthfulness, individual rationality, and efficiency in the one-shot game. These bounds reflect the worst-case outcome from the mechanism's perspective, for example, the maximum utility an agent could gain from misreporting preferences. To show individual rationality, we make the natural assumption that agents do not incur any cost when they are not part of an allocation.

Algorithm 1 One-shot game: VCG mechanism with preference feedback

1: **Input:** K
2: Select K comparisons for each agent i: $\{(a_{i,k}, a'_{i,k})\}_{k=1}^{K} \leftarrow$ D-Optimal-Design(K).
3: Collect preferences $\{y_{i,k}\}_{k=1}^{K}$ from each agent i.
4: Estimate $\hat{\theta}_i$ from $\mathcal{D}_i$ via Equation (4).
5: Compute $\hat{a}(\mathcal{D})$ and $p_i(\hat{a}(\mathcal{D}))$ for all i via (6).

Assumption 2. *There exists an allocation* $a^0 \in \mathcal{A}$ *with* $\phi_i(a^0) = 0$*, for all* $i \in [N]$.

Next, we show that Algorithm 1 is approximately truthful, individually rational, and efficient. We use $\bar{\mathcal{D}}$ to denote truthful feedback according to the Bradley–Terry model (3), and $\mathcal{D}$ for arbitrary feedback.

Theorem 1. *Let Assumption 1 hold, and suppose the mechanism makes* $K > d(d+1)/2$ *preference queries to each agent* $i \in [N]$*. Then, with probability at least* $1-\delta$*, Algorithm 1 satisfies:*

1) ***Truthfulness:*** *An agent's utility gain for arbitrary (possibly strategic) feedback is at most:*

$$u_i(\hat{a}(\mathcal{D}_i, \mathcal{D}_{-i})) - u_i(\hat{a}(\bar{\mathcal{D}}_i, \mathcal{D}_{-i})) \leq \epsilon_K(\delta).$$

2) ***Individual rationality:*** *Under Assumption 2, a truthful agent's utility is at least:*

$$u_i(\hat{a}(\bar{\mathcal{D}}_i, \mathcal{D}_{-i})) \geq -\epsilon_K(\delta).$$

3) ***Efficiency:*** *With all agents truthful, the gap to the optimal social cost* $J(a^*)$ *is at most:*

$$J(\hat{a}(\bar{\mathcal{D}})) - J(a^*) \leq N\epsilon_K(\delta/N).$$

The bound $\epsilon_K(\cdot) \in \tilde{\mathcal{O}}\left(d\sqrt{\log(1/\cdot)/K}\right)$ *holds after* K *queries per agent.*

This theorem shows that truthfulness, individual rationality, and efficiency are preserved for sufficient preference queries. The error terms $\epsilon_K(\delta)$ are in line with statistical rates for learning from preferences (Scheid et al., 2024). Note that the error decreases as we increase the number of queries.

Proof idea. The full proof is given in Appendix B. *1) Truthfulness:* The argument is based on the payments of the VCG mechanism. In particular, the payment rule p_i aligns agent i's utility with the negative

estimated social cost, that is, maximizing individual utility is equivalent to minimizing $\hat{J}(a; \mathcal{D})$. As the mechanism selects the socially efficient allocation under truthful feedback, agents have no incentive to misreport. The only inefficiencies arise from the learning error $\epsilon_K(\delta)$, which we control using the bounds for preference-based learning from Lemma 1.

2) Individual rationality: The VCG payment ensures that a truthful agent has non-negative utility, deteriorated only by the learning error $\epsilon_K(\delta)$. Unlike for numerical feedback, we need Assumption 2 to rule out constant terms in the cost function, which would be unidentifiable from preferences.

3) Efficiency: Finally, when considering the social cost, learning errors of the N agents add up. Taking the union bound over all agents results in at most a gap of $N\epsilon_K(\delta/N)$, relative to the optimal social cost.

4 MULTI-ROUND GAME: ALGORITHM AND ANALYSIS

Next, we consider a multi-round game, which extends the one-shot game to an online setting. The mechanism proceeds in stages that consist of an *exploration phase*, where the planner queries agents' preferences, followed by an *exploitation phase*, where the planner implements the allocation according to the VCG mechanism. This setting is closely related to no-regret learning in linear bandits (Lattimore and Szepesvári, 2020), where the learner must carefully balance exploration and exploitation to select informative queries while ensuring low regret. An example of such a setting is local electricity markets, where the planner must learn consumers' preferences for energy adjustments online, while also deploying such adjustments in real-time to ensure grid stability.

In order to study the multi-round game, we extend the results from the one-shot game. As before, we assume a linear cost model (Assumption 1) for agents, with the cost parameters fixed across rounds.

4.1 Algorithm

We introduce the multi-round algorithm, summarized in Algorithm 2. Inspired by Kandasamy et al. (2023), the algorithm alternates between exploration phases for learning and progressively longer exploitation phases with VCG allocations and payments. In contrast to their work, which assumes that agents give numerical feedback over finitely many allocations, we consider preference feedback over a compact allocation space. Their analysis does therefore not extend trivially to our setting and requires the methods developed in Section 3.

The multi-round algorithm differs from the one-shot setting in several ways. During exploration, the planner makes payments to the agents to ensure individual rationality in every round. These payments are set to $2c_{\max} = 2BL$, the maximum cost under Assumption 1, for each queried allocation pair. Moreover, each stage s consists of K exploration rounds and M_s exploitation rounds, where $M_s = \lfloor \frac{5}{6} K \sqrt{s} \rfloor$, so that exploitation phases progressively become longer as estimates improve.

Algorithm 2 Multi-round game: VCG mechanism with preference feedback

Input: K, T
$t \leftarrow 1$, $s \leftarrow 1$
while $t \leq T$ **do**
  /* Exploration phase */
  Select K comparisons for each agent i: $\{(a_{i,k}, a'_{i,k})\}_{k=1}^{K} \leftarrow$ D-Optimal-Design(K).
  Collect queried preferences $\{y^s_{i,k}\}_{k=1}^{K}$ from each agent i.
  Pay $2Kc_{\max}$ to each agent i.
  $t \leftarrow t + K$
  Estimate $\hat{\theta}^s_i$ from $\mathcal{D}^s_i$ via Equation (4).
  /* Exploitation phase */
  Compute $\hat{a}_t(\mathcal{D}^s)$ and $p_i(\hat{a}_t(\mathcal{D}^s))$ for all i via (6) during $M_s = \lfloor \frac{5}{6} K \sqrt{s} \rfloor$ rounds.
  $t \leftarrow t + M_s$
  $s \leftarrow s + 1$
end while

4.2 Theoretical guarantees

In the multi-round game, we focus on cumulative quantities over rounds $t = 1, \ldots, T$. Let the cumulative utility of agent i be defined as $\bar{U}_i(T) = \sum_{t=1}^{T} u_i(\hat{a}_t(\bar{\mathcal{D}}_i, \mathcal{D}_{-i}))$, where agent i is truthful in every round. Similarly, let $U_i(T) = \sum_{t=1}^{T} u_i(\hat{a}_t(\mathcal{D}_i, \mathcal{D}_{-i}))$ be the cumulative utility of agent i, when deviating from truthful feedback in at least one round. Furthermore, we define welfare regret as:

$$R^w(T) = \sum_{t=1}^{T} \big(J(\hat{a}_t(\bar{\mathcal{D}})) - J(a^*) \big), \tag{7}$$

where $\hat{a}_t(\bar{\mathcal{D}})$ is the mechanism's allocation computed from truthful feedback and a^* is the optimal allocation. Note that $R^w(T)$ also includes allocation costs of agents during exploration rounds.

Our goal is to provide regret bounds for these cumulative quantities, again from the mechanism's perspective. We show sublinear regret bounds, which imply asymptotic guarantees for truthfulness, individual rationality, and efficiency.

Theorem 2. *Let Assumption 1 hold, and suppose the mechanism makes $K > d(d+1)/2$ preference queries to each agent $i \in [N]$. Then, after T rounds with probability $1-\delta$, Algorithm 2 satisfies:*

1) ***Truthfulness:*** *An agent's utility gain from deviating in at least one round is upper bounded by:*

$$U_i(T) - \bar{U}_i(T) \in \tilde{\mathcal{O}}\left(dT^{2/3}\sqrt{\log(1/\delta)}\right).$$

2) ***Individual rationality:*** *Under Assumption 2, a truthful agent's utility is lower bounded by:*

$$-\bar{U}_i(T) \in \tilde{\mathcal{O}}\left(dT^{2/3}\sqrt{\log(1/\delta)}\right).$$

3) ***Efficiency:*** *With all agents truthful, the welfare regret (7) is upper bounded by:*

$$R^w(T) \in \tilde{\mathcal{O}}\left(NdT^{2/3}\sqrt{\log(N/\delta)}\right).$$

Compared to the approximate guarantees in the one-shot game, this theorem implies asymptotic truthfulness, individual rationality and efficiency. This follows directly from sublinearity of the upper bounds, for example, $R^w(T)/T \to 0$ as $T \to \infty$ for welfare regret. These rates are in line with works that consider numerical feedback and show sublinear upper bounds of $\tilde{O}(T^{2/3})$ (Babaioff et al., 2010). This aligns with the broader insight that, under the Bradley–Terry model, learning from preference feedback is not fundamentally harder than learning from numerical feedback (Ailon et al., 2014).

Proof idea. The full proof is given in Appendix C. *1) Truthfulness:* Utilities in exploration rounds are not affected by strategic behavior as allocations and payments are selected independently of the preference feedback. In exploitation rounds, the utility gain per round is upper bounded by Theorem 1 with $\epsilon_{Q_t}(\delta)$, where Q_t denotes the number of preference queries at time t. Summing over all rounds T yields the finite-time bound.
2) Individual rationality: Fixed payments of $2c_{\max}$ for a pairwise comparison guarantee non-negative utility in exploration rounds. In exploitation rounds, learning errors of $\epsilon_{Q_t}(\delta)$ per round sum up as above.
3) Efficiency: Due to repeated exploration, the learning error decreases over time, thus the mechanism converges to efficient allocations. The convergence rate is given by M_s, which ensures sublinear welfare regret by progressively longer exploitation phases as estimates get better.

Our theoretical results show that even when agents only provide preference feedback, a mechanism can be implemented that converges to a socially efficient allocation. A practical application for this setting arises in local electricity markets, which we discuss next.

5 EXPERIMENTS

An emerging challenge with the increased penetration of intermittent renewable energy in electric grids is the balancing of energy supply and demand. A promising approach is demand response, where a grid operator procures energy flexibility from domestic consumers through recurring *demand response events*. The allocation of this flexibility can be organized as local electricity markets (Tsaousoglou et al., 2022). This naturally raises questions about strategic behavior among consumers in these markets. Accordingly, a range of works study this problem and propose various market mechanisms (Tsaousoglou et al., 2021; Fochesato et al., 2022; Crowley et al., 2025). However, an open challenge is how consumers can communicate the costs they incur when providing energy flexibility, which are difficult to quantify (Abedrabboh and Al-Fagih, 2023). We address this challenge in our setting, where consumers express preferences over allocations of energy flexibility.

5.1 Simulation setup

We consider a group of consumers who adjust their thermal loads, such as heating and cooling, during demand response events. The consumers participate in a local electricity market, where a central planner coordinates the demand response. The planner allocates the energy flexibility $a = (a_1, \dots, a_N)$ with the goal to minimize the social cost. The individual cost of consumers is given by their discomfort when the room temperature deviates from their preferred temperature due to their energy flexibility a_i (kWh). We assume that each consumer has up to d rooms, which each contributes to the discomfort as follows:

$$c_i(a) = \langle \theta_i^*, \phi_i(a_i) \rangle = \sum_{l=1}^{d} \theta_{i,l}^* a_{i,l}^2,$$

where θ_i^* (\$/kWh^2) are positive cost parameters and $\phi_i(a_i)$ (kWh^2) is a quadratic feature map (Li et al., 2011).[2] When a demand response event occurs, the central planner must choose an allocation from the feasible set $\mathcal{F}$, defined as

$$\mathcal{F} = \left\{ a \in \mathcal{A} \;\middle|\; \sum_{i=1}^{N}\sum_{l=1}^{d} a_{i,l} = P \right\},$$

[2]While our theoretical guarantees allow costs c_i to depend on the entire allocation a, our simulations consider restricted feature maps $\phi_i(a_i)$ for illustration purposes.

Table 1: Maximum utility gain of a given agent with untruthful feedback for $K = 1000$ queries.

Deviation	Pay-as-Bid		VCG	
	Max	Std	Max	Std
-0.2	-0.40	0.07	-0.04	0.03
-0.1	-0.14	0.06	$+0.01$	0.03
$+0.1$	$+0.19$	0.05	$+0.03$	0.02
$+0.2$	$+0.22$	0.05	$+0.01$	0.03

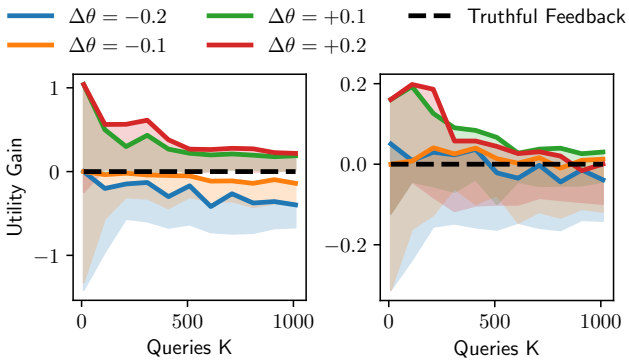
where P (kWh) is the total energy flexibility required by the grid. In this setting, preference queries can be thought of as simulated demand response events, such as questionnaires or tests.

5.2 Numerical results

In the simulation, energy flexibilities belong to a discrete allocation set of size $|\mathcal{A}_i| = 15$ for each agent i. Note that this implies 15^2 possible comparisons per agent, making exhaustive exploration infeasible. The discretization allows us to compute an exact D-optimal design, even though the theoretical results hold for any compact allocation space. See Appendix A.2 for a discussion of approximate optimal designs in the case of infinite sets. The cost parameters are generated synthetically with random samples between 0.1 and 0.5 \$/kWh^2 (Safdar et al., 2019), reflecting heterogeneity across consumers and rooms. We assume a constant flexibility requirement $P = 15$ kWh for each demand response event. An overview of parameters and runtimes is given in Appendix E. Next, we present our simulation results, illustrating truthfulness and efficiency.

To assess truthfulness in the one-shot game, we compare a given agent's utility under truthful and untruthful preference feedback. We focus on untruthful feedback where preferences are sampled from the Bradley–Terry model (3), but for biased cost parameters $\theta_i = \theta_i^* + \Delta\theta$, where $\Delta\theta \in \{-0.2, -0.1, 0.1, 0.2\}$. As a benchmark, we compare utilities for allocations and payments computed under the pay-as-bid (first-price) mechanism based on the learned costs. Figure 1 shows that, under the pay-as-bid mechanism, agents benefit from overstating their costs. Under the VCG mechanism, however, untruthful feedback can only increase an agent's utility when the number of queries K is small.

Table 1 provides an overview of a given agent's maximum utility gain for $K = 1000$ preference queries. Under the VCG mechanism, the incentive to deviate is negligible, once sufficient preferences are queried. Standard deviations are small, indicating that the agent's utility gains across runs become concentrated as the number of queries increases. This trend generalizes to any agent, confirming that the mechanism elicits truthful preference feedback.



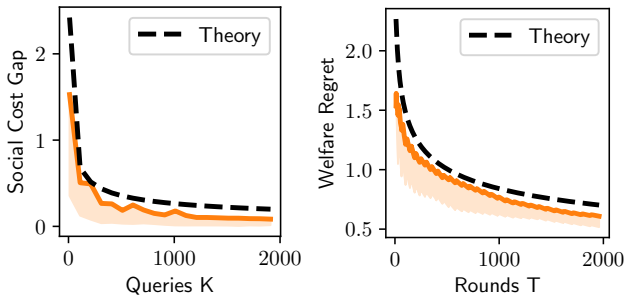
(a) Pay-as-Bid (b) VCG

Figure 1: Impact of misreporting on a given agent's utility in the one-shot game under (a) the pay-as-bid mechanism and (b) the VCG mechanism. We show the utility gain $u_i(\hat{a}(\mathcal{D}_i, \mathcal{D}_{-i})) - u_i(\hat{a}(\bar{\mathcal{D}}_i, \mathcal{D}_{-i}))$ across independent runs, with the worst-case outcome highlighted.

Finally, we assess efficiency under truthful feedback both in the one-shot and multi-round settings. In the one-shot game, Figure 2a shows that the worst-case social cost gap decreases at a rate of $\tilde{\mathcal{O}}(K^{-1/2})$. In the multi-round game, Figure 2b shows that the average welfare regret decreases a rate of $R^w(T)/T \in \tilde{\mathcal{O}}(T^{-1/3})$. Both rates are consistent with our theoretical guarantees. Overall, these results confirm that the mechanism converges to a socially efficient allocation by learning agents' costs from preference feedback.

6 CONCLUSION

In this paper, we studied the problem of designing preference queries and payment rules to ensure truthfulness, individual rationality, and efficiency. We proposed an algorithm that integrates preference-based learning with VCG, and proved that it satisfies these properties approximately in a one-shot setting and asymptotically in an online setting. Our method improves practicality by removing the need for bid functions or numerical feedback, while accommodating large allocation spaces. Lastly, we validated our method in a numerical case study on demand response in local electricity markets, demonstrating its ability to elicit truthful preferences and converge to socially efficient allocations.



(a) One-Shot Game (b) Multi-Round Game

Figure 2: Mechanism's performance in terms of (a) social cost gap $J(\hat{a}(\bar{\mathcal{D}})) - J(a^*)$ in the one-shot game and (b) average welfare regret $R^w(T)/T$ in the multi-round game. Both are normalized by the optimal social cost $J(a^*)$.

Based on these results, several directions for future work emerge. First, our mechanism builds on VCG, which is known to suffer from limitations such as lack of budget balance and potential revenue loss for the central planner (Ausubel et al., 2006). Future work could therefore explore alternative mechanism designs, such as core-selecting or ascending auctions (Ausubel and Milgrom, 2002), within a preference-based learning setup. Second, while the Bradley–Terry model is a standard cardinal preference model widely used in the literature, it remains restrictive in settings where preferences exhibit richer structure. An interesting direction would be to incorporate models that account for full reward distributions or alignment distortions, as recently studied by Gölz et al. (2025). Extending the theoretical guarantees to more general models, including time-varying costs or richer function classes (e.g., RKHS), and establishing robustness guarantees under model misspecification remain important open challenges. Finally, validating the approach using real-world preference data in simulations can highlight its practical effectiveness.

Acknowledgments

The authors thank the reviewers for their constructive and insightful feedback. The authors also thank the members of the Sycamore lab for valuable discussions and feedback.

References

Abedrabboh, K. and Al-Fagih, L. (2023). Applications of mechanism design in market-based demand-side management: A review. *Renewable and Sustainable Energy Reviews*, 171(113016).

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

Ausubel, L. and Milgrom, P. (2002). Ascending auctions with package bidding. *Frontiers of theoretical economics*, 1(1):1019.

Ausubel, L. M., Milgrom, P., et al. (2006). The lovely but lonely vickrey auction. *Combinatorial auctions*, 17(3):22–26.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. (2024). A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Babaioff, M., Kleinberg, R. D., and Slivkins, A. (2010). Truthful mechanisms with implicit payment computation. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 43–52.

Babaioff, M., Sharma, Y., and Slivkins, A. (2009). Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 79–88.

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

Chowdhury, S. R., Zhou, X., and Natarajan, N. (2024). Differentially private reward estimation with preference feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 4843–4851. PMLR.

Chremos, I. V. and Malikopoulos, A. A. (2024). Mechanism design theory in control engineering: A tutorial and overview of applications in communication, power grid, transportation, and security systems. *IEEE Control Systems Magazine*, 44(1):20–45.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Clarke, E. H. (1971). Multipart pricing of public goods. *Public choice*, pages 17–33.

Crowley, B., Kazempour, J., and Mitridati, L. (2025). How can energy communities provide grid services? a dynamic pricing mechanism with budget balance, individual rationality, and fair allocation. *Applied Energy*, 382(125154).

Devanur, N. R. and Kakade, S. M. (2009). The price of truthfulness for pay-per-click auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 99–106.

Fallah, A., Jordan, M., and Ulichney, A. (2024). Fair allocation in dynamic mechanism design. *Advances in Neural Information Processing Systems*, 37:125935–125966.

Fochesato, M., Cenedese, C., and Lygeros, J. (2022). A stackelberg game for incentive-based demand response in energy markets. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2487–2492. IEEE.

Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

Góis, A., Mofakhami, M., Santos, F. P., Lacoste-Julien, S., and Gidel, G. (2025). Performative prediction on games and mechanism design. In *International Conference on Artificial Intelligence and Statistics*, pages 1855–1863. PMLR.

Gölz, P., Haghtalab, N., and Yang, K. (2025). Distortion of ai alignment: Does preference optimization optimize for preferences? *arXiv preprint arXiv:2505.23749*.

Groves, T. (1973). Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631.

Hazan, E. and Karnin, Z. (2016). Volumetric spanners: an efficient exploration basis for learning. *The Journal of Machine Learning Research*, 17(1):4062–4095.

Kandasamy, K., Gonzalez, J. E., Jordan, M. I., and Stoica, I. (2023). VCG mechanism design with unknown agent values under stochastic bandit feedback. *Journal of Machine Learning Research*, 24(53):1–45.

Karlin, A. R. and Peres, Y. (2017). *Game theory, alive*, volume 101. American Mathematical Soc.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. (2023). Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

Li, N., Chen, L., and Low, S. H. (2011). Optimal demand response based on utility maximization in power networks. In *2011 IEEE power and energy society general meeting*, pages 1–8. IEEE.

Li, S., Lian, J., Conejo, A. J., and Zhang, W. (2020). Transactive energy systems: The market-based coordination of distributed energy resources. *IEEE Control Systems Magazine*, 40(4):26–52.

Luce, R. D. et al. (1959). *Individual choice behavior*, volume 4. Wiley New York.

Pereira, B. L., Ueda, A., Penha, G., Santos, R. L., and Ziviani, N. (2019). Online learning to rank for sequential music recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 237–245.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Safdar, M., Hussain, G. A., and Lehtonen, M. (2019). Costs of demand response from residential customers' perspective. *Energies*, 12(9):1617.

Scheid, A., Boursier, E., Durmus, A., Jordan, M. I., Ménard, P., Moulines, E., and Valko, M. (2024). Optimal design for reward modeling in rlhf. *arXiv preprint arXiv:2410.17055*.

Schlaginhaufen, A., Ouhamma, R., and Kamgarpour, M. (2025). Efficient preference-based reinforcement learning: Randomized exploration meets experimental design. *arXiv preprint arXiv:2506.09508*.

Strzalecki, T. (2025). Stochastic choice theory. *Cambridge Books*.

Tsaousoglou, G., Giraldo, J. S., and Paterakis, N. G. (2022). Market mechanisms for local electricity markets: A review of models, solution concepts and algorithmic techniques. *Renewable and Sustainable Energy Reviews*, 156(111890).

Tsaousoglou, G., Giraldo, J. S., Pinson, P., and Paterakis, N. G. (2021). Mechanism design for fair and efficient dso flexibility markets. *IEEE transactions on smart grid*, 12(3):2249–2260.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37.

Yazdani-Damavandi, M., Neyestani, N., Shafie-khah, M., Contreras, J., and Catalao, J. P. (2017). Strategic behavior of multi-energy players in electricity markets as aggregators of demand side resources using a bi-level approach. *IEEE Transactions on Power Systems*, 33(1):397–411.

Yue, Y. and Joachims, T. (2009). Interactively optimizing information retrieval systems as a dueling

bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208.

Zhang, G. and Duan, J. (2024). VickreyFeedback: Cost-efficient data construction for reinforcement learning from human feedback. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 351–366. Springer.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

A Technical results

This section presents the technical results necessary for our theorems' proofs.

A.1 Confidence set

First, we state an established lemma for maximum likelihood estimation given preference feedback according to the Bradley–Terry model (3). We include it for completeness, as it is used to control the learning error in our theorems.

Lemma 2 (Schlaginhaufen et al. (2025), Lemma A.1)**.** *Let Assumption 1 hold and let $\hat{\theta}_i$ be defined as in Equation (4). Then, after K queries, we get with probability at least $1-\delta$:*

$$\|\hat{\theta}_i - \theta_i^*\|_{V_{i,K}} \leq \gamma_{i,K}(\delta).$$

The bound is given as

$$\gamma_{i,K}^2(\delta) = \kappa_i \left[\log\left(\frac{1}{\delta}\right) + d \log\left(\max\left\{ e, \frac{4eBL(K-1)}{d} \right\} \right) \right].$$

where $\kappa_i = \max_{\|\theta_i\| \leq B,\, x_i \in \mathcal{X}_i} \frac{1}{\dot{\sigma}(\langle \theta_i, x_i \rangle)}$.

Here, $\|x\|_A := \langle x, Ax \rangle$ denotes the Mahalanobis norm given a positive definite matrix $A \in \mathbb{R}^{d \times d}$. This norm with respect to the empirical design matrix $V_{i,K} := \sum_{k=1}^{K} x_{i,k}\, x_{i,k}^\top$ measures the learning error for a given query $x_{i,k} := \phi_i(a_{i,k}) - \phi_i(a'_{i,k})$. The leading factor κ_i characterizes the difficulty of learning from preferences when the feedback is nearly deterministic.

A.2 Optimal experimental design

Since the set of all queries $\mathcal{X}_i = \{\phi_i(a') - \phi_i(a) : a, a' \in \mathcal{A}\}$ could be large or infinite, we leverage optimal experimental design to select a set of queries that optimally explore the allocation space $\mathcal{A}$. We want to control the worst-case prediction variance

$$\max_{x_i \in \mathcal{X}_i} \|x_i\|^2_{V_i(\pi_i)^{-1}}, \quad \text{with} \quad V_i(\pi_i) := \int_{\mathcal{X}_i} x_i x_i^\top d\pi_i(x_i),$$

where $V_i(\pi_i)$ is referred to as the theoretical design matrix associated with a distribution of queries π_i. While directly optimizing this G-optimality criterion is in general not straightforward, the Kiefer–Wolfowitz theorem (Kiefer and Wolfowitz, 1960) states the equivalence between G-optimality and D-optimality. We thus compute a D-optimal design

$$\pi_i^* \in \operatorname*{argmax}_{\pi_i \in \Delta_{\mathcal{X}_i}} \log\det V_i(\pi_i), \tag{8}$$

where $\Delta_{\mathcal{X}_i}$ is the set of all probability measures supported on the compact subset $\mathcal{X}_i \subset \mathbb{R}^d$. Note that compactness of the design space $\mathcal{X}_i$ follows from continuity of ϕ_i and compactness of $\mathcal{A}$. We can solve (8) efficiently for finite allocation sets $\mathcal{A}$, e.g., via the Frank-Wolfe algorithm (Frank et al., 1956). For infinite sets, Corollary 4.1 of (Hazan and Karnin, 2016) shows that, given access to an argmax oracle that can identify the optimal allocation, one can find a $\mathcal{O}(\sqrt{d})$-approximate G-optimal design of support d using $\mathcal{O}(d^2 \log d)$ calls to the argmax oracle. Algorithm 3 details the subroutine D-Optimal-Design(K) to select the queries.

Our lemma below bounds the prediction variance when queries are selected based on a D-optimal design. The result is shown given an exact π_i^*, but can be extended for approximations. This lemma is used to control the learning error in our theorems, together with Lemma 2.

Algorithm 3 D-Optimal-Design(K)

1: **Input:** K
2: Compute optimal design π_i^* (8) with $|\text{Supp}(\pi_i^*)| \leq d(d+1)/2$.
3: Set $n \leftarrow K - d(d+1)/2$.
4: Round number of queries $n_{x_i} \leftarrow \lceil n\pi_i^*(x_i) \rceil$ for comparison x_i.
5: Select queries $\{(a_{i,k}, a'_{i,k})\}_{k=1}^K$ from the rounded design.

Lemma 3. *Suppose $K > d(d+1)/2$ queries are selected using Algorithm 3. Then, the worst-case prediction variance is bounded as:*

$$\max_{x_i \in \mathcal{X}_i} \|x_i\|^2_{V_{i,K}^{-1}} \leq \frac{d}{K - d(d+1)/2}.$$

Proof. Our goal is to bound the worst-case prediction variance

$$\max_{x_i \in \mathcal{X}_i} \|x_i\|^2_{V_{i,K}^{-1}},$$

where $V_{i,K} = \sum_{k=1}^K x_{i,k} x_{i,k}^\top$ is the *empirical* design matrix. Our argument builds on the Kiefer–Wolfowitz theorem (Kiefer and Wolfowitz, 1960), which characterizes the properties of D-optimal designs with respect to the *theoretical* design matrix $V_i(\pi_i)$, and follows the ideas in Lattimore and Szepesvári (2020, Section 21.1) to relate the theoretical and empirical design matrices. By the Kiefer–Wolfowitz theorem, a D-optimal design π_i^* satisfies

$$\max_{x_i \in \mathcal{X}_i} \|x\|^2_{V_i(\pi_i^*)^{-1}} = d, \tag{9}$$

where $V_i(\pi_i) = \int_{\mathcal{X}_i} x_i x_i^\top d\pi_i(x_i)$ is the theoretical design matrix, based on the distribution π_i rather than sampled queries. Moreover, the theorem states that there exists a D-optimal design π_i^* with support

$$|\text{Supp}(\pi_i^*)| \leq d(d+1)/2. \tag{10}$$

This result bounds the worst-case prediction variance for the theoretical design matrix, which we now relate to the empirical design matrix. Let $n_{x_i} = \lceil n\pi_i^*(x_i) \rceil$ denote the number of times we query x_i. Then, the empirical design matrix is

$$V_{i,K} = \sum_{x_i \in \text{Supp}(\pi_i^*)} n_{x_i} x_i x_i^\top = \sum_{x_i \in \text{Supp}(\pi_i^*)} \lceil n\pi_i^*(x) \rceil x_i x_i^\top \succeq n V_i(\pi_i^*) = n \sum_{x_i \in \text{Supp}(\pi_i^*)} \pi_i^*(x_i) x_i x_i^\top.$$

By (9), it follows that

$$\max_{x_i \in \mathcal{X}_i} \|x_i\|^2_{V_{i,K}^{-1}} \leq \frac{1}{n} \max_{x_i \in \mathcal{X}_i} \|x_i\|^2_{V_i(\pi_i^*)^{-1}} = \frac{d}{n}.$$

The rounded design results in a gap between n and K, which we express using (10) as

$$K = \sum_{x_i \in \text{Supp}(\pi_i^*)} n_{x_i} \leq \sum_{x_i \in \text{Supp}(\pi_i^*)} (n\pi_i^*(x) + 1) \leq n + \frac{d(d+1)}{2},$$

so that $n \geq K - d(d+1)/2$. Substituting into the previous bound yields the lemma. □

A.3 Proof of Lemma 1

Proof. We want to bound the learning error for pairwise preferences:

$$\begin{aligned}
\left|(c_i(a) - c_i(a')) - \left(\hat{c}_i(a; \bar{\mathcal{D}}_i)\right) - \hat{c}_i(a'; \bar{\mathcal{D}}_i)\right)\right| &= \left|\left(\hat{c}_i(a'; \bar{\mathcal{D}}_i) - c_i(a')\right) + \left(\hat{c}_i(a; \bar{\mathcal{D}}_i) - c_i(a)\right)\right| \\
&= \left|\langle \hat{\theta}_i - \theta_i^*, \phi_i(a') \rangle - \langle \hat{\theta}_i - \theta_i^*, \phi_i(a) \rangle\right| \\
&= \left|\langle \hat{\theta}_i - \theta_i^*, \phi_i(a') - \phi_i(a) \rangle\right|, \qquad (11)
\end{aligned}$$

which we rewrote using Assumption 1. Note that $\hat{\theta}_i$ is the cost parameter learned from truthful feedback. Then, by the Cauchy-Schwarz inequality, we have:

$$\begin{aligned}\left|(c_i(a) - c_i(a')) - \left(\hat{c}_i(a; \bar{\mathcal{D}}_i)) - \hat{c}_i(a'; \bar{\mathcal{D}}_i)\right)\right| &\leq \|\hat{\theta}_i - \theta_i^*\|_{V_{i,K}} \|\phi_i(a') - \phi_i(a)\|_{V_{i,K}^{-1}} \\ &\leq \epsilon_K(\delta),\end{aligned}$$

where

$$\epsilon_K(\delta) := \gamma_K(\delta) \sqrt{\frac{d}{K - d(d+1)/2}}.$$

Here, we combined the confidence bound for $\|\hat{\theta}_i - \theta_i^*\|_{V_{i,K}}$ (Lemma 2) and the prediction variance bound for $\|\phi_i(a') - \phi_i(a)\|_{V_{i,K}^{-1}}$ (Lemma 3). We set $\gamma_K(\delta) = \max_i \gamma_{i,K}(\delta)$, with

$$\gamma_{i,K}(\delta) = \sqrt{\kappa_i \left[\log\left(\frac{1}{\delta}\right) + d \log\left(\max\left\{e, \frac{4eBL(K-1)}{d}\right\}\right)\right]}. \tag{12}$$

Hence, with probability at least $1 - \delta$, after K comparisons we have

$$\epsilon_K(\delta) \in \mathcal{O}\left(\sqrt{\left[\frac{d \log(1/\delta) + d^2 \log(K)}{K}\right]}\right).$$

□

B Proof of Theorem 1

Proof. We prove approximate truthfulness, individual rationality, and efficiency in the one-shot game. For each property, the VCG mechanism reduces the analysis to bounding the learning error, which we control via Lemma 1.

Truthfulness. We aim to bound agent i's utility gain from giving untruthful feedback, given other agents' arbitrary feedback $\mathcal{D}_{-i}$. On the one hand, if agent i provides truthful feedback $\bar{\mathcal{D}}_i$, the allocation $\hat{a} := \hat{a}(\bar{\mathcal{D}}_i, \mathcal{D}_{-i})$ minimizes

$$\hat{a} \in \operatorname*{argmin}_{a \in \mathcal{F}} \left(\hat{J}_{-i}(a; \mathcal{D}_{-i}) + \hat{c}_i(a; \bar{\mathcal{D}}_i)\right). \tag{13}$$

On the other hand, if agent i gives arbitrary feedback $\mathcal{D}_i$, the allocation $\hat{a}' := \hat{a}(\mathcal{D}_i, \mathcal{D}_{-i})$ minimizes

$$\hat{a}' \in \operatorname*{argmin}_{a \in \mathcal{F}} \left(\hat{J}_{-i}(a; \mathcal{D}_{-i}) + \hat{c}_i(a; \mathcal{D}_i)\right).$$

Now, consider agent i's utility difference under arbitrary and truthful feedback:

$$\begin{aligned}u_i(\hat{a}') - u_i(\hat{a}) &\overset{(i)}{=} \left[\min_{a \in \mathcal{A}} \hat{J}_{-i}(a; \mathcal{D}_{-i}) - \hat{J}_{-i}(\hat{a}'; \mathcal{D}_{-i}) - c_i(\hat{a}')\right] \\ &\quad - \left[\min_{a \in \mathcal{A}} \hat{J}_{-i}(a; \mathcal{D}_{-i}) - \hat{J}_{-i}(\hat{a}; \mathcal{D}_{-i}) - c_i(\hat{a})\right] \\ &\overset{(ii)}{=} -\hat{J}_{-i}(\hat{a}'; \mathcal{D}_{-i}) - c_i(\hat{a}') + \hat{J}_{-i}(\hat{a}; \mathcal{D}_{-i}) + c_i(\hat{a}) \\ &\overset{(iii)}{=} \left[-\hat{J}_{-i}(\hat{a}'; \mathcal{D}_{-i}) - \hat{c}_i(\hat{a}'; \bar{\mathcal{D}}_i) + \hat{c}_i(\hat{a}'; \bar{\mathcal{D}}_i) - c_i(\hat{a}')\right] \\ &\quad + \left[\hat{J}_{-i}(\hat{a}; \mathcal{D}_{-i}) + \hat{c}_i(\hat{a}; \bar{\mathcal{D}}_i) - \hat{c}_i(\hat{a}; \bar{\mathcal{D}}_i) + c_i(\hat{a})\right].\end{aligned}$$

In (i), we applied the VCG payment rule (6). The term $\min_{a \in \mathcal{F}} \hat{J}_{-i}(a; \mathcal{D}_{-i})$ canceled in (ii), as it is independent from agent i's feedback. Then, (iii) added and subtracted the terms $\hat{c}_i(\hat{a}'; \bar{\mathcal{D}}_i)$ and $\hat{c}_i(\hat{a}; \bar{\mathcal{D}}_i)$, the agent's cost estimated under truthful feedback, but for allocations $\hat{a}'$ (untruthful feedback) and $\hat{a}$ (truthful feedback). Since $\hat{a}$ is the truthful minimizer in (13), we have $\hat{J}_{-i}(\hat{a}; \mathcal{D}_{-i}) + \hat{c}_i(\hat{a}; \bar{\mathcal{D}}_i) \leq \hat{J}_{-i}(\hat{a}'; \mathcal{D}_{-i}) + \hat{c}_i(\hat{a}'; \bar{\mathcal{D}}_i)$, and hence:

$$\begin{aligned}u_i(\hat{a}') - u_i(\hat{a}) &\leq \hat{c}_i(\hat{a}'; \bar{\mathcal{D}}_i) - c_i(\hat{a}') - \hat{c}_i(\hat{a}; \bar{\mathcal{D}}_i) + c_i(\hat{a}) \\ &\leq \left|(c_i(\hat{a}) - c_i(\hat{a}')) - \left(\hat{c}_i(\hat{a}; \bar{\mathcal{D}}_i) - \hat{c}_i(\hat{a}'; \bar{\mathcal{D}}_i)\right)\right|.\end{aligned}$$

But this is exactly the term we bound in Lemma 1, so with probability at least $1-\delta$,

$$u_i(\hat{a}') - u_i(\hat{a}) \leq \epsilon_K(\delta). \tag{14}$$

Individual rationality. Next, we show that a truthful agent's utility is lower bounded. Letting $\hat{a} := \hat{a}(\bar{\mathcal{D}}_i, \mathcal{D}_{-i})$, we have

$$\begin{aligned} u_i(\hat{a}) &\overset{(i)}{=} \left(\min_{a\in\mathcal{F}} \hat{J}_{-i}(a;\mathcal{D}_{-i}) - \hat{J}_{-i}(\hat{a};\mathcal{D}_{-i})\right) - c_i(\hat{a}) \\ &\overset{(ii)}{=} \min_{a\in\mathcal{F}} \hat{J}_{-i}(a;\mathcal{D}_{-i}) - \hat{J}_{-i}(\hat{a};\mathcal{D}_{-i}) - \hat{c}_i(\hat{a};\bar{\mathcal{D}}_i) + \hat{c}_i(\hat{a};\bar{\mathcal{D}}_i) - c_i(\hat{a}), \end{aligned}$$

where in (i) we substituted the VCG payments in the utility and in (ii) $\hat{c}_i(\hat{a};\bar{\mathcal{D}}_i)$ was added and subtracted. Observe that $\min_{a\in\mathcal{F}} \left(\hat{J}_{-i}(\hat{a};\mathcal{D}_{-i}) + \hat{c}_i(a;\bar{\mathcal{D}}_i)\right) \leq \min_{a\in\mathcal{F}} \hat{J}_{-i}(a;\mathcal{D}_{-i})$, because the mechanism can always ignore agent i if including them does not reduce total cost. Their participation can only lower or maintain the social cost, thus

$$\begin{aligned} u_i(\hat{a}) &\geq \hat{c}_i(\hat{a};\bar{\mathcal{D}}_i) - c_i(\hat{a}) \\ &= \hat{c}_i(\hat{a};\bar{\mathcal{D}}_i) - c_i(\hat{a}) + c_i(a^0) - \hat{c}_i(a^0;\bar{\mathcal{D}}_i) \\ &\geq -\left|\left(c_i(a^0) - c_i(\hat{a})\right) - \left(\hat{c}_i(a^0;\bar{\mathcal{D}}_i) - \hat{c}_i(\hat{a};\bar{\mathcal{D}}_i)\right)\right|, \end{aligned}$$

where we added $c_i(a^0) - \hat{c}_i(a^0;\bar{\mathcal{D}}_i)$, both zero under Assumption 2. Then, by Lemma 1, we get with probability at least $1-\delta$:

$$u_i(\hat{a}) \geq -\epsilon_K(\delta). \tag{15}$$

Efficiency. Finally, we upper bound the gap to the optimal social cost $J(a^*)$ under truthful feedback. Defining $\hat{a} := \hat{a}(\bar{\mathcal{D}})$, we have:

$$\begin{aligned} J(\hat{a}) - J(a^*) &\overset{(i)}{=} J(\hat{a}) - \hat{J}(a^*;\bar{\mathcal{D}}) + \hat{J}(a^*;\bar{\mathcal{D}}) - J(a^*) \\ &\overset{(ii)}{\leq} J(\hat{a}) - \hat{J}(\hat{a};\bar{\mathcal{D}}) + \hat{J}(a^*;\bar{\mathcal{D}}) - J(a^*). \end{aligned}$$

In (i), we added and subtracted $\hat{J}(a^*;\bar{\mathcal{D}})$. Then, (ii) used that $\hat{J}(\hat{a};\mathcal{D}) \leq \hat{J}(a^*;\mathcal{D})$, because $\hat{a}$ is the minimizer of $J(\hat{a};\bar{\mathcal{D}})$. Separating the sum into individual costs gives us

$$\begin{aligned} J(\hat{a}) - J(a^*) &\leq \sum_{i=1}^{N} \left(c_i(\hat{a}) - \hat{c}_i(\hat{a};\bar{\mathcal{D}}_i) + \hat{c}_i(a^*;\bar{\mathcal{D}}_i) - c_i(a^*)\right) \\ &\leq \sum_{i=1}^{N} \left|\left(c_i(\hat{a}) - c_i(a^*)\right) - \left(\hat{c}_i(\hat{a};\bar{\mathcal{D}}_i) - \hat{c}_i(a^*;\bar{\mathcal{D}}_i)\right)\right|. \end{aligned}$$

By Lemma 1 and taking the union bound over N agents, we obtain with probability at least $1-\delta$,

$$J(\hat{a}) - J(a^*) \leq N\epsilon_K(\delta/N). \tag{16}$$

□

C Proof of Theorem 2

Proof. We prove asymptotic truthfulness, individual rationality, and efficiency in the multi-round game. Similar to Theorem 1, the analysis builds on the properties of the VCG mechanism and the learning error bound from Lemma 1. However, since we now consider an online setting, the properties need to be analyzed both for exploration and exploitation phases.

We write $\mathcal{T}_{\text{explore}}$ for the set of exploration rounds and $\mathcal{T}_{\text{exploit}}$ for the set of exploitation rounds. The key idea to achieve sublinear welfare regret is to balance exploration length K and exploitation length M_s. In our algorithm, which separates exploration and exploitation, the exploitation length M_s must progressively become longer as estimates improve. We use the following lemma for our proofs.

Lemma 4 (Kandasamy et al. (2023), Lemma 18)**.** *Let $M_s = \lfloor \frac{5}{6} K \sqrt{s_t} \rfloor$ for stage s_t in round t. Then,*

$$s_t \leq 3K^{-2/3} t^{2/3}, \quad \text{if } t \in \mathcal{T}_{explore},$$
$$\frac{1}{2} K^{-2/3} t^{2/3} \leq s_t, \quad \text{if } t \in \mathcal{T}_{exploit}.$$

Each stage consists of an exploration phase followed by an exploitation phase. Hence, the number of queries Q_t up to round t is given by:

$$Q_t \leq K s_t, \quad \text{if } t \in \mathcal{T}_{explore}, \tag{17}$$
$$Q_t = K s_t, \quad \text{if } t \in \mathcal{T}_{exploit}. \tag{18}$$

Now, we can control the learning error for pairwise preferences using $\epsilon_{Q_t}(\delta)$.

Truthfulness. We want to bound the cumulative utility gain when agent i is untruthful. In exploration rounds, utilities are independent of agents' feedback, because allocations are selected based on a D-optimal design and payments are constant. For exploitation rounds, we have:

$$U_i(T) - \bar{U}_i(T) = \sum_{t \in \mathcal{T}_{\text{exploit}}} \left(u_i(\hat{a}_t(\mathcal{D}_i^{s_t}, \mathcal{D}_{-i}^{s_t})) - u_i(\hat{a}_t(\bar{\mathcal{D}}_i^{s_t}, \mathcal{D}_{-i}^{s_t})) \right),$$

We apply the one-shot truthfulness bound (14) and get

$$U_i(T) - \bar{U}_i(T) \leq \sum_{t \in \mathcal{T}_{\text{exploit}}} \epsilon_{Q_t}(\delta),$$

where Q_t is the number of queries up to round t. As $\epsilon_{Q_t}(\delta) \geq 0$, we extend the sum to all rounds such that

$$U_i(T) - \bar{U}_i(T) \leq \sum_{t=1}^{T} \epsilon_{Q_t}(\delta) = \sum_{t=1}^{T} \gamma_t(\delta) \sqrt{\frac{d}{Q_t - d(d+1)/2}}. \tag{19}$$

Using Equation (18) with Lemma 4 to lower bound Q_t, we get with probability at least $1 - \delta$:

$$U_i(T) - \bar{U}_i(T) \lesssim \sum_{t=1}^{T} \gamma_t(\delta) \sqrt{2d} K^{-1/6} t^{-1/3} \leq \frac{3}{2} \gamma_T(\delta) \sqrt{2d} K^{-1/6} T^{2/3}. \tag{20}$$

Here, $\lesssim$ indicates that we dropped the additive rounding term $d(d+1)/2$ in the denominator, since it does not affect asymptotic behavior. Then, we used monotonicity of $\gamma_t(\delta)$ (12) and the identity $\sum_{t=1}^{T} t^{-1/3} \leq \frac{3}{2} T^{2/3}$.

Individual rationality. Next, we consider a truthful agent's utility under allocations $\hat{a}_t := \hat{a}_t(\bar{\mathcal{D}}_i^{s_t}, \mathcal{D}_{-i}^{s_t})$. In exploration rounds, the mechanism makes constant payments $p_i(\hat{a}_t) = c_{\max}$ for each queried allocation. We set $c_{\max} = BL$ under Assumption 1, then agents have utility

$$u_i(a_t) = p_i(\hat{a}_t) - c_i(\hat{a}_t) = c_{\max} - c_i(\hat{a}_t) \geq 0.$$

For exploitation rounds, we apply the one-shot individual rationality bound (15) to obtain

$$\bar{U}_i(T) = \sum_{t \in \mathcal{T}_{\text{exploit}}} u_i(\hat{a}_t) \geq \sum_{t \in \mathcal{T}_{\text{exploit}}} -\epsilon_{Q_t}(\delta).$$

Note that this bound uses Assumption 2. By the same arguments as in (19) and (20), we get with probability at least $1 - \delta$:

$$\bar{U}_i(T) \gtrsim \sum_{t=1}^{T} -\epsilon_{Q_t}(\delta) \geq -\frac{3}{2} \gamma_T(\delta) \sqrt{2d} K^{-1/6} T^{2/3}.$$

Efficiency. Finally, we want to upper bound the welfare regret (7) when all agents provide truthful feedback. We decompose the total regret as

$$R^w(T) = \sum_{t=1}^{T} r_t = \sum_{t \in \mathcal{T}_{\text{explore}}} r_t + \sum_{t \in \mathcal{T}_{\text{exploit}}} r_t,$$

where $r_t = J(\hat{a}_t(\bar{\mathcal{D}}^{s_t})) - J(a^*)$ is the instantaneous regret at time t. In exploration rounds, the social cost for a pairwise comparison is at most $2\,J_{\max}$, with an upper bound $J_{\max} = NBL$ under Assumption 1. Then, we have

$$\begin{aligned}\sum_{t \in \mathcal{T}_{\text{explore}}} r_t &\overset{(i)}{\leq} 2J_{\max} Q_T \\ &\overset{(ii)}{\leq} 6NBLK^{1/3}T^{2/3}.\end{aligned}$$

In (i), we used that $r_i \leq 2J_{\max}$ for Q_T total queries, and (ii) used Equation (17) with Lemma 4 to upper bound Q_T. For exploitation rounds, we apply the one-shot efficiency bound (16). By the same arguments as in (19) and (20), and taking a union bound over N agents, we get with probability at least $1-\delta$:

$$\sum_{t \in \mathcal{T}_{\text{exploit}}} r_t \lesssim \sum_{t=1}^{T} N\epsilon_{Q_t}(\delta/N) \leq \frac{3}{2} N\gamma_T(\delta/N)\sqrt{2d}K^{-1/6}T^{2/3}.$$

Combining the bounds yields with probability at least $1-\delta$:

$$R^w(T) \lesssim 6NBLK^{1/3}T^{2/3} + \frac{3}{2} N\gamma_T(\delta/N)\sqrt{2d}K^{-1/6}T^{2/3}.$$

□

D Unknown rationality parameter

We extend the Bradley–Terry model (3) by introducing an agent-specific rationality parameter $\beta_i > 0$, also known as the inverse temperature. The probability that agent i prefers allocation $a \in \mathcal{A}$ over allocation $a' \in \mathcal{A}$ is then given by:

$$\begin{aligned}\Pr(a \succ a') &= \sigma\big(\beta_i[u_i(a) - u_i(a')]\big) \\ &= \sigma\big(\beta_i[p_i(a) - c_i(a) - p_i(a') + c_i(a')]\big),\end{aligned}$$

where $\sigma(x) = 1/(1+e^{-x})$ denotes the sigmoid function. Under Assumption 1 and defining $\Delta p_i := p_i(a) - p_i(a')$, we equivalently write:

$$\Pr(a \succ a') = \sigma\big(\langle \beta_i\theta_i^*, \phi_i(a') - \phi_i(a)\rangle + \beta_i \Delta p_i\big). \tag{21}$$

In the main paper and the subsequent analysis, we assumed β_i to be known and, without loss of generality, set $\beta_i = 1$. This assumption in the context of agent-specific $\beta_i > 0$ can also be interpreted as agents with different levels of rationality who adjust their thinking times, so that the central planner can learn θ_i^* (Strzalecki, 2025).

In practice, however, such an adjustment may be difficult for agents. To learn θ_i^*, the central planner then also needs to estimate β_i. To this end, we outline a simple subroutine for the mechanism to learn β_i in the one-shot game. Following the earlier procedure in Algorithm 1, the mechanism first computes a scaled cost parameter $\tilde{\theta}_i \approx \beta_i\theta_i^*$. Now, the subroutine queries preferences over two allocations $a, a' \in \mathcal{A}$ and makes payments with $\Delta p_i \neq 0$. To compute $\hat{\beta}_i$, in principle any pair of allocations could be used, as all quantities except β_i in Equation (21) are known or have been estimated. To simplify the analysis, we consider the case $a = a'$, which removes the dependence on $\tilde{\theta}_i$. Then, the agent's truthful preference is expressed by

$$\bar{y}_i \sim \text{Bernoulli}\big(\sigma(\beta_i \Delta p_i)\big).$$

From these preferences, the social planner can estimate $\hat{\beta}_i$ via maximum likelihood. By Lemma 2, we can find high-probability bounds after K_1 and K_2 queries, respectively, such that:

$$\Pr(\|\tilde{\theta}_i - \beta_i\theta_i^*\|_{V_{i,K_1}} \leq \gamma_1) \geq 1 - \delta_1, \tag{22}$$

$$\Pr(|\hat{\beta}_i - \beta_i| \leq \gamma_2) \geq 1 - \delta_2, \tag{23}$$

where $\gamma_1 \in \mathcal{O}(\log(K_1))$ and $\gamma_2 \in \mathcal{O}(\log(K_2)K_2^{-1/2})$. The estimation error of the scaled parameter $\hat{\theta}_i$ is

$$\begin{aligned}\hat{\theta}_i - \theta_i^* &\overset{(i)}{=} \frac{\tilde{\theta}_i}{\hat{\beta}_i} - \frac{\beta_i\theta_i^*}{\beta_i} \\ &\overset{(ii)}{=} \frac{\tilde{\theta}_i - \beta_i\theta_i^*}{\hat{\beta}_i} + \beta_i\theta_i^*\left(\frac{1}{\hat{\beta}_i} - \frac{1}{\beta_i}\right).\end{aligned}$$

In (i), we used that $\hat{\theta}_i = \tilde{\theta}_i/\hat{\beta}_i$ and (ii) added and subtracted $\beta_i\theta_i^*/\hat{\beta}_i$. Plugging this term in (11), we get

$$\begin{aligned}\left|(c_i(a) - c_i(a')) - \left(\hat{c}_i(a;\bar{\mathcal{D}}_i)) - \hat{c}_i(a';\bar{\mathcal{D}}_i)\right)\right| &= \left|\langle\hat{\theta}_i - \theta_i^*, \phi_i(a') - \phi_i(a)\rangle\right| \\ &= \left|\left\langle\frac{\tilde{\theta}_i - \beta_i\theta_i^*}{\hat{\beta}_i} + \beta_i\theta_i^*\left(\frac{1}{\hat{\beta}_i} - \frac{1}{\beta_i}\right), \phi_i(a') - \phi_i(a)\right\rangle\right|.\end{aligned}$$

Applying the triangle inequality gives

$$\begin{aligned}\left|(c_i(a) - c_i(a')) - \left(\hat{c}_i(a;\bar{\mathcal{D}}_i)) - \hat{c}_i(a';\bar{\mathcal{D}}_i)\right)\right| \leq &\left|\left\langle\frac{\tilde{\theta}_i - \beta_i\theta_i^*}{\hat{\beta}_i}, \phi_i(a') - \phi_i(a)\right\rangle\right| \\ &+ \left|\left\langle\beta_i\theta_i^*\left(\frac{1}{\hat{\beta}_i} - \frac{1}{\beta_i}\right), \phi_i(a') - \phi_i(a)\right\rangle\right|.\end{aligned}$$

For the first term, we apply the Cauchy-Schwarz inequality and get with probability at least $1 - \delta_1$:

$$\left|\left\langle\frac{\tilde{\theta}_i - \beta_i\theta_i^*}{\hat{\beta}_i}, \phi_i(a') - \phi_i(a)\right\rangle\right| \leq \frac{\|\tilde{\theta}_i - \beta_i\theta_i^*\|_{V_{i,K_1}}}{|\hat{\beta}_i|}\|\phi_i(a') - \phi_i(a)\|_{V_{i,K_1}^{-1}} \leq \frac{\gamma_1}{|\hat{\beta}_i|}\sqrt{\frac{d}{K_1 - d(d+1)/2}},$$

using the bound (22) and Lemma 3. For the second term, we have with probability at least $1 - \delta_2$:

$$\left|\left\langle\beta_i\theta_i^*\left(\frac{1}{\hat{\beta}_i} - \frac{1}{\beta_i}\right), \phi_i(a') - \phi_i(a)\right\rangle\right| = \frac{|\hat{\beta}_i - \beta_i|}{|\hat{\beta}_i|}|\langle\theta_i^*, \phi_i(a') - \phi_i(a)\rangle| \leq \frac{\gamma_2}{|\hat{\beta}_i|}2BL,$$

using the bound (23) and Assumption 1. Finally, combining the bounds and setting $K_1 = K_2 = K$, we get with probability at least $1 - \delta_1 - \delta_2$:

$$\left|(c_i(a) - c_i(a')) - \left(\hat{c}_i(a;\bar{\mathcal{D}}_i)) - \hat{c}_i(a';\bar{\mathcal{D}}_i)\right)\right| \leq \frac{2\gamma_1}{\beta_i}\sqrt{\frac{d}{K - d(d+1)/2}} + \frac{4BL\gamma_2}{\beta_i} = \tilde{\mathcal{O}}\left(\frac{\epsilon_K(\delta_1 + \delta_2)}{\beta_i}\right),$$

where we assumed that $\gamma_2 \leq \beta_i/2$ to lower bound $\hat{\beta}_i \geq \beta_i/2$. Overall, this discussion shows that, given that agents are somewhat rational, unknown rationality parameters do not affect the properties of the mechanism.

E Experimental Configurations

Table 2 provides the configurations used for experiments in Section 5. All experiments were run on a MacBook Pro 2023 with an M2 Pro chip and 16 GB of RAM.

Table 2: Overview of experiment parameters and runtimes.

	Truthfulness Fig. 1	**Social Cost Gap** Fig. 2a	**Welfare Regret** Fig. 2b
Parameters	`K = 10,110,...,1010`	`K = 10,110,...,1910`	`K = 5, T = 2000`
Repetitions	20	50	20
Runtime	1:36 h	0:23 h	3:46 h