POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

# A Theory-Driven Approach to Large Language Models Alignment with Human Feedback

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

**Author:** Michele Simeone

**Advisor:** Prof. Alberto Maria Metelli

**Co-advisors:** Tommaso Bianchi, Simone Drago, Gianmarco Genalti

**Academic year:** 2023-2024

---

## 1. Introduction

Decision-making in artificial intelligence (AI) often requires selection among multiple potential outcomes, each of which carries distinct implications and trade-offs. Chatbots and Large Language Models (LLMs) usually generate numerous possible responses to a user query, and the most appropriate one must be selected based on contextual understanding and predefined objectives. Human feedback mechanisms, such as Reinforcement Learning from Human Feedback (RLHF) [1], play an integral role in refining these selections to ensure alignment with user expectations and ethical considerations [2]. With the advancement of increasingly powerful LLMs, the financial and computational resources required for aligning high-performance models have become prohibitively expensive for all but a few organizations, which rely on executing techniques that have no theoretical guarantees that they will work. This implies the need to explore alternative approaches to align LLMs, leveraging response characteristics as a source to perform this process, and provide associated sample complexity. The primary objective of this work is to understand the theoretical aspects of the LLM alignment problem, providing an algorithm with theoretical guarantees and good experimental performance at least in a simplified scenario. We focus on the study of the sample complexity associated with the proposed approach and empirically evaluate its effectiveness in real world data. In Section 2, we introduce mathematical modeling and formal problem formulation. Specifically, we model the offline feasibility problem, where the constraints are represented by the contexts associated with the two inputs, and the output is an estimator of the target's preference. In Section 3, we present our proposed algorithm for aligning an LLM and derive its sample complexity. Finally, in Section 4, we provide a numerical evaluation of our algorithm and discuss the results.

## 2. Problem Formulation and Methods

Our interaction protocol begins with a user providing a query $Q$, in response to which two possible answers are generated: $A_1$ and $A_2$. In general, the entities indicated as $Q$, $A_1$ and $A_2$ can correspond to objects of different types such as text, images and even music files. However, within the scope of this study, these elements will be conceptualized as an interaction with an

LLM that acts as a chatbot. Each answer is also associated with a vector, $\mathbf{c}_1$ for $A_1$ and $\mathbf{c}_2$ for $A_2$, both belonging to $\mathbb{R}^n$. They represent the properties of the corresponding responses and are called *contexts*, these can be easily extracted from the responses using existing models. Specifically, considering $n$ different properties of the answers (for example the length or the stylistic register), the $j$-th element of the vector represents a score for that category called $[c_{i,1}, \ldots, c_{i,n}]$ with $i \in \{1, 2\}$. The set of contexts fully characterizes each response. Note that the user which gives us the preference cannot explicitly observe either $\mathbf{c}_1$ or $\mathbf{c}_2$.

Furthermore, we consider a unknown preference vector $\mathbf{v}^* \in \mathbb{R}^n$, which represents the preference for each context of the human decision-maker. Each element of this vector, $v_i^*$, can assume a value within the interval $[-1, 1]$.

We assume that there exists a function $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, which, given an input vector $\mathbf{c}$ (representing a response) and the preference vector $\mathbf{v}^*$, returns a numerical score. This score represents how well the response characterized by $\mathbf{c}$ aligns with human preferences as described by $\mathbf{v}^*$. The comparison of the scores $f(\mathbf{c}_1, \mathbf{v}^*)$ and $f(\mathbf{c}_2, \mathbf{v}^*)$ determines which response the human would prefer, with three possible outcomes:

- $f(\mathbf{c}_1, \mathbf{v}^*) < f(\mathbf{c}_2, \mathbf{v}^*)$ indicating a preference for response $A_2$.
- $f(\mathbf{c}_1, \mathbf{v}^*) > f(\mathbf{c}_2, \mathbf{v}^*)$ indicating a preference for response $A_1$.
- $f(\mathbf{c}_1, \mathbf{v}^*) = f(\mathbf{c}_2, \mathbf{v}^*)$ indicating indifference between the two responses.

Our objective is to identify a feasible region $\mathbf{V} \in \mathbb{R}^n$ within the $n$-dimensional hyperspace that contains the preference vector $\mathbf{v}^*$.

## 2.1. Offline Feasibility Problem

To define the feasibility problem rigorously, we introduce the necessary notation and underlying structure. Let $\mathbb{D} = \{(\mathbf{c}_1^i, \mathbf{c}_2^i, p_i)\}_{i=1}^k$ be a dataset consisting of $k$ samples, where each sample $i$ is associated with two feature vectors, $\mathbf{c}_1^i$ and $\mathbf{c}_2^i$. Additionally, each sample is accompanied by a preference label $p_i$, which takes a binary value in $\{0, 1\}$, indicating the preferred option between the two, specifically if the $p_i = 0$ the first option is preferred, otherwise the second is the chosen one. There also also two unknown distributions $\mathcal{G}, \mathcal{Z}$ such that $\mathbf{c}_1 \sim \mathcal{G}, \mathbf{c}_2 \sim \mathcal{Z}$, however, the preferences associated with each row are deterministic given $\mathbf{c}_1$ and $\mathbf{c}_2$.

The goal is to determine a preference vector $\hat{\mathbf{v}} \in \mathbb{R}^n$ that best fits a given set of pairwise comparisons. The problem is structured so that the preference vector $\hat{\mathbf{v}}$ must satisfy constraints that encode the preference relationships observed in the data. Specifically, suppose a given preference label is $p_i = 0$, the difference of the application of $f$ with context vectors and $\mathbf{v}^*$ must be nonnegative, ensuring that the first vector is at least as preferred as the second. Mathematically, this is expressed as $p_i = \mathbb{1}_{\{f(\mathbf{c}_1,\mathbf{v}^*) - f(\mathbf{c}_2,\mathbf{v}^*) < 0\}}$.

The feasibility problem is formulated by enforcing these conditions as finding $\hat{\mathbf{v}}$ that best satisfies the observed preference relationships.

## 2.2. Model Assumption

We introduce a set of assumptions to restrict the focus of this work while preserving its applicability to real-world scenarios. Each assumption reflects the problem's underlying structure and defines the preference function's mathematical properties, ensuring a reusable theoretical background to provide mathematical analyses of the problem.

One fundamental assumption is that the function $f(\mathbf{c}, \mathbf{v}) = \mathbf{c}^\top \mathbf{v}$, exhibits linearity with respect to both input vectors. In this way the human preference scores can be interpreted as a weighted sum of categorical attributes, with $\mathbf{v}$ encoding the corresponding weight coefficients. This assumption is justified by the fact that features can be extracted from contexts in order to make the preference function approximately linear.

We also consider that $\|\mathbf{v}^*\| = 1$ . This assumption is justified by the fact that, within the given setting, all points that are scalar multiples of the same unit vector have equivalent solutions. In other words, they share the same proportional relationships among the dimensions under consideration.

In preference-based learning framework it is crucial that we can always distinguish between options. Thus, we assume that we never encounter situations where two samples are indistinguishable in terms of their alignment with $\mathbf{v}^*$, in other words this is equivalent $\mathbb{P}_{\mathcal{G},\mathcal{Z}}((\mathbf{c}_1 - \mathbf{c}_2)^\top \mathbf{v}^* \neq 0) = 1$ .

Finally, we assume that the elements of the feature vectors are mutually independent and this simplifies both the modeling process and computational complexity.

## 3. Algorithm and Sample Complexity

We now present the Cutting Plane Preference Learner (CPPL) algorithm that leverages the feasibility region derived from the dataset to construct a preference estimation model and make informed decisions on new samples.

**Algorithm 1** Cutting Plane Preference Learner

```
1:  Collect a dataset $\mathbb{D}$: $\mathbb{D} \leftarrow \{(\mathbf{c}_1^i, \mathbf{c}_2^i, p_i)\}_{i=1}^k$
2:  Find the feasibility region $\mathbf{V}$ defined by the dataset $\mathbb{D}$: find_FP($\mathbb{D}$) $\rightarrow \mathbf{V}$
3:  Compute the mean of $\mathbf{V}$: mean($\mathbf{V}$) $\rightarrow \mathbf{v}$
4:  Normalize $\mathbf{v}$: normalize($\mathbf{v}$) $\rightarrow \hat{\mathbf{v}}$
5:  Receive new sample: $(a_1, \tilde{\mathbf{c}}_1, a_2, \tilde{\mathbf{c}}_2)$
6:  Compute preference prediction:
        $\hat{y}_{\hat{\mathbf{v}}} \leftarrow \begin{cases} 0, & \text{if } (\tilde{\mathbf{c}}_1 - \tilde{\mathbf{c}}_2)^T \hat{\mathbf{v}} > 0 \\ 1, & \text{otherwise} \end{cases}$
7:  if $\hat{y}_{\hat{\mathbf{v}}} = 0$ then
8:      Assign $A \leftarrow a_1$
9:  else
10:     Assign $A \leftarrow a_2$
11: end if
```

Given that multiple valid preference vectors may exist in $\mathbf{V}$, in line 3, the algorithm computes their mean to derive a representative preference vector $\mathbf{v}$. The choice of the midpoint, defined as the mean for each coordinate of all feasible vectors extracted from $\mathbf{V}$, minimizes the worst-case error that arises when the true preference vector $\mathbf{v}^*$ resides at the extremes of the identified region. In line 4, the preference vector $\mathbf{v}$ is normalized to ensure it has unit length. This normalization step is crucial for maintaining consistency in comparative computations. At this stage, the algorithm is equipped with an estimated preference vector $\hat{\mathbf{v}}$, which serves as the foundation for predicting the user's choice. Specifically, the algorithm is designed to identify and suggest the response most aligned with the user's preferences. Once the feasibility region $\mathbf{V}$ has been sufficiently reduced and the candidate to represent the user's preferences has been chosen, for each new sample the algorithm does not offer the user a choice but directly suggests the most suitable answer by providing the output $A$.

### 3.1. Context Vector Distribution Dissection

The feasibility problem is influenced by the statistical properties of the feature vectors $\mathbf{c}_1$ and $\mathbf{c}_2$, whose underlying distribution is unknown and dependent on the LLM. This introduces complexity, requiring an investigation into the associated random variables. The distribution of the difference $\mathbf{c}_1 - \mathbf{c}_2$ is crucial, as it directly impacts feasibility conditions and sample complexity.
Assuming each element $c_{1j}$ and $c_{2j}$ results from $d$ independent Bernoulli trials, their values follow a binomial distribution:$c_{1j} \sim \text{Binomial}(d, m_{1j})$, $c_{2j} \sim \text{Binomial}(d, m_{2j})$ where $m_{ij}$ is the success probability in each trial. The distribution of the difference $c_{1j} - c_{2j}$ is fully determined by $m_{1j}$ and $m_{2j}$, governing the probabilities of the binomial processes.

### 3.2. Sample Complexity

To establish an upper bound on the number of samples $k$ required to achieve a small error, we use the feasibility region $\mathbf{V}$ obtained by solving the offline feasibility problem as a measure of error. This region represents the set of feasible solutions within which the true preference vector $\mathbf{v}^*$ is expected to lie.
By quantifying how $\mathbf{V}$ shrinks as the number of samples increases, we can determine the conditions under which the estimated preference vector $\hat{\mathbf{v}}$ remains sufficiently close to $\mathbf{v}^*$.
The Euclidean norm between the optimal solution $\mathbf{v}^*$ and its corresponding estimator $\hat{\mathbf{v}}$ is defined as the distance $\hat{d}$:

$$\hat{d} = \|\hat{\mathbf{v}} - \mathbf{v}^*\| \ . \tag{1}$$

Accordingly, our objective is to ensure that the probability $\mathbb{P}(\hat{d} < \epsilon)$ exceeds $1 - \delta$, i.e.,

$$\mathbb{P}(\hat{d} < \epsilon) > 1 - \delta \ . \tag{2}$$

In other words, we seek to guarantee, with high probability, that the error associated with $\hat{d}$ remains below a predefined threshold $\epsilon$.

### 3.3. 2d Setting

We initially restrict our consideration to the two-dimensional (2d) scenario to analyze this case study. To derive the sample complexity we need to approximate the constraints' vector distribution with a Gaussian one using the Berry-Esseen theorem [3, 4]. This simplification introduces two irreducible error terms $\Delta_{d,m}$ and $I_d$,both decreasing as a function of $d$:

$$I_d, \Delta_{d,m} = \mathcal{O}\left(\frac{1}{\sqrt{dm(1-m)}}\right) . \tag{3}$$

The sample complexity presented is the following:

$$k \geq \frac{\log\left(\frac{2}{\delta}\right)}{\log\left(\frac{1}{\frac{2\pi-\frac{\theta}{2}}{2\pi}+2\Delta_{d,m}+I_d}\right)} , \tag{4}$$

this is valid only when $\frac{2\pi-\frac{\theta}{2}}{2\pi} + 2\Delta_{d,m} + I_d \leq 1$.

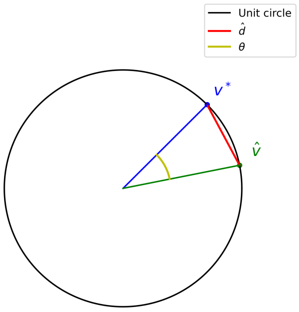


Figure 1: 2d sample complexity setting visualization.

As can be observed in Figure 1, the distance $\hat{d}$ depends mainly on the two closest constraints to $\mathbf{v}^*$. To simplify our approach, we introduce $\theta$, which represents the angle corresponding to $\hat{d}$ in a unit circle. Since the constraints are derived from the arctangent of a ratio of two Gaussian variables, the term $I_d$ in Equation (4) accounts for cases where the denominator of this ratio is zero in the original Binomial distribution and $\Delta_{d,m}$ come from approximating the Binomial with a Gaussian. The presence of $\Delta_{d,m}$ and $I_d$ impacts the sample complexity by imposing a lower bound on the angle $\theta$, preventing it from being reduced arbitrarily. This constraint arises because there exists a region $\mathbf{V}^*$ in which all points represent equivalent solutions to our problem. Now using basic geometric arguments, we deduce that $\|\hat{\mathbf{v}} - \mathbf{v}^*\| \leq 2\sin\left(\frac{\theta}{2}\right)$.

### 3.4. Generalization in $n$-dimension

To extend sample complexity analysis from the two-dimensional case to an arbitrary $n$-dimensional setting, we begin by examining the 2-norm of the difference between $\mathbf{v}^*$ and $\hat{\mathbf{v}}$ in higher dimensions. In an $n$-dimensional space, we can divide this 2-norm using $\lceil \frac{n}{2} \rceil$ different angular parameters $(\theta_1, \theta_2, \ldots, \theta_{\lceil \frac{n}{2} \rceil})$, where each of these angles determines the orientation of constraints. Moreover each angular parameter $\theta_i$ is independently and uniformly distributed over the interval $[-\pi, \pi]$, which forms the basis for our probabilistic analysis.
As done in the 2d case, we set the amplitude for each angle to $\theta$. Geometrically, this condition ensures that the feasible configuration is constrained within a hypercube whose volume we want proportional to $\epsilon$. Figure 2 shows this setting in a 3d space.

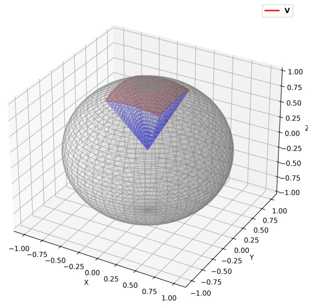


Figure 2: 3d sample complexity setting visualization.

This led to our generalized sample complexity:

$$k \geq \frac{\sqrt{\lceil \frac{n}{2} \rceil}\log\left(\frac{2\lceil \frac{n}{2} \rceil}{\delta}\right)}{\log\left(\frac{1}{\frac{2\pi-\frac{\theta}{2}}{2\pi}+2\Delta_{d,m}+I_d}\right)} , \tag{5}$$

to ensure that $\|\hat{\mathbf{v}} - \mathbf{v}^*\| \leq \theta$. This result demonstrates that the sample complexity scales as square root in the dimensionality $n$. This reflects the intrinsic geometric complexity of higher-dimensional spaces: as the number of angular parameters increases, the difficulty of obtaining high-probability configurations increases.

# 4. Experiment and Results

## 4.1. Comparison between Binomial Dataset and Real Dataset

We now move to the experimental analysis by conducting an in-depth examination of the dataset selected for evaluating the proposed algorithm.

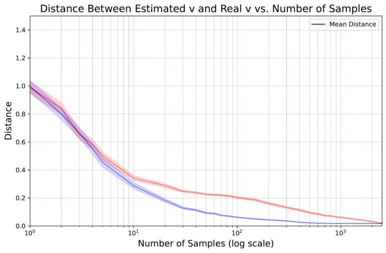
The Preference Dissection dataset [5] comprises 5,240 observations distributed across 18 columns, each representing various AI-generated responses and their corresponding evaluations under different contextual scenarios.

We conducted a series of experiments using hybrid data, consisting both of real samples from the dataset and artificially generated samples derived from fitted distributions. Specifically, we generated new data points based on the statistical properties of selected categories, yielding two new vectors: $\mathbf{c}_{1,g}$ and $\mathbf{c}_{2,g}$ (obtained from Gaussian fits).

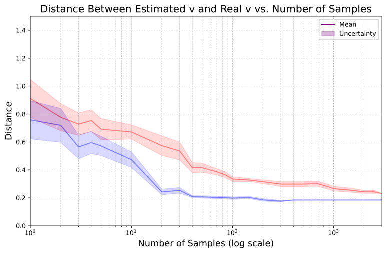
The objective of this experiment is to analyze the impact of using different representations of the response vectors in reconstructing preferences. To this end, we compared the performance of the system when using different variations of $\mathbf{c}$ while keeping a fixed reference vector $\mathbf{v}^*$, which is manually selected. The preference is then reconstructed on a sample-by-sample basis according to the corresponding values of $\mathbf{c}_1$ and $\mathbf{c}_2$.

To ensure statistical robustness, we repeated this experiment 100 times, permuting the dataset in each run. This was necessary because the sequential order in which samples are processed influences the partitioning of the feasibility region. The final results were obtained by averaging across all runs.

Figure 3 quantifies the variation in the distance between the fixed reference vector $\mathbf{v}^*$ and the estimated preference vector $\hat{\mathbf{v}}$. This metric is directly related to the sample complexity considerations discussed in the previous section. The x-axis, which represents the number of evaluated samples, is displayed on a logarithmic scale because as the number of processed samples increases, further reductions in the feasibility area become increasingly difficult to achieve. Consequently, the earliest samples contribute more significantly to shaping the feasibility region. The experiment was carried out in both a 2d and a 3d setting to demonstrate how the dimensionality of the context vector plays a fundamental role in scaling the number of samples needed to shrink the V region.



(a) 2d setting.



(b) 3d setting.

Figure 3: Distance between $\hat{\mathbf{v}}$ and $\mathbf{v}^*$ vs number of samples with 100 permutations of the dataset using Generated Binomial values (in blue) vs Original Dataset (in red).

As expected, in the two-dimensional case, we achieve a smaller average distance compared to its three-dimensional counterpart but it is notable to highlight that in all experimental settings, the initial feasibility region is effectively constrained using only a small subset of the dataset.

A less intuitive but notable result is that our experimental setting is also well-aligned with the structural characteristics of the original dataset. Specifically, the results obtained using the binomial distribution closely resemble those derived from the original dataset, indicating that our approximated approach successfully captures key contextual properties of the data.

## 4.2. Test the Cutting Plane Preference Learner

As a final experiment, we assess the predictive performance of the Cutting Plane Preference Learner algorithm. The experimental setup consisted of dividing the circumference into eight equal segments, deriving eight fixed points, de-

noted as $\mathbf{v}^*$. For each phase, we reconstructed the preference vector associated with the respective fixed point and trained the algorithm using the training portion of the dataset. The trained model was then evaluated on the test set to assess its generalization capabilities. As a baseline, we used a majority voting algorithm, which aggregated the preference vectors of the 31 LLMs extracted from the original dataset. Our algorithm exhibits desirable approximation capabilities in the two-dimensional case, accurately estimating the fixed point $\mathbf{v}^*$ within each sector. Even in sectors where the discrepancy between the estimated and actual points increases, the model successfully captures the underlying structural patterns in the data.
In contrast, the majority voting approach, despite leveraging the expressiveness of 31 LLMs, shows systematic biases in certain regions of the circumference. The results are summarized in Table 1 and Table 2. The high values observed in the Cutting Plane Preference Learner algorithm metrics can be attributed to the low dimensionality and our assumptions, which render the problem linearly separable.

| **$v^*$ (coordinates)** | **Cutting Plane Preference Learner** | | |
|---|---|---|---|
| | Accuracy;Precision;Recall | $\hat{v}$ (coordinates) | Distance |
| $(1; 0.01)$ | 1;1;1 | $(0.92; 0.38)$ | 0.38 |
| $(0.71; 0.71)$ | 1;1;1 | $(0.71; 0.71)$ | 0 |
| $(0.01; 1)$ | 1;1;1 | $(0.16; 0.99)$ | 0.15 |
| $(-0.7; 0.72)$ | 1;1;1 | $(-0.59; 0.81)$ | 0.15 |
| $(-1; 0.01)$ | 1;1;1 | $(-0.92; 0.38)$ | 0.38 |
| $(-0.71; -0.71)$ | 1;1;1 | $(-0.71; -0.71)$ | 0 |
| $(0.01; -1)$ | 1;1;1 | $(0.2; -0.97)$ | 0.22 |
| $(0.72; -0.70)$ | 1;1;1 | $(0.92; -0.38)$ | 0.37 |

Table 1: Cutting Plane Preference Learner Performance.

| **$v^*$ (coordinates)** | **Majority Voting** | | |
|---|---|---|---|
| | Accuracy | Precision | Recall |
| $(1; 0.01)$ | 0.59 | 0.62 | 0.65 |
| $(0.71; 0.71)$ | 0.56 | 0.63 | 0.62 |
| $(0.01; 1)$ | 0.50 | 0.53 | 0.58 |
| $(-0.7; 0.72)$ | 0.25 | 0.26 | 0.32 |
| $(-1; 0.01)$ | 0.31 | 0.30 | 0.38 |
| $(-0.71; -0.71)$ | 0.47 | 0.49 | 0.55 |
| $(0.01; -1)$ | 0.75 | 0.74 | 0.81 |
| $(0.72; -0.70)$ | 0.69 | 0.70 | 0.75 |

Table 2: Majority Voting Performance.

## 5. Conclusions

In this work, we investigated the problem of selecting between two alternatives within the context of response alignment for LLMs. This problem was formulated as an offline feasibility problem, wherein the objective was to delineate the user's preference space by leveraging the context vectors associated with the two options as constraints. We provided theoretical guarantees regarding its resolution through the formulation of sample complexity. Additionally, we introduced an algorithm, *The Cutting Plane Preference Learner*, which was designed to manage the entire pipeline for aligning an LLM with human preferences. We developed a series of validation tests to ensure its correct functioning. These tests were executed on a real dataset, which we modified with synthetic data that adhered to our initial assumptions. A limitation of our approach stemmed from the constrained number of samples available for validation, which restricted our ability to extend the analysis to higher spaces. Nevertheless, in comparison to the existing state of the art, the research direction pursued in this work appeared promising, particularly concerning its sample efficiency in learning human preferences throughout the entire space.

## References

[1] Paul F Christiano et al. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[2] Long Ouyang et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[3] Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

[4] Carl-Gustav Esseen. *A moment inequality with an application to the central limit theorem*, volume 1956. Taylor & Francis, 1956.

[5] Junlong Li et al. Dissecting human and llm preferences. *arXiv preprint arXiv:2402.11296*, 2024.