

Strategyproof Reinforcement Learning from Human Feedback

Thomas Kleine Buening¹, Jiarui Gan², Debmalya Mandal³, Marta Kwiatkowska²

¹The Alan Turing Institute

²University of Oxford

³University of Warwick

Abstract

We study Reinforcement Learning from Human Feedback (RLHF), where multiple individuals with diverse preferences provide feedback strategically to sway the final policy in their favor. We show that existing RLHF methods are not strategyproof, which can result in learning a substantially misaligned policy even when only one out of k individuals reports their preferences strategically. In turn, we also find that any strategyproof RLHF algorithm must perform k -times worse than the optimal policy, highlighting an inherent trade-off between incentive alignment and policy alignment. We then propose a pessimistic median algorithm that, under appropriate coverage assumptions, is approximately strategyproof and converges to the optimal policy as the number of individuals and samples increases.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) has become a widely used approach for aligning AI systems with human preferences. By leveraging human-labeled comparisons, RLHF enables policy optimization in applications such as robotics, recommendation systems, and large language models (LLMs) [5, 18]. This approach has led to significant improvements in usability and alignment with intended objectives. However, RLHF also introduces new challenges, particularly in settings where preferences are diverse and subjective.

Recently, pluralistic alignment—the notion that different individuals or groups may have conflicting or varying preferences over AI behavior—has emerged as an active area of research [3, 6, 12, 19, 27]. Unlike traditional reinforcement learning, where a single, well-defined reward function governs optimal policy learning, pluralistic settings require reconciling multiple human perspectives. This raises questions about whose preferences should shape AI decisions and how to aggregate diverse inputs fairly and effectively. In such pluralistic settings, existing methods often optimize policies based on aggregated human preferences but typically assume that individuals provide honest feedback, ignoring the possibility of strategic misreporting [10].

When human preferences influence the final policy, individuals may have incentives to manipulate their feedback in ways that benefit them at the expense of broader alignment [5, 17, 24]. For example, in LLM fine-tuning, human labelers might systematically misreport preferences to amplify specific biases and reinforce narratives favorable to their views. More broadly, strategic behavior can distort AI alignment and undermine the fairness, robustness, and efficiency of the AI system [2, 11]. Despite its importance, this issue remains largely unaddressed in existing RLHF methodologies.

This paper aims to bridge this gap by studying RLHF through the lens of mechanism design and proposing solutions that ensure robustness against strategic behavior. We formalize the problem, analyze the conditions and the cost under which strategyproofness can be achieved, and propose a mechanism that mitigates incentive misalignment while maintaining policy performance.

In summary, our main contributions are:

- We formally introduce the problem of offline RLHF with strategic human labelers, where each labeler can misreport preference feedback to steer the final policy toward the maximization of their own objectives, i.e., their reward function (Section 3). We focus on linear reward functions and social welfare maximization and study the tensions that arise between individual incentive alignment and policy alignment with social welfare.
- We show that existing RLHF methods are not strategyproof (Proposition 3.3), and even a single strategic labeler can almost arbitrarily degrade policy performance of existing methods (Proposition 3.4). Moreover, without additional assumptions, we find that any strategyproof RLHF method suffers from constant suboptimality (Theorem 3.5) and performs at least k -times worse compared to the optimal policy (Corollary 3.6), suggesting a fundamental trade-off between incentive alignment and policy alignment.
- We propose an RLHF method called Pessimistic Median of MLEs, which combines pessimistic estimates with a median rule to incentivize truthful preference reporting (Section 4). Interestingly, we find that Pessimistic Median of MLEs is *approximately* strategyproof due to the uncertainty in reward estimation. Notably, the incentive strength depends on the *uniform* policy coverage of each labeler’s data. This stands in contrast to standard RLHF guarantees, which rely only on the coverage of the optimal policy. More precisely, under additional domain restrictions, we show:
 - Pessimistic Median of MLEs is $\tilde{\mathcal{O}}(\kappa_i \sqrt{d/n})$ -strategyproof for labeler $i \in [k]$, where κ_i quantifies uniform policy coverage (Theorem 4.1).
 - The computed policy’s suboptimality is bounded by $\tilde{\mathcal{O}}(\sqrt{d/k} + \max_{i \in [k]} \kappa_i^* \cdot k \sqrt{d/n})$, where κ_i^* denotes the optimal policy coverage for labeler i (Theorem 4.2, Proposition 4.3).

We establish these results for both contextual bandits and Markov decision processes (Section 5).

2 Related Work

Reinforcement Learning from Human Feedback. RLHF has emerged as a powerful framework for aligning AI systems with human values by leveraging human feedback to guide policy learning [5, 9, 18]. Most relevant to our work is the growing literature on RLHF in settings with diverse and possibly conflicting preferences among individuals or demographic groups [7, 23, 30]. Some of these works focus on maximizing the worst-case utility across labelers (or groups) [7, 23], whereas others optimize welfare functions such as social welfare [30]. Some other recent work has also explicitly taken a social choice perspective on pluralistic alignment and studies how to ensure that aggregation methods satisfy desirable properties inspired by social choice theory [1, 10, 13]. All of these works assume truthful feedback, without explicitly accounting for the incentives of human labelers. However, aggregating and trading off preferences naturally invites strategic or malicious behavior, as labelers may manipulate the aggregation process to better align the final policy with their own beliefs and goals. For example, Siththaranjan et al. [26] highlighted how standard RLHF methods implicitly aggregate preferences using the Borda count voting rule, which can create incentives for annotators to misreport their preferences to influence model behavior.

Another body of work considers robustness against adversarial corruption in RLHF. Mandal et al. [20] assume that an ε -fraction of samples is adversarially manipulated, allowing for both manipulation of trajectory features and human feedback. Similarly, Bukharin et al. [4] and Cheng et al. [8] also consider the case where a fraction of samples is manipulated but restrict their attention to adversaries flipping preferences. This line of work differs from ours notably in both perspective (strategic vs. adversarial) and techniques (mechanism design vs. robust offline RL).

Mechanism Design for RLHF. Recently, several works have incorporated mechanism design principles into RLHF to incentivize truthful feedback [22, 28, 29]. These approaches design payment rules to align labelers’ incentives, often extending VCG-style mechanisms to the RLHF setting. In contrast, we propose a strategy-robust RLHF method that does not rely on payments or financial incentives, which are often impractical in real-world applications. Also closely related is the work of Hao and Duan [16], which studies an online RLHF framework where labelers sequentially provide preference feedback, aggregated using a linearly weighted average. Their approach focuses on identifying the most accurate labeler over time and adjusting weights to incentivize truthful reporting. In contrast, we study the offline RLHF setting and do not impose linear weighting assumptions on labelers. Moreover, unlike [16], we assume that labelers seek to influence the final policy rather than just the estimated reward function, which better reflects real-world strategic behavior, where individuals care about the actual policy outcomes rather than intermediate preference estimates.

3 Problem Formulation

We consider episodic Markov Decision Processes (MDPs) and the special case of contextual bandits. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, H, \rho)$ be an MDP without reward function, where \mathcal{S} is the state space, \mathcal{A} is the action space, H is the horizon and ρ is the initial state distribution. $\mathcal{P} = (\mathcal{P}_1, \dots, \mathcal{P}_H)$ denotes the tuple of transition functions, where $\mathcal{P}_h: \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ determines the transitions in step $h \in [H]$. A history-independent policy $\pi = (\pi_h)_{h \leq H}$ maps from states to a distribution over actions in every time step, i.e., $\pi_h: \mathcal{S} \rightarrow \Delta(\mathcal{A})$, and we let Π denote its policy space. A trajectory in MDP \mathcal{M} is given by a sequence of actions and states $\tau = (a_1, s_2, a_2, \dots, s_H, a_H)$. The MDP reduces to a contextual bandit problem when $H = 1$, in which case a trajectory consists only of the action taken in the initial state and the initial states are interpreted as contexts sampled from ρ .

Multiple Labelers with Diverse Preferences. We consider the situation where $k \geq 1$ many human labelers provide preference data to the RLHF algorithm. In particular, each labeler $i \in [k]$ is associated with a reward function $r_i: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. The expected return of a policy π w.r.t. a reward function r is given by $V_r^\pi(s) := \mathbb{E}[\sum_{h=1}^H r(s_h, \pi_h(s_h)) \mid s_1 = s]$.¹ Accordingly, we define the *utility* of labeler $i \in [k]$ w.r.t. the initial state distribution ρ and a policy π as

$$J_i(\rho, \pi) := \mathbb{E}_{s \sim \rho} [V_{r_i}^\pi(s)].$$

Note that this simplifies to $J_i(\rho, \pi) = \mathbb{E}_{s \sim \rho} [r_i(s, \pi(s))]$ in the contextual bandit.

We focus on the linearly realizable case, where the reward function of each labeler is a linear function $r_\theta(s, a) = \langle \theta, \phi(s, a) \rangle$ of a known feature embedding ϕ of the state (i.e., context) and action.

Assumption 1 (Linear Realizability). Every labeler’s reward function r_i is given by a linear function $r_{\theta_i^*}(s, a) := \langle \theta_i^*, \phi(s, a) \rangle$, where θ_i^* is sampled from $\{\theta \in \mathbb{R}^d: \|\theta\|_2 \leq B\}$ and ϕ is a known mapping.

Offline RLHF. We focus on the offline RLHF setting, where each labeler $i \in [k]$ is given a predetermined set of n examples $(s^{i,j}, \tau_0^{i,j}, \tau_1^{i,j})_{j \leq n}$, where $s^{i,j}$ denotes the initial state and $(\tau_0^{i,j}, \tau_1^{i,j})$ are two subsequent trajectories. For each example, labeler i provides a preference $o^{i,j} \in \{0, 1\}$, where $o^{i,j} = 0$ means that trajectory $\tau_0^{i,j}$ is preferred over $\tau_1^{i,j}$ given initial state $s^{i,j}$, and vice versa.²

As the comparison model, we consider the widely used Bradley-Terry-Luce (BTL) model under which a labeler with reward parameter θ (i.e., reward function r_θ) prefers trajectory $\tau_0 = (a_1, s_2, a_2, \dots, s_H, a_H)$

¹With slight abuse of notation we let $r(s, \pi(s)) = \mathbb{E}_{a \sim \pi(\cdot | s)} [r(s, a)]$ if the policy π is stochastic.

²To ease notation, we assume that every labeler provides preferences for the same number of examples n . This can be straightforwardly relaxed at the cost of additional notation and more cumbersome statements.

over trajectory $\tau_1 = (\tilde{a}_1, \tilde{s}_2, \tilde{a}_2, \dots, \tilde{s}_H, \tilde{a}_H)$ with probability

$$\mathbb{P}_\theta(o = 0 \mid s, \tau_0, \tau_1) := \frac{\exp(r_\theta(s, \tau_0))}{\exp(r_\theta(s, \tau_0)) + \exp(r_\theta(s, \tau_1))}, \quad (1)$$

where $r_\theta(s, \tau_0) := \sum_{h=1}^H r_\theta(s_h, a_h)$ is the total reward of the trajectory τ_0 with initial state $s_1 = s$. In a contextual bandit setting, where each trajectory consists of a single action, e.g., $\tau_0 = a_0$ and $\tau_1 = a_1$, the comparison model simplifies to

$$\mathbb{P}_\theta(o = 0 \mid s, a_0, a_1) \propto \exp(r_\theta(s, a_0)).$$

Strategic Misreporting. We assume that if labeler $i \in [k]$ reports their preferences *truthfully*, then the preference labels $o^{i,j}$ are sampled according to their true reward parameter θ_i^* , while the examples $(s^{i,j}, \tau_0^{i,j}, \tau_1^{i,j})$ remain fixed. Thus, the collected preference dataset under truthful reporting is given by

$$\mathcal{D}_i^* = (s^{i,j}, \tau_0^{i,j}, \tau_1^{i,j}, o^{i,j})_{j \leq n} \text{ with } o^{i,j} \sim \mathbb{P}_{\theta_i^*}(\cdot \mid s^{i,j}, \tau_0^{i,j}, \tau_1^{i,j}).$$

Since labelers aim to maximize their utility, they may strategically misreport their preferences to influence the final policy. To model this behavior, we allow each labeler i to report labels according to a manipulated reward parameter $\tilde{\theta}_i$, resulting in reported preferences

$$\tilde{\mathcal{D}}_i = (s^{i,j}, \tau_0^{i,j}, \tau_1^{i,j}, \tilde{o}^{i,j})_{j \leq n} \text{ with } \tilde{o}^{i,j} \sim \mathbb{P}_{\tilde{\theta}_i}(\cdot \mid s^{i,j}, \tau_0^{i,j}, \tau_1^{i,j}).$$

Given a reported preference dataset $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_k)$, an RLHF algorithm computes a policy $\hat{\pi}_{\text{RLHF}}(\mathcal{D}) \in \Pi$. We omit the argument \mathcal{D} when the preference data is clear from context.

Now, an RLHF algorithm is *strategyproof* if for every labeler truthfully sampling preferences according to their true reward function is a (weakly) dominant strategy.

Definition 3.1 (Strategyproofness). We say that the mapping $\hat{\pi}_{\text{RLHF}}(\mathcal{D})$ is strategyproof if for all $i \in [k]$, other labelers' data $\mathcal{D}_{-i} = (\mathcal{D}_1, \dots, \mathcal{D}_{-i}, \mathcal{D}_{i+1}, \dots, \mathcal{D}_k)$ and deviating $\tilde{\theta}_i \neq \theta_i^*$ it holds that

$$\mathbb{E}_{o^{i,j} \sim \mathbb{P}_{\theta_i^*}} [J_i(\hat{\pi}_{\text{RLHF}}(\mathcal{D}_i^*, \mathcal{D}_{-i}))] \geq \mathbb{E}_{\tilde{o}^{i,j} \sim \mathbb{P}_{\tilde{\theta}_i}} [J_i(\hat{\pi}_{\text{RLHF}}(\tilde{\mathcal{D}}_i, \mathcal{D}_{-i}))].$$

Note that this property is also referred to as *dominant strategy incentive compatibility (DSIC)*.

We can relax the strict incentive constraint by allowing labelers to have a limited incentive to misreport, leading to the notion of ε -strategyproofness.

Definition 3.2 (ε -Strategyproofness). We say that the mapping $\pi_{\text{RLHF}}(\mathcal{D})$ is ε -strategyproof with $\varepsilon > 0$ if for all $i \in [k]$, other labelers' data \mathcal{D}_{-i} and deviating $\tilde{\theta}_i \neq \theta_i^*$ it holds that

$$\mathbb{E}_{o^{i,j} \sim \mathbb{P}_{\theta_i^*}} [J_i(\pi_{\text{RLHF}}(\mathcal{D}_i^*, \mathcal{D}_{-i}))] \geq \mathbb{E}_{\tilde{o}^{i,j} \sim \mathbb{P}_{\tilde{\theta}_i}} [J_i(\pi_{\text{RLHF}}(\tilde{\mathcal{D}}_i, \mathcal{D}_{-i}))] - \varepsilon.$$

A few comments are in place. The careful reader might wonder why we define strategyproofness at the distributional level (ex ante) rather than at the level of realized preference labels, e.g., by allowing labelers to flip preferences *after* sampling. The main reason is that defining misreporting at the data level would blur the line between strategic manipulation and post hoc noise correction, which is not the focus of our analysis. Defining strategies over distributions ensure a more meaningful comparison between truthful and strategic reporting and avoids these complications.

Learning Objective. We here assume that the group of labelers is representative of the population whose preferences we wish to align to. Our objective is then to compute a policy maximizing the *average social welfare*

$$\mathcal{W}(\rho, \pi) := \frac{1}{k} \sum_{i=1}^k J_i(\rho, \pi).$$

Let $\pi^* := \operatorname{argmax}_{\pi \in \Pi} \mathcal{W}(\rho, \pi)$ be the optimal policy. The *suboptimality* of a policy π is defined as

$$\operatorname{SubOpt}(\rho, \pi) := \mathcal{W}(\rho, \pi^*) - \mathcal{W}(\rho, \pi).$$

In addition to the standard notion of suboptimality, it can also be insightful to consider the multiplicative *approximation ratio* of a policy π that is frequently studied in the computational social choice literature and defined as the ratio

$$\alpha(\rho, \pi) := \frac{\mathcal{W}(\rho, \pi)}{\mathcal{W}(\rho, \pi^*)}.$$

By definition, this ratio satisfies $\alpha(\rho, \pi) \leq 1$ and the larger the ratio the better the policy. In the following, we primarily use the approximation ratio as a secondary metric to understand the convergence behavior of an RLHF method, i.e., when the number of samples is sufficiently large.

3.1 Existing RLHF is not Strategyproof

Unsurprisingly, we find that existing RLHF algorithms are not strategyproof. We consider two recently proposed RLHF methods for learning from diverse human preferences [7, 30]. Note that Maxmin-RLHF [7] consider a maximin objective, that is, they wish to maximize the worst-case utility across all labelers. While this is different from the social welfare objective that we consider, it does not prevent us from analyzing their strategyproofness.

Proposition 3.3. *Existing RLHF methods such as Pessimistic Social Welfare [30] and MaxMin-RLHF [7] are not strategyproof.*

Next, we wish to understand what consequences being manipulable has on the policy performance of the RLHF algorithm. After all, one could imagine failing to guarantee strategyproofness but still learning a nearly optimal policy. This is in general not the case and we show that the performance can degrade arbitrarily in the worst-case even if only a single labeler is strategic. We show this exemplarily for the Pessimistic Social Welfare approach from Zhong et al. [30].

Proposition 3.4. *Let at least one out of the k labelers report strategically. Let $\hat{\pi}$ denote the output of the Pessimistic Social Welfare [30]. In the worst-case, for n sufficiently large, $\mathcal{W}(\hat{\pi}) \leq \varepsilon$ while $\mathcal{W}(\pi^*) \geq BL - 2\varepsilon$ for any $\varepsilon > 0$. In other words, $\operatorname{SubOpt}(\hat{\pi}) = \mathcal{W}(\pi^*) - \varepsilon \geq BL - 3\varepsilon$. This means that the policy learned by Pessimistic Social Welfare can be almost arbitrarily bad.*

Proof Sketch. We provide a simple example for a contextual bandit where the first labeler strongly disagrees with all other labelers, but can exert significant influence on the computed policy by overstating its preference in a dimension of the features that is otherwise irrelevant to all labelers' utility (i.e., θ_i^* is zero in said dimension for all labelers). \square

3.2 Inherent Limitations of Strategyproof RLHF

We have seen that existing RLHF approaches are not strategyproof, but can be manipulated by labelers to the detriment of its policy performance. We now also show that any RLHF algorithm that satisfies strategyproofness must suffer at least constant suboptimality (irrespective of the number of samples or policy coverage) and has an approximation ratio of at most $1/k$. We thereby observe a fundamental trade-off between incentive alignment (strategyproofness) and policy alignment (social welfare maximization) in RLHF with strategic preference labeling.

Theorem 3.5. *The output $\hat{\pi}$ of any strategyproof RLHF algorithm has worst-case expected suboptimality at least $\text{SubOpt}(\hat{\pi}) \geq \frac{k-1}{k}$, where k denotes the number of labelers.*

Proof Sketch. We can map each RLHF instance to a voting instance and map $\hat{\pi}$ to a voting rule f for the latter, such that f always outputs the same alternative (or distribution of alternatives) as $\hat{\pi}$ does. This construction ensures that if $\hat{\pi}$ is strategyproof, then f is, too. The Gibbard–Satterthwaite theorem [14, 25] says that any strategyproof voting rule must be either a dictatorial rule or a “dupe”, i.e., either it always selects the most preferred alternative of a fixed voter, or selects among a fixed pair of alternatives. Hence, if $\hat{\pi}$ is strategyproof, it must behave either as a dictatorial rule, always selecting the most preferred action of a fixed labeler, or as a dupe, always selecting the outcome among a fixed pair of actions. The former case leads to low social welfare values for instances in which all the other labelers’ rewards are negatively correlated with that of the fixed labeler. The latter leads to low social welfare values for instances in which the fixed pair of actions have almost zero value to all labelers. In both cases, the suboptimality gaps are at least $(k-1)/k$. \square

Theorem 3.5 implies that even with infinitely many samples, no strategyproof RLHF algorithm converges to the optimal policy in the worst-case. This is also reflected in the following upper bound on the multiplicative approximation ratio of strategyproof algorithms.

Corollary 3.6. *The approximation ratio of any strategyproof RLHF method is $\alpha(\rho, \hat{\pi}) \leq \frac{1}{k}$.*

In other words, any strategyproof RLHF algorithm achieves k -times worse social welfare compared to the optimal policy that maximizes social welfare.

4 Approximate Strategyproofness: Pessimistic Median of MLEs

We first consider the contextual bandit problem and discuss the extension to MDPs in Section 5. Our previous Theorem 3.5 suggests that without additional assumptions about the problem instance, we cannot reconcile strategyproofness with social welfare maximization. For this reason, we here introduce an additional assumption about the structure of the initial state distribution (i.e., context distribution) and the policy space.

Assumption 2. The set $\{\mathbb{E}_{s \sim \rho} [\phi(s, \pi(s))] : \pi \in \Pi\}$ spans a hyperrectangle in \mathbb{R}^d .

Specifically, in the simplest case where $\mathbb{E}_{s \sim \rho} [\phi(s, \pi(s))] \in [-1, 1]^d$, this means that for any $\mathbf{z} \in [-1, 1]^d$ there exists $\pi \in \Pi$ such that $\|\mathbb{E}_{s \sim \rho} [\phi(s, \pi(s))] - \mathbf{z}\|_2 = 0$.

We propose to use a median rule over learned reward parameters in combination with pessimistic estimates to achieve approximate strategyproofness while maximizing social welfare. To do so, we must first introduce a few key concepts and quantities.

Algorithm 1 Pessimistic Median of MLEs (Pessimistic MoMLE)

input Offline preference data sets $\mathcal{D}_1, \dots, \mathcal{D}_k$

- 1: **for** every labeler $i \in [k]$ **do**
- 2: Compute the MLE $\hat{\theta}_i^{\text{MLE}}$
- 3: Construct confidence set $C_i := \{\theta \in \mathbb{R}^d : \|\hat{\theta}_i^{\text{MLE}} - \theta\|_{\Sigma_{\mathcal{D}_i}} \leq f(d, n, \delta)\}$
- 4: **end for**
- 5: Get median confidence set $\mathcal{C} := \{\text{med}(\theta_1, \dots, \theta_k) : \theta_i \in C_i \text{ for } i \in [k]\}$
- 6: Get pessimistic estimate of the social welfare w.r.t. \mathcal{C} given by

$$\underline{\mathcal{W}}(\pi) := \min_{\theta \in \mathcal{C}} \mathbb{E}_{s \sim \rho} [\langle \theta, \phi(s, \pi(s)) \rangle]$$

- 7: **return** $\hat{\pi} = \text{argmax}_{\pi \in \Pi} \underline{\mathcal{W}}(\pi)$
-

MLEs and Confidences. Let $\mathcal{D}_i = (s^{i,j}, a_0^{i,j}, a_1^{i,j}, o^{i,j})_{1 \leq j \leq n}$ be the preference data reported by labeler $i \in [k]$, where $o^{i,j} \sim \mathbb{P}_{\theta_i}(\cdot | s^{i,j}, a_0^{i,j}, a_1^{i,j})$ (cf. Section 3). In other words, labeler i samples their preferences w.r.t. the labeler-chosen (and potentially manipulated) reward parameter $\theta_i \in \mathbb{R}^d$. Given the observations \mathcal{D}_i , the Maximum Likelihood Estimate (MLE) of the underlying reward parameter θ_i is the maximizer of the log-likelihood

$$\hat{\theta}_i^{\text{MLE}} \in \underset{\theta}{\text{argmax}} \sum_{j=1}^n \log \mathbb{P}_{\theta}(o^{i,j} | s^{i,j}, a_0^{i,j}, a_1^{i,j}).$$

We wish to establish confidences around the MLE. To this end, let $x^{i,j} = \phi(s^{i,j}, a_0^{i,j}) - \phi(s^{i,j}, a_1^{i,j})$ and consider the covariance matrix $\Sigma_{\mathcal{D}_i} = \frac{1}{n} \sum_{j=1}^n x^{i,j} (x^{i,j})^\top$.³ The confidence ellipsoid is then given by

$$C_i := \{\theta \in \mathbb{R}^d : \|\hat{\theta}_i^{\text{MLE}} - \theta\|_{\Sigma_{\mathcal{D}_i}} \leq f(d, n, \delta)\}$$

for some function $f(d, n, \delta)$. It is well-known that when choosing $f(d, n, \delta) \approx \sqrt{\frac{d + \log(k/\delta)}{n}}$, it holds with probability at least $1 - \delta$ that $\theta_i \in C_i$ (see Appendix ?? for details).

Median Rule. A fundamental insight from social choice theory is that under certain conditions aggregating preferences according to a median rule is strategyproof [21]. However, in our case, the *high-dimensionality* of features and reward parameters, the *uncertainty* about rewards, and the *policy optimization* pose additional unique challenges that can cause a median rule to become manipulable by the labelers.

To incorporate our uncertainty about the reward parameters, we consider the *pessimistic median return* of a policy defined as the return of a policy w.r.t. the worst-case *coordinate-wise* median over confidence sets C_1, \dots, C_k . In other words, we consider the worst-case performance of policies π with respect to $\text{med}(\theta_1, \dots, \theta_k)$, where med denotes the coordinate-wise median and $\theta_i \in C_i$. We outline the Pessimistic Median of MLEs approach in Algorithm 1.

4.1 Approximate Strategyproofness

We begin the analysis by showing that the Pessimistic Median of MLEs is approximately strategyproof. In particular, the strategyproofness guarantee involves an interesting, and perhaps surpris-

³We assume that $\Sigma_{\mathcal{D}_i}$ is positive definite. Otherwise, we can always consider $\Sigma_{\mathcal{D}_i} + \lambda_i I$ and choose $\lambda_i > 0$ depending on the preference data set \mathcal{D}_i .

ing, dependence on the *uniform policy coverage* of every labeler's data \mathcal{D}_i .

Theorem 4.1. *Pessimistic Median of MLEs is $\tilde{\mathcal{O}}(\kappa_i \sqrt{d/n})$ -strategyproof, where κ_i measures the uniform policy coverage of \mathcal{D}_i . More precisely, for every labeler $i \in [k]$, any other labelers' data \mathcal{D}_{-i} and manipulated reward parameter $\tilde{\theta}_i \neq \theta_i^*$, with probability at least $1 - \delta$ the gain from misreporting is bounded as*

$$\mathbb{E}_{\tilde{\theta}^{i,j} \sim \mathbb{P}_{\tilde{\theta}_i}} [J_i(\pi_{\text{RLHF}}(\tilde{\mathcal{D}}_i, \mathcal{D}_{-i}))] - \mathbb{E}_{\theta^{i,j} \sim \mathbb{P}_{\theta_i^*}} [J_i(\pi_{\text{RLHF}}(\mathcal{D}_i^*, \mathcal{D}_{-i}))] \leq \text{const} \cdot \kappa_i \sqrt{\frac{d + \log(k/\delta)}{n}},$$

where $\kappa_i = \max_{\pi \in \Pi} \|\mathbb{E}_{s \sim \rho} [\phi(s, \pi(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}}$ is the uniform policy coverage of \mathcal{D}_i .⁴

Proof Sketch. The key challenge is that estimation errors of the reward parameters may unintentionally alter the median computation and thereby create unintended incentives for misreporting. To bound the gain from misreporting, we analyze the effect of deviations in the estimated parameters on the learned policy. Using concentration inequalities, we show that the deviation in each labeler's expected return is proportional to the estimation error, which scales as $\sqrt{d/n}$. The worst-case impact on strategyproofness is then controlled by the uniform coverage coefficient κ_i , which measures how well the labeler's data constrains policy choices. \square

Whereas the $\sqrt{d/n}$ factor may be expected due to the construction of confidence ellipsoids of corresponding size, the uniform policy coverage coefficient κ_i is more surprising. In offline RL and RLHF, performance bounds typically depend on the coverage of the optimal policy π^* only, i.e., they feature a term of the form $\|\mathbb{E}_{s \sim \rho} [\phi(s, \pi^*(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}}$, where π^* is the optimal policy.

However, in our case, we are not bounding the suboptimality of a learned policy but rather analyzing the strategic incentives of labelers. This shifts the focus to the range of possible policies that could result from different labeler behavior. Since labelers can, in principle, report arbitrarily misleading reward parameters—potentially inducing policies far from π^* —bounding their incentive to deviate requires uniform policy coverage rather than coverage of any single specific policy. This ensures that no matter what policy is induced by a misreport, the confidence set remains well-constrained and bounds the potential gain from misreporting.

4.2 Social Welfare Maximization

We have shown that being truthful is approximately optimal for all labelers. Next, we provide guarantees on the suboptimality and the approximation ratio of the Pessimistic Median of MLEs algorithm when the labelers are either truthful or act according to their (potentially manipulating) weakly dominant strategy, which we show to exist. We begin with the case when the labelers are truthful, which is a $\tilde{\mathcal{O}}(\kappa_i \sqrt{d/n})$ -dominant strategy according to Theorem 4.1.

Theorem 4.2. *Let $\hat{\pi}$ be the output of the Pessimistic Median of MLEs algorithm and suppose that all labelers report truthfully. With probability at least $1 - \delta$:*

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot \sqrt{\frac{d \log(k/\delta)}{k}} + \text{const} \cdot k \sqrt{\frac{d + \log(k/\delta)}{n}} \max_{i \in [k]} \|\mathbb{E}_{s \sim \rho} [\phi(s, \pi^*(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}}. \quad (2)$$

Note that $\|\mathbb{E}_{s \sim \rho} [\phi(s, \pi^*(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}} = \|\Sigma_{\mathcal{D}_i}^{-1/2} \mathbb{E}_{s \sim \rho} [\phi(s, \pi^*(s))]\|_2$.

⁴Note that for any positive definite matrix Σ and vector x , we can write $\|x\|_{\Sigma^{-1}} = \|\Sigma^{-1/2}x\|_2$. It is also worth noting that labeler i cannot influence the coverage coefficient κ_i as it only depends on the state-action pairs and not the preference labels.

Proof Sketch. The suboptimality arises from two sources: (1) the deviation of the pessimistic median from the average, and (2) the deviation of each true reward parameter from its worst-case estimate in its respective confidence set. The first term follows from median concentration around the mean, contributing an error of $\mathcal{O}(\sqrt{d \log(k/\delta)/k})$. The second term is upper bounded by $\mathcal{O}(\sqrt{(d + \log(k/\delta))/n})$, scaled by the worst-case policy coverage coefficient. Here, taking the median over confidence sets introduces an additional factor of k . \square

We also show that the Pessimistic Median of MLEs algorithm enjoys a suboptimality upper bound matching the one from Theorem 4.2 under any weakly dominant strategy it induces.

Proposition 4.3. *When the labelers report their preferences according to any weakly dominant strategy under Pessimistic Median of MLEs, with probability at least $1 - \delta$, the output $\hat{\pi}$ satisfies:*

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot \sqrt{\frac{d \log(k/\delta)}{k}} + \text{const} \cdot k \sqrt{\frac{d + \log(k/\delta)}{n}} \max_{i \in [k]} \|\mathbb{E}_{s \sim \rho}[\phi(s, \pi^*(s))]\|_{\Sigma_{D_i}^{-1}}.$$

The bounds in Theorem 4.2 and Proposition 4.3 suggest two sources of suboptimality. The first term, of order $\mathcal{O}(\sqrt{d \log(k/\delta)/k})$, stems from approximating the social welfare function using the coordinate-wise median, which improves as the number of labelers increases. The second term results from the estimation of the underlying reward parameters, where the use of a median rule introduces an additional factor of k . Overall, as the number of samples per labeler increases and provides sufficient coverage, and as the number of labelers grows, the Pessimistic Median of MLEs algorithm converges to the optimal policy.

Remark 4.4 (Suboptimality Lower Bound). *The worst-case suboptimality of any RLHF algorithm in our problem setup is lower bounded by $\Omega(\sqrt{d/n})$. This can be derived using a similar worst-case problem instance construction to the one in Zhu et al. [31].*

We want to highlight the performance bounds of the Pessimistic Median of MLEs algorithm in two interesting special cases: (1) when there is only a single labeler so that $k = 1$, and (2) when all k labelers have identical reward functions.

Corollary 4.5. *When there is only a single labeler, with probability at least $1 - \delta$:*

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot \sqrt{\frac{d + \log(k/\delta)}{n}} \|\mathbb{E}_{s \sim \rho}[\phi(s, \pi^*(s))]\|_{\Sigma_{D_1}^{-1}}.$$

When all k labelers have the same reward function, with probability at least $1 - \delta$:

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot k \sqrt{\frac{d + \log(k/\delta)}{n}} \max_{i \in [k]} \|\mathbb{E}_{s \sim \rho}[\phi(s, \pi^*(s))]\|_{\Sigma_{D_i}^{-1}}.$$

The result for the single labeler matches the existing bounds in the offline RLHF literature. Interestingly, we observe that in the special case of k labelers with identical reward functions, the Pessimistic Median of MLEs avoids the additive $\mathcal{O}(\sqrt{d \log(k/\delta)/k})$ term but still suffers from an additional factor of k as it takes the median over the confidence sets.

Finally, we can also derive a lower bound on the approximation ratio of Algorithm 1.

Corollary 4.6. *Suppose $\mathcal{W}(\pi^*) > 0$ is constant. When the number of samples is sufficiently large and provide sufficient coverage of the optimal policy, with probability at least $1 - \delta$, the approximation ratio of the Pessimistic Median of MLEs algorithm is given by $\alpha(\rho, \hat{\pi}) \geq 1 - \mathcal{O}(\sqrt{d \log(k/\delta)/k})$.*

5 Extension to Markov Decision Processes

We now extend our algorithm and our previous results to MDPs. Recall that we consider trajectory-wise preferences so that labeler i provides a preferences $o^{i,j}$ over two trajectories $\tau_0^{i,j}$ and $\tau_1^{i,j}$ given initial state $s^{i,j}$ according to a BTL model \mathbb{P}_{θ_i} as defined in Section 3. This results in reported preference data $\mathcal{D}_i = (s^{i,j}, \tau_0^{i,j}, \tau_1^{i,j}, o^{i,j})_{1 \leq j \leq n}$.

Like before, the MLE of θ_i is given by the maximizer of the log-likelihood

$$\theta_i^{\text{MLE}} := \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^n \log \mathbb{P}_{\theta}(o^{i,j} \mid s^{i,j}, \tau_0^{i,j}, \tau_1^{i,j}),$$

where $\mathbb{P}_{\theta}(\cdot \mid s, \tau_0, \tau_1)$ is defined in (1). To construct the confidence ellipsoid around the MLE, let $x^{i,j} = \sum_{h=1}^H (\phi(s_h^{i,j}, a_h^{i,j}) - \phi(\bar{s}_h^{i,j}, \bar{a}_h^{i,j}))$ with $s_1^{i,j} = \bar{s}_1^{i,j} = s^{i,j}$ and consider the adapted covariance matrix $\Sigma_{\mathcal{D}_i} = \sum_{j=1}^n x^{i,j} (x^{i,j})^\top$. Note that this definition is consistent with our previous definition for the contextual bandit setup when $H = 1$.

To derive the pessimistic estimate of the median social welfare, we now consider the state occupancy of a policy π defined as $q_{\pi}(s \mid \rho) := \frac{1}{H} \sum_{h=1}^H \mathcal{P}_h(s_h = s \mid \rho, \pi)$, which defines a distribution over states. We can then express the expected return of policy π w.r.t. reward parameter θ as $\mathbb{E}_{s \sim \rho}[V_{\theta}^{\pi}(s)] = \mathbb{E}_{s \sim q_{\pi}}[\langle \theta, \phi(s, \pi(s)) \rangle]$ and the pessimistic estimate of the median social welfare is given by $\mathcal{W}(\pi) := \min_{\theta \in \mathcal{C}} \mathbb{E}_{s \sim q_{\pi}}[\langle \theta, \phi(s, \pi(s)) \rangle]$. The remainder of the Pessimistic Median of MLEs algorithm proceeds the same.

We assume the analogue of Assumption 3 for MDPs.

Assumption 3. The set $\{\mathbb{E}_{s \sim q_{\pi}}[\phi(s, \pi(s))]: \pi \in \Pi\}$ spans a hyperrectangle in \mathbb{R}^d .

Under Assumption 3, we obtain the following analogue of Theorem 4.1, showing that the Pessimistic Median of MLEs algorithm is approximately strategyproof.

Theorem 5.1. *Pessimistic Median of MLEs is $\mathcal{O}(\nu_i \sqrt{d/n})$ -strategyproof with uniform policy coverage coefficient $\nu_i = \max_{\pi \in \Pi} \|\mathbb{E}_{s \sim q_{\pi}}[\phi(s, \pi(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}}$. More precisely, for every labeler $i \in [k]$, any other labelers' data \mathcal{D}_{-i} and manipulated reward parameter $\tilde{\theta}_i \neq \theta_i^*$, with probability at least $1 - \delta$, the gain from misreporting is bounded as*

$$\mathbb{E}_{\tilde{o}^{i,j} \sim \mathbb{P}_{\tilde{\theta}_i}} [J_i(\pi_{\text{RLHF}}(\tilde{\mathcal{D}}_i, \mathcal{D}_{-i}))] - \mathbb{E}_{o^{i,j} \sim \mathbb{P}_{\theta_i^*}} [J_i(\pi_{\text{RLHF}}(\mathcal{D}_i^*, \mathcal{D}_{-i}))] \leq \text{const} \cdot \nu_i \sqrt{\frac{d + \log(k/\delta)}{n}}.$$

The suboptimality upper bounds under truthful or weakly dominant reporting also take a similar form to their counterparts in Theorem 4.2 and Proposition 4.3. Similarly to before, the coverage of the optimal policy, i.e., social welfare maximizing policy, is enough.

Theorem 5.2. *When all labelers report truthfully or report according to their weakly dominant strategies, then with probability at least $1 - \delta$:*

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot \sqrt{\frac{d \log(k/\delta)}{k}} + \text{const} \cdot k \sqrt{\frac{d + \log(k/\delta)}{n}} \max_{i \in [k]} \|\mathbb{E}_{s \sim q_{\pi^*}}[\phi(s, \pi^*(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}}.$$

Note that we can also extend the corollaries from Section 4 to MDPs in a similar fashion.

6 Discussion

We studied how to robustify offline RLHF against strategic preference reporting in a pluralistic alignment setting with multiple diverse labelers. We demonstrated an inherent trade-off between incentive alignment and policy alignment and proposed the Pessimistic Median of MLEs algorithm based on pessimistic estimates of the median return of a policy. We show that this approach is $\tilde{\mathcal{O}}(\sqrt{d/n})$ -strategyproof while guaranteeing suboptimality of at most $\tilde{\mathcal{O}}(\sqrt{d/k} + k\sqrt{d/n})$. In future work, it will be interesting to study strategyproof RLHF for non-linear reward functions, parameterized or otherwise restricted policy classes, as well as different preference models to the here used BTL preference model.

Acknowledgements

This work was supported by the EPSRC Prosperity Partnership FAIR (grant number EP/V056883/1). MK receives funding from the ERC under the European Union’s Horizon 2020 research and innovation programme (FUN2MODEL, grant agreement No. 834115).

References

- [1] Parand A Alamdari, Soroush Ebadian, and Ariel D Procaccia. Policy aggregation. *Advances in Neural Information Processing Systems*, 37:68308–68329, 2025.
- [2] Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Vallian, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, pages 1–9, 2025.
- [3] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.
- [4] Alexander Bukharin, Ilgee Hong, Haoming Jiang, Zichong Li, Qingru Zhang, Zixuan Zhang, and Tuo Zhao. Robust reinforcement learning from corrupted human feedback. *arXiv preprint arXiv:2406.15568*, 2024.
- [5] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémie Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.
- [6] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*, 2024.
- [7] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. In *Forty-first International Conference on Machine Learning*, 2024.
- [8] Jie Cheng, Gang Xiong, Xingyuan Dai, Qinghai Miao, Yisheng Lv, and Fei-Yue Wang. Rime: Robust preference-based reinforcement learning with noisy preferences. *arXiv preprint arXiv:2402.17257*, 2024.

- [9] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- [10] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.
- [11] Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting llm quantization. *Advances in Neural Information Processing Systems*, 37:41709–41732, 2025.
- [12] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951*, 2024.
- [13] Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for ai alignment from human feedback. *arXiv preprint arXiv:2405.14758*, 2024.
- [14] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.
- [15] Allan Gibbard. Straightforwardness of game forms with lotteries as outcomes. *Econometrica: Journal of the Econometric Society*, pages 595–614, 1978.
- [16] Shugang Hao and Lingjie Duan. Online learning from strategic human feedback in llm fine-tuning. *arXiv preprint arXiv:2412.16834*, 2024.
- [17] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt’s pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.
- [18] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.
- [19] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.
- [20] Debmalya Mandal, Andi Nika, Parameswaran Kamalaruban, Adish Singla, and Goran Radanović. Corruption robust offline reinforcement learning with human feedback. *arXiv preprint arXiv:2402.06734*, 2024.
- [21] Hervé Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35(4):437–455, 1980.
- [22] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E Ozdaglar. Rlhf from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.
- [23] Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. *arXiv preprint arXiv:2405.20304*, 2024.

- [24] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.
- [25] Mark Allen Satterthwaite. Strategy-proofness and arrow’s conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2):187–217, 1975.
- [26] Anand Siththanjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.
- [27] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.
- [28] Ermis Soumalias, Michael J Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising. *arXiv preprint arXiv:2405.05905*, 2024.
- [29] Haoran Sun, Yurong Chen, Siwei Wang, Wei Chen, and Xiaotie Deng. Mechanism design for llm fine-tuning with multiple reward models. *arXiv preprint arXiv:2405.16276*, 2024.
- [30] Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*, 2024.
- [31] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

A Proofs

A.1 Proof of Proposition 3.3

Proposition 3.3. *Existing RLHF methods such as Pessimistic Social Welfare [30] and MaxMin-RLHF [7] are not strategyproof.*

Proof. We construct straightforward examples for each of the algorithms and show that the algorithms are not strategyproof. W.l.o.g. we assume that n is sufficiently large with appropriate policy coverage so that the algorithms are able to obtain perfect estimates. We thereby also avoid redefining the preference model \mathbb{P}_θ to fit the models considered in the respective work. W.l.o.g. we also ignore the KL-regularization w.r.t. the reference policy π_{ref} in MaxMin-RLHF, which would only add notational burden. In the following examples, the horizon is set to $H = 1$, that is, we consider contextual bandit problems.

Pessimistic Social Welfare: Suppose there are two labelers with true reward parameters $\theta_1^* = (1, 0)$ and $\theta_2^* = (0, 1)$. Moreover, suppose that for all s , we have $\phi(s, a) = (1/2, 1/2)$ and $\phi(s, b) = (3/4, 0)$. If both labelers report truthfully, Pessimistic Social Welfare reports a policy $\pi(s) = a$ for all s as action a is maximizing social welfare. In this case, labeler 1 receives utility 1/2. Suppose that labeler 1 misreports as $\tilde{\theta}_1 = (1, -1)$. As a result, the social welfare maximizing policy is $\pi(s) = b$, which yields utility 3/4 for labeler 1. Hence, misreporting is beneficial to labeler 1 and Pessimistic Social Welfare not strategyproof.

MaxMin-RLHF: Consider the simple example where $\theta_1^* = (1, 0)$ and $\theta_2^* = (1/2, 1/2)$ as well as $\phi(s, a) = (1/2, 1/2)$ and $\phi(s, b) = (3/4, 0)$ for all s . If both labelers report truthfully, the MaxMin-RLHF would compute the policy $\pi(s) = a$ which yields a return of 1/2 for both labelers. However, suppose labeler 1 reports $\tilde{\theta}_1 = (1, -1)$ while labeler 2 truthfully reports $\theta_2^* = (1/2, 1/2)$. In this case, MaxMin-RLHF returns a policy $\tilde{\pi}(s) = b$ as this maximizes the minimal utility w.r.t. the reported reward parameters. The return for labeler 1 under policy $\tilde{\pi}$ is 3/4, which means that misreporting is to the benefit of labeler 1 and MaxMin-RLHF not strategyproof.

□

A.2 Proof of Proposition 3.4

Proposition 3.4. *Let at least one out of the k labelers report strategically. Let $\hat{\pi}$ denote the output of the Pessimistic Social Welfare [30]. In the worst-case, for n sufficiently large, $\mathcal{W}(\hat{\pi}) \leq \varepsilon$ while $\mathcal{W}(\pi^*) \geq BL - 2\varepsilon$ for any $\varepsilon > 0$. In other words, $\text{SubOpt}(\hat{\pi}) = \mathcal{W}(\pi^*) - \varepsilon \geq BL - 3\varepsilon$. This means that the policy learned by Pessimistic Social Welfare can be almost arbitrarily bad.*

Proof. We construct a contextual bandit problem, where Pessimistic Social Welfare is arbitrarily bad if even a single labeler is strategic. We here assume that Pessimistic Social Welfare receives infinitely many samples from each labeler with full policy coverage.

Let $\theta_1^* = (0, 1, 0)$ and $\theta_i^* = (\frac{B}{k-1}, 0, 0)$ for all $i \neq 1$. Moreover, suppose that $\phi(s, a) = (\sqrt{L^2 - 2\varepsilon^2}, 0, 0)$ and $\phi(s, b) = (0, \varepsilon, \sqrt{L^2 - \varepsilon^2})$ for all s . Under truthful reporting, Pessimistic Social Welfare computes the policy $\pi(s) = a$, which clearly maximizes social welfare. In particular, the optimal social welfare is given by $\mathcal{W}(\pi^*) = B\sqrt{L^2 - 2\varepsilon^2} \geq B(L - \sqrt{2}\varepsilon)$. In this case, labeler 1 receives utility zero. However, suppose that labeler 1 misreports its reward parameter as $\tilde{\theta}_1^* = (0, 0, B)$. In this case, Pessimistic Social Welfare returns the policy $\tilde{\pi}(s) = b$, which has social welfare $\mathcal{W}(\tilde{\pi}) = \varepsilon$, whereas the utility of labeler 1 is ε which is the best possible outcome for labeler 1.

As a result, even if only labeler 1 misreports, then $\mathcal{W}(\hat{\pi}) = \varepsilon$ and the suboptimality is at least $\text{SubOpt}(\hat{\pi}) = \mathcal{W}(\pi^*) - \varepsilon \geq BL - 3\varepsilon$. We can choose $\varepsilon > 0$ arbitrarily small (in particular, note that

ε does not depend on the reward parameter so that any estimation error would have no effect). This means that the suboptimality of Pessimistic Social Welfare can be maximal. \square

A.3 Proof of Theorem 3.5

Theorem 3.5. *The output $\hat{\pi}$ of any strategyproof RLHF algorithm has worst-case expected suboptimality at least $\text{SubOpt}(\hat{\pi}) \geq \frac{k-1}{k}$, where k denotes the number of labelers.*

Proof. Suppose for the sake of contradiction that the (worst-case) suboptimality gap of $\hat{\pi}$ is $\frac{k-1}{k} - \epsilon$ for some $\epsilon > 0$. We show that if this were true, we could construct a voting rule based on $\hat{\pi}$ that is *strategyproof*, *non-dictatorial*, and *onto at least three alternatives*, contradicting the Gibbard–Satterthwaite theorem [14, 25]. The Gibbard–Satterthwaite theorem asserts that such a voting rule does not exist. We will consider the case where $\hat{\pi}$ is deterministic and discuss how the proof generalizes to the case with randomized $\hat{\pi}$, too.

Specifically, consider a voting instance with k voters and m alternatives a_1, \dots, a_m . A voting rule f maps every possible preference profile of the voters to one of the alternatives.

To construct a voting rule based on $\hat{\pi}$, we first map every voting instance $I_V = \prec := (\prec_1, \dots, \prec_k)$ to an RLHF instance I_{RLHF} as follows.

- Let there be one state (so we omit the state in what follows) and m actions a_1, \dots, a_m , each corresponding to an alternative in I_V .
- The feature embedding ϕ maps each action a_ℓ to the unit vector whose ℓ -th component is 1.
- Let there be k labelers, each corresponding to a voter in I_V . Each labeler i 's parameter θ_i^* is a vector in which the ℓ -th component is defined as follows:
 - 1 if a_ℓ is the most preferred alternative according to voter i 's preference order \prec_i in I_V .
 - $1 - \delta$ (for a sufficiently small δ) if a_ℓ is the second most preferred alternative according to \prec_i .
 - $\delta \cdot (m - j)$ if a_ℓ is the j -th most preferred alternative, $j > 2$, according to \prec_i .

The parameters ensure that each labeler i 's preference over the actions is the same as \prec_i .

With the above map from I_V to I_{RLHF} , we then let f be a voting rule that outputs alternative a_ℓ if $\hat{\pi}$ outputs action a_ℓ in I_{RLHF} . Clearly, f must be strategyproof given our assumption that $\hat{\pi}$ is strategyproof and the fact that I_{RLHF} preserves the preference orders in I_V . We next argue that, given the assumption that $\hat{\pi}$ has a suboptimality gap of at most $\frac{k-1}{k} - \epsilon$, f must be non-dictatorial and onto at least three alternatives.

f is Non-Dictatorial. Suppose that f is dictatorial; say, it always outputs the most preferred alternative of voter 1. This means that $\hat{\pi}$ must output a_1 on the RLHF instance that the following $I_V = \prec$ instance maps to:

$$\begin{aligned} a_1 &\prec_1 a_2 \prec_1 \cdots \prec_1 a_{m-1} \prec_1 a_m \\ a_2 &\prec_i a_3 \prec_i \cdots \prec_i a_m \prec_i a_1 \quad \text{for all } i = 2, \dots, k. \end{aligned}$$

However, this contradicts the assumption that the suboptimality gap of $\hat{\pi}$ is at most $\frac{k-1}{k} - \epsilon$. To see this, note that a_1 achieves social welfare 1 in I_{RLHF} , while a_2 achieves social welfare $k(1 - \delta)$. The suboptimality is then at least $\frac{k-1-\delta k}{k-\delta k} > \frac{k-1}{k} - \epsilon$ when $\delta \rightarrow 0$.

f is Onto. Similarly, we can argue that f must be onto at least three alternatives by considering the set of all possible voting instances where all voters' preferences are identical. In this case, $\hat{\pi}$ must always output either the most or the second preferred actions of the labelers in the corresponding RLHF instances; otherwise, the suboptimality gap of $\hat{\pi}$ can be arbitrarily close to 1 when $\delta \rightarrow 0$. Consequently, when all possible preference orders are considered, $\hat{\pi}$, and hence f , must be onto at least three different alternatives (suppose that $k \geq 4$).

As a result, we obtain a voting rule that is strategyproof, non-dictatorial, and onto at least three alternatives. This contradicts the Gibbard–Satterthwaite theorem. The stated lower bound on the suboptimality gap then follows for deterministic policies.

Randomized Policies. To further argue that the same bound holds even when $\hat{\pi}$ is randomized, we invoke a generalization of the Gibbard–Satterthwaite theorem to randomized voting rules [15], which states that a voting rule, if strategyproof, must be a probability mixture of dictatorial rules and “duples”. A voting rule is a duple if it restricts its outcomes, over all possible instances, to a fixed pair of alternatives.

Similarly to our approach above, we construct a voting rule f based on $\hat{\pi}$ the same way we did above and argue that, if $\hat{\pi}$ is strategyproof and has a suboptimality gap of $\frac{k-1}{k} - \epsilon$ for some $\epsilon > 0$, then f cannot be a probability mixture of dictatorial rules and duples.

Suppose for the sake of contradiction that f is a mixture of dictatorial rules and duples. Consider the following set of voting instances $(\prec^j)_j$, each involving a set of alternatives $\{a_1, \dots, a_k, b_1, \dots, b_K\}$, $K \geq 4/\epsilon$ (hence, $m = k + K$). In each instance $\prec^j = (\prec_1^j, \dots, \prec_k^j)$, the preference order \prec_i^j ranks a_i first, b_j second, and all other alternatives according to the order $a_1, \dots, a_k, b_1, \dots, b_k$.

By assumption, f is a mixture of dictatorial rules and duples. Now that there are more than K alternatives while each duple selects at most two alternatives, there must be at least one alternative among b_1, \dots, b_K that is selected by the duples with probability at most $2/K$. W.l.o.g., let this alternative be b_1 and consider the instance \prec^1 , in which each voter i ranks a_i the first and b_1 the second.

Clearly, in this instance, the policy that outputs b_1 deterministically achieves social welfare $k(1 - \delta)$. Moreover, any other alternatives yields a social welfare of at most $1 + \delta m k$, as each of these alternatives is ranked first by at most one labeler and third or even lower by all other labelers. Since f selects b_1 with probability at most $2/K$, so does $\hat{\pi}$. This means that the social welfare achieved by $\hat{\pi}$ is at most

$$\frac{2}{K} \cdot k(1 - \delta) + \left(1 - \frac{2}{K}\right)(1 + \delta m k) < \epsilon k / 2 + 1 + \delta m k.$$

This gives a suboptimality gap of $1 - \frac{\epsilon k / 2 + 1 + \delta m k}{k(1 - \delta)} > \frac{k-1}{k} - \epsilon$ when $\delta \rightarrow 0$. □

A.4 Proof of Corollary 3.6

Corollary 3.6. *The approximation ratio of any strategyproof RLHF method is $\alpha(\rho, \hat{\pi}) \leq \frac{1}{k}$.*

Proof. This is a direct consequence of Theorem 3.5. □

A.5 Proof of Theorem 4.1

Before we begin with some preliminaries that we will repeatedly use in the proofs of both Theorem 4.1 and Theorem 4.2. Firstly, we recall a standard MLE concentration bound, which can be found, for instance, in [31].

Lemma A.1 (MLE Concentration Bound). *In the contextual bandit problem, with probability at least $1 - \delta$,*

$$\|\hat{\theta}_i^{\text{MLE}} - \theta_i^*\|_{\Sigma_{\mathcal{D}_i}} \leq \text{const} \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}},$$

where $\gamma := 1/(2 + \exp(-LB) + \exp(LB))$. Note that we here assume that the covariance matrix \mathcal{D}_i is positive definite. Otherwise, consider $\Sigma_{\mathcal{D}_i} + \lambda I$, which adds an additive term λB^2 in the square root.

Proof. See, e.g., [31]. \square

Lemma A.2 (Median Concentration Bound). *Suppose that $\theta_1, \dots, \theta_k \in \mathbb{R}^d$ are sampled i.i.d. from some σ -sub Gaussian distribution. Let $\hat{\theta}_{\text{med}}$ be the coordinate-wise median and $\hat{\theta}_{\text{avg}}$ the average of $\theta_1, \dots, \theta_k$. Then, for a universal constant $c > 0$, it holds that, for every $t > 0$,*

$$\mathbb{P}\left(\|\hat{\theta}_{\text{med}} - \hat{\theta}_{\text{avg}}\|_2 \geq t\right) \leq 2 \exp\left(-\frac{ckt^2}{d\sigma^2}\right).$$

Hence, in other words, with probability at least $1 - \delta$:

$$\|\hat{\theta}_{\text{med}} - \hat{\theta}_{\text{avg}}\|_2 \leq \mathcal{O}\left(\sigma \sqrt{\frac{d \log(1/\delta)}{k}}\right).$$

Proof. We begin by proving the median concentration in one dimension. To this end, let θ_{avg}^* denote the mean of the distribution. Since each θ_i is σ -sub-Gaussian with mean θ_{avg}^* , the centered variables $X_i = \theta_i - \theta_{\text{avg}}^*$ satisfy

$$\mathbb{P}(|X_i| \geq u) \leq 2 \exp\left(-\frac{c_2 u^2}{\sigma^2}\right)$$

for some constant $c_2 > 0$. To control $\mathbb{P}(\hat{\theta}_{\text{med}} \geq \theta_{\text{avg}}^* + t)$, note that if $\hat{\theta}_{\text{med}} \geq \theta_{\text{avg}}^* + t$, at least half of the θ_i are at least $\theta_{\text{avg}}^* + t$. Define

$$p = \mathbb{P}(\theta_i \geq \theta^* + t) = \mathbb{P}(X_i \geq t).$$

By sub-Gaussianity, $p \leq \exp\left(-\frac{c_2 t^2}{\sigma^2}\right)$. Let $Y = \sum_{i=1}^k \mathbf{1}\{\theta_i \geq \theta^* + t\}$, which follows a Binomial(k, p) distribution. Then $\mathbb{P}(\hat{\theta}_{\text{med}} \geq \theta^* + t) \leq \mathbb{P}(Y \geq k/2)$. A Chernoff or Hoeffding bound implies

$$\mathbb{P}(Y \geq k/2) \leq \exp\left(-k D\left(\frac{1}{2} \| p\right)\right),$$

where $D(\frac{1}{2} \| p)$ is the Kullback–Leibler divergence between Bernoulli(1/2) and Bernoulli(p). For $p \ll 1/2$, $D(\frac{1}{2} \| p)$ is bounded below by a constant times $(1/2 - p)^2$. Hence, there exists $c_1 > 0$ such that

$$\mathbb{P}(\hat{\theta}_{\text{med}} \geq \theta_{\text{avg}}^* + t) \leq \exp\left(-\frac{c_1 k t^2}{\sigma^2}\right).$$

By a symmetric argument, $\mathbb{P}(\hat{\theta}_{\text{med}} \leq \theta_{\text{avg}}^* - t) \leq \exp\left(-\frac{c_1 k t^2}{\sigma^2}\right)$. Combining these, we obtain

$$\mathbb{P}(|\hat{\theta}_{\text{med}} - \theta_{\text{avg}}^*| \geq t) \leq \mathbb{P}(\hat{\theta}_{\text{med}} \geq \theta_{\text{avg}}^* + t) + \mathbb{P}(\hat{\theta}_{\text{med}} \leq \theta_{\text{avg}}^* - t) \leq 2 \exp\left(-\frac{c_1 k t^2}{\sigma^2}\right).$$

From Hoeffding's inequality we get an analogous bound for $\mathbb{P}(|\hat{\theta}_{\text{avg}} - \theta_{\text{avg}}^*| \geq t)$ so that we get the desired result for $d = 1$ using the triangle inequality $|\hat{\theta}_{\text{med}} - \hat{\theta}_{\text{avg}}| \leq |\hat{\theta}_{\text{med}} - \theta_{\text{avg}}^*| + |\theta_{\text{avg}}^* - \hat{\theta}_{\text{avg}}|$.

Finally, this translates to a bound in $d > 1$ dimensions by using Jensen's inequality

$$\|\hat{\theta}_{\text{med}} - \hat{\theta}_{\text{avg}}\|_2 = \sqrt{\sum_{j=1}^d (\hat{\theta}_{\text{med},j} - \hat{\theta}_{\text{avg},j})^2} \leq \sqrt{d} \max_{j \in [d]} |\hat{\theta}_{\text{med},j} - \hat{\theta}_{\text{avg},j}|$$

and applying the previous bound for each dimension. \square

We are now ready to prove Theorem 4.1

Theorem 4.1. *Pessimistic Median of MLEs is $\tilde{\mathcal{O}}(\kappa_i \sqrt{d/n})$ -strategyproof, where κ_i measures the uniform policy coverage of \mathcal{D}_i . More precisely, for every labeler $i \in [k]$, any other labelers' data \mathcal{D}_{-i} and manipulated reward parameter $\tilde{\theta}_i \neq \theta_i^*$, with probability at least $1 - \delta$ the gain from misreporting is bounded as*

$$\mathbb{E}_{\tilde{o}^{i,j} \sim \mathbb{P}_{\tilde{\theta}_i}} [J_i(\pi_{\text{RLHF}}(\tilde{\mathcal{D}}_i, \mathcal{D}_{-i}))] - \mathbb{E}_{o^{i,j} \sim \mathbb{P}_{\theta_i^*}} [J_i(\pi_{\text{RLHF}}(\mathcal{D}_i^*, \mathcal{D}_{-i}))] \leq \text{const} \cdot \kappa_i \sqrt{\frac{d + \log(k/\delta)}{n}},$$

where $\kappa_i = \max_{\pi \in \Pi} \|\mathbb{E}_{s \sim \rho} [\phi(s, \pi(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}}$ is the uniform policy coverage of \mathcal{D}_i .⁵

Proof. We begin first with the case where every individual can directly report their reward parameter to the algorithm, hence, removing the noise and uncertainty from the process. In this case, we show that Pessimistic Median of MLEs is exactly strategyproof.

Case 1 (direct access to $\theta_1, \dots, \theta_k$): Let us begin with the case where we obtain infinitely many samples with appropriate coverage so that $C_i = \{\theta_i\}$ for all individuals $i \in [k]$. We need to show that reporting θ_i^* is the optimal strategy for individual i irrespective of the other individuals' strategies.

The following two basic lemmas will prove useful.

Lemma A.3. *Let $\theta_{-i} \in \mathbb{R}^{(k-1) \times d}$ be fixed arbitrarily. For any $j \in [d]$ the following holds:*

- If $\theta_{i,j}^* > 0$ and $\text{med}(\theta_{-i,j}, \theta_{i,j}^*) < 0$, then $\text{med}(\theta_{-i,j}, \theta_{i,j}) < 0$ for all $\theta_i \in \mathbb{R}^d$.
- Analogously, if $\theta_{i,j}^* < 0$ and $\text{med}(\theta_{-i,j}, \theta_{i,j}^*) > 0$, then $\text{med}(\theta_{-i,j}, \theta_{i,j}) > 0$ for all $\theta_i \in \mathbb{R}^d$.

Proof. W.l.o.g. let $\theta_{i,j}^* > 0$ and let $\theta_i \in \mathbb{R}^d$. Suppose that $\theta_{i,j} < 0$. It follows directly that $\text{med}(\theta_{-i,j}, \theta_{i,j}) < \text{med}(\theta_{-i,j}, \theta_{i,j}^*)$. Alternatively, suppose that $\theta_{i,j} > 0$. Since $\text{med}(\theta_{-i,j}, \theta_{i,j}^*) < 0$, it means that the median equals some $\theta_{l,j} < 0$ with $l \neq i$. Hence, the median does not change for any alternative choice $\theta_{i,j} > 0$. \square

We assume hyperrectangularity, which allows use to decompose the reward-maximizing policy as follows. For a given policy π , let $\mathbf{z}_\pi := \mathbb{E}_{s \sim \rho} [\phi(s, \pi(s))] \in \mathbb{R}^d$ denote its feature occupancy and let $\mathbf{z}_{\pi,j}$ be its j -th entry. W.l.o.g. we here assume $\mathbf{z}_{\pi,j} \in [-1, 1]$, but any other lower and upper bounds can be considered the same way.

⁵Note that for any positive definite matrix Σ and vector x , we can write $\|x\|_{\Sigma^{-1}} = \|\Sigma^{-1/2}x\|_2$. It is also worth noting that labeler i cannot influence the coverage coefficient κ_i as it only depends on the state-action pairs and not the preference labels.

We denote the optimal policy w.r.t. a reward parameter θ as $\pi^*(\theta) := \operatorname{argmax}_{\pi \in \Pi} J_\theta(\pi)$. From Assumption 2 it follows that the optimal policy $\pi^*(\theta)$ is such that $z_{\pi^*(\theta),j} = -1$ for $\theta_j < 0$ and $z_{\pi^*(\theta),j} = +1$ for $\theta_j > 0$. This yields an equivalence between reward parameters that have identical signs. In particular, this provides us with a class of reward parameters that induce an optimal policy w.r.t. the true reward parameter θ_i^* .

Lemma A.4. *Let $\theta \in \mathbb{R}^d$. If $\operatorname{sign}(\theta_{i,j}^*) = \operatorname{sign}(\theta_j)$, then $\theta_{i,j}^* \cdot z_{\pi^*(\theta),j} \geq \theta_{i,j}^* \cdot z_{\pi^*(\tilde{\theta}),j}$ for all $\tilde{\theta} \in \mathbb{R}^d$.*

Proof. This follows from the structure of the optimal policies $\pi^*(\theta)$ under Assumption 2. \square

We fix everyone's reported parameter θ_{-i} except for individual i . Moreover, let $\tilde{\theta}_i \neq \theta_i^*$ and let $\tilde{\mu} := \operatorname{med}(\theta_{-i}, \tilde{\theta}_i)$ be the coordinate-wise median w.r.t. $\tilde{\theta}_i$. Similarly, let $\mu^* = \operatorname{med}(\theta_{-i}, \theta_i^*)$ be the coordinate-wise median w.r.t. θ^* . We will now show that $J_{\theta_i^*}(\hat{\pi}(\mu^*)) \geq J_{\theta_i^*}(\hat{\pi}(\tilde{\mu}))$, i.e., reporting θ_i^* is the optimal strategy for individual i under the Pessimistic Median of MLEs algorithm.

Since we here assume direct access to the reported parameters, given reported parameters $\theta_1, \dots, \theta_k$, Pessimistic Median of MLEs computes the optimal policy w.r.t. the median $\mu = \operatorname{med}(\theta_1, \dots, \theta_k)$, i.e., $\hat{\pi} = \pi^*(\mu) = \operatorname{argmax}_{\pi \in \Pi} J_\mu(\pi)$. We here assume that the μ -maximizing policy is unique and otherwise use lexicographic tie-breaking. Clearly, if $\operatorname{signs}(\mu^*) = \operatorname{signs}(\tilde{\mu})$, then the policies $\hat{\pi}(\mu^*)$ and $\hat{\pi}(\tilde{\mu})$ are identical.

Next, consider any $j \in [d]$ so that $\operatorname{sign}(\mu_j^*) \neq \operatorname{sign}(\tilde{\mu}_j)$. Suppose that $\operatorname{sign}(\mu_j^*) = \operatorname{sign}(\theta_{i,j}^*)$. In this case, Lemma A.4 tells us that $\theta_{i,j}^* \cdot z_{\hat{\pi}(\mu^*),j} \geq \theta_{i,j}^* \cdot z_{\hat{\pi}(\tilde{\mu}),j}$. Hence, in any such dimension j , μ^* implies a policy that outperforms the policy maximizing $\tilde{\mu}$ w.r.t. labeler i 's true reward parameter θ_i^* . Hence, misreporting $\theta_{i,j} \neq \theta_{i,j}^*$ cannot be a strictly better strategy than truthfully reporting in dimension j .

Suppose that $\operatorname{sign}(\mu_j^*) \neq \operatorname{sign}(\theta_{i,j}^*)$. In this case, Lemma A.3 implies that $\operatorname{sign}(\tilde{\mu}_j) = \operatorname{sign}(\mu_j^*)$, which implies $\theta_{i,j}^* \cdot z_{\hat{\pi}(\tilde{\mu}),j} = \theta_{i,j}^* \cdot z_{\hat{\pi}(\mu^*),j}$. Once again misreporting is never a strictly better strategy than truthfully reporting θ_i^* .

We have thus confirmed that reporting θ_i^* is optimal irrespective of the other individuals' reports θ_{-i} .

Case 2 (direct access to θ_i , but not θ_{-i}): Let C_{-i} denote the product space of confidence sets derived from the preference data \mathcal{D}_{-i} of all labelers but labeler i . Once again, the key lies in the observation that the assumption of hyperrectangularity implies that the labelers get to strategize over each dimension independently.

The main concern that we must alleviate is that, by taking the minimum over the confidence sets, misreporting becomes beneficial for the labelers. To this end, let θ_i be the report of individual i , which we for now assume to be directly observable. Given preference data \mathcal{D}_{-i} and θ_i , the Pessimistic Median of MLEs computes a policy maximizing

$$\min_{\theta_{-i} \in C_{-i}} \sum_{j=1}^d \langle \operatorname{med}(\theta_{-i,j}, \theta_{i,j}), \mathbb{E}_{s \sim \rho} [\phi(s, \pi(s))] \rangle$$

Suppose that $\theta_{i,j}^* > 0$ for $j \in [d]$. Clearly, by design of the median, for any $\theta_{-i,j}$, it follows from the same argument as in Lemma A.3 that misreporting either has no effect on the policy (if the report is $\theta_{i,j} > 0$), or can only have an adverse effect for labeler i (if the report is $\theta_{i,j} < 0$). Hence, it is optimal for individual i to report θ_i^* irrespective of the other individuals' reported preference data \mathcal{D}_{-i} and the confidence sets that we construct.

Case 3 (no direct access to $\theta_1, \dots, \theta_k$): In the previous cases, we have shown that truthfully reporting is a dominant strategy for every individual $i \in [k]$. We will now see that this is in general no longer true when an individual cannot directly share their reward parameter with the algorithm. The reason lies in unintentional changes in the sign due to estimation errors and confidence sizes.

In the following, we assume that the preference data \mathcal{D}_{-i} of all labelers but labeler i are fixed arbitrarily and C_{-i} are the corresponding confidence sets that Pessimistic Median of MLEs constructs. From Case 2 we know that the policy

$$\hat{\pi}_i(\theta_i^*) := \operatorname{argmax}_{\pi \in \Pi} \min_{\theta_{-i} \in C_{-i}} \langle \operatorname{med}(\theta_i^*, \theta_{-i}), \mathbb{E}_{s \sim \rho}[\phi(s, \pi(s))] \rangle$$

is preferred over any other policy $\hat{\pi}(C_i)$ computed w.r.t. any confidence set C_i given by

$$\hat{\pi}_i(C_i) := \operatorname{argmax}_{\pi \in \Pi} \min_{\theta_i \in C_i} \min_{\theta_{-i} \in C_{-i}} \langle \operatorname{med}(\theta_i, \theta_{-i}), \mathbb{E}_{s \sim \rho}[\phi(s, \pi(s))] \rangle,$$

i.e., $J_{\theta_i^*}(\hat{\pi}_i(\theta_i^*)) \geq J_{\theta_i^*}(\hat{\pi}_i(C_i))$ for any confidence set C_i .

Let us now consider the confidence set C_i^* derived from \mathcal{D}_i^* , which is sampled according to the true reward parameter θ_i^* . By construction of the confidence sets, with probability at least $1 - \delta$, it follows from Lemma A.1 that for any $\theta_i \in C_i^*$:

$$\|\theta_i^* - \theta_i\|_{\Sigma_{\mathcal{D}_i}} \leq \|\theta_i^* - \hat{\theta}_i^{\text{MLE}}\|_{\Sigma_{\mathcal{D}_i}} + \|\hat{\theta}_i^{\text{MLE}} - \theta_i\|_{\Sigma_{\mathcal{D}_i}} \leq 2c\sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}}.$$

We now compare the difference in return w.r.t. θ_i^* of policy $\hat{\pi}_i(\theta_i^*)$ and $\hat{\pi}_i(C_i^*)$. To do so, we decompose the difference as follows:

$$\begin{aligned} & J_{\theta_i^*}(\hat{\pi}_i(\theta_i^*)) - J_{\theta_i^*}(\hat{\pi}_i(C_i^*)) \\ &= \left(J_{\theta_i^*}(\hat{\pi}_i(\theta_i^*)) - \min_{\theta_i \in C_i^*} J_{\theta_i}(\hat{\pi}_i(\theta_i^*)) \right) + \left(\min_{\theta_i \in C_i^*} J_{\theta_i}(\hat{\pi}_i(\theta_i^*)) - J_{\theta_i^*}(\hat{\pi}_i(C_i^*)) \right). \end{aligned}$$

Using Cauchy-Schwarz, the first difference can be rewritten and bounded as

$$\max_{\theta_i \in C_i^*} \langle \theta_i^* - \theta_i, \mathbf{z}_{\hat{\pi}_i(\theta_i^*)} \rangle \leq \|\theta_i^* - \theta_i\|_{\Sigma_{\mathcal{D}_i}} \|\mathbf{z}_{\hat{\pi}_i(\theta_i^*)}\|_{\Sigma_{\mathcal{D}_i}^{-1}}.$$

We then further decompose the second difference into

$$\begin{aligned} & \min_{\theta_i \in C_i^*} J_{\theta_i}(\hat{\pi}_i(\theta_i^*)) - J_{\theta_i^*}(\hat{\pi}_i(C_i^*)) \\ &= \left(\min_{\theta_i \in C_i^*} J_{\theta_i}(\hat{\pi}_i(\theta_i^*)) - \min_{\theta_i \in C_i^*} J_{\theta_i}(\hat{\pi}_i(C_i^*)) \right) + \left(\min_{\theta_i \in C_i^*} J_{\theta_i}(\hat{\pi}_i(C_i^*)) - J_{\theta_i^*}(\hat{\pi}_i(C_i^*)) \right). \end{aligned}$$

By definition of $\hat{\pi}_i(C_i^*)$, we have $\min_{\theta_i \in C_i^*} \langle \theta_i, \mathbf{z}_{\hat{\pi}_i(C_i^*)} \rangle \geq \min_{\theta_i \in C_i^*} \langle \theta_i, \mathbf{z}_{\hat{\pi}_i(\theta_i^*)} \rangle$ so that the first expression on the right hand side is less or equal to zero. Since $\theta_i^* \in C_i^*$, we also know that $\min_{\theta_i \in C_i^*} J_{\theta_i}(\pi) \leq J_{\theta_i^*}(\pi)$ for all $\pi \in \Pi$. Thus, on the good event when $\theta_i^* \in C_i^*$, we obtain:

$$\begin{aligned} J_{\theta_i^*}(\hat{\pi}_i(\theta_i^*)) - J_{\theta_i^*}(\hat{\pi}_i(C_i^*)) &\leq c\sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}} \cdot \|\mathbb{E}_{s \sim \rho}[\phi(s, \hat{\pi}_i(\theta_i^*)(s))] \|_{\Sigma_{\mathcal{D}_i}^{-1}} \\ &\leq c\sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}} \cdot \max_{\pi \in \Pi} \|\mathbb{E}_{s \sim \rho}[\phi(s, \pi(s))] \|_{\Sigma_{\mathcal{D}_i}^{-1}}. \end{aligned}$$

Note that the coverage coefficient on the right can be written as $\|\Sigma_{\mathcal{D}_i}^{-1/2} \mathbb{E}_{s \sim \rho}[\phi(s, \pi(s))] \|_2$.

We have here (arguably coarsely) upper bounded the coverage of $\hat{\pi}(\theta_i^*)$ by the uniform policy coverage $\kappa_i := \max_{\pi} \|\mathbb{E}_{s \sim \rho}[\phi(s, \pi(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}}$ of labeler's i data. We must do this here as the policy $\hat{\pi}_i(\theta_i^*)$ notably depends on the other labeler's reported preferences \mathcal{D}_{-i} and is thus hard to control or express explicitly. Overall, we have thus shown that being truthful is an approximately dominant strategy for labeler i under Pessimistic Median of MLEs. Hence, Pessimistic MoMLEs is $\mathcal{O}(\kappa_i \sqrt{d/n})$ -strategyproof.

Remark A.5. *As we wish to ensure strategyproofness, i.e., truthfulness is a dominant strategy, we could not control the needed coverage carefully, but had to take a worst-case perspective and consider uniform coverage of all policies as quantified by κ_i . Naturally, we would expect to improve upon this when considering incentive-compatibility instead of strategyproofness, i.e., showing that truthfulness forms an equilibrium but is not necessarily a dominant strategy profile. In that case, one can show that Pessimistic Median of MLEs is approximately incentive-compatible where instead of the uniform policy coverage the coverage of the output $\hat{\pi}^*$ of Pessimistic Median of MLEs given that everyone reports truthfully is enough. In other words, the coverage coefficient is given by $\|\mathbb{E}_{s \sim \rho}[\phi(s, \hat{\pi}^*(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}} \leq \kappa_i$.*

□

A.6 Proof of Theorem 4.2

Theorem 4.2. *Let $\hat{\pi}$ be the output of the Pessimistic Median of MLEs algorithm and suppose that all labelers report truthfully. With probability at least $1 - \delta$:*

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot \sqrt{\frac{d \log(k/\delta)}{k}} + \text{const} \cdot k \sqrt{\frac{d + \log(k/\delta)}{n}} \max_{i \in [k]} \|\mathbb{E}_{s \sim \rho}[\phi(s, \pi^*(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}}. \quad (2)$$

Note that $\|\mathbb{E}_{s \sim \rho}[\phi(s, \pi^*(s))]\|_{\Sigma_{\mathcal{D}_i}^{-1}} = \|\Sigma_{\mathcal{D}_i}^{-1/2} \mathbb{E}_{s \sim \rho}[\phi(s, \pi^*(s))]\|_2$.

Proof. We will decompose the suboptimality in various ways. To this end, let π^* denote the policy that maximizes social welfare and let $\hat{\pi}$ denote the policy computed by Pessimistic Median of MLEs. Recall the definition of the set of medians w.r.t. confidence sets C_1, \dots, C_k as $\mathcal{C} := \{\text{med}(\theta_1, \dots, \theta_k) : \theta_i \in C_i\}$ and let $\mathcal{A} := \{\frac{1}{k} \sum_{i=1}^k \theta_i : \theta_i \in C_i\}$ denote the set of averages. For convenience, we define for any π :

$$\mathbf{z}_\pi := \mathbb{E}_{s \sim \rho}[\phi(s, \pi(s))].$$

Moreover, we let

$$\theta_{\text{avg}}^* := \text{avg}(\theta_1^*, \dots, \theta_k^*) \quad \text{and} \quad \theta_{\text{med}}^* := \text{med}(\theta_1^*, \dots, \theta_k^*)$$

correspond to the true average and median, respectively. We now decompose the suboptimality as follows:

$$\begin{aligned} \text{SubOpt}(\hat{\pi}) &= \frac{1}{k} \sum_{i=1}^k \langle \theta_i^*, \mathbf{z}_{\pi^*} \rangle - \langle \theta_i^*, \mathbf{z}_{\hat{\pi}} \rangle \\ &= \langle \theta_{\text{avg}}^*, \mathbf{z}_{\pi^*} \rangle - \langle \theta_{\text{avg}}^*, \mathbf{z}_{\hat{\pi}} \rangle \\ &= \underbrace{\left(\langle \theta_{\text{avg}}^*, \mathbf{z}_{\pi^*} \rangle - \min_{\theta \in \mathcal{A}} \langle \theta, \mathbf{z}_{\pi^*} \rangle \right)}_{(I)} + \underbrace{\left(\min_{\theta \in \mathcal{A}} \langle \theta, \mathbf{z}_{\pi^*} \rangle - \langle \theta_{\text{avg}}^*, \mathbf{z}_{\hat{\pi}} \rangle \right)}_{(II)} \end{aligned}$$

In the following, we work on the good event such that $\theta_i^* \in C_i$ for all $i \in [k]$. Using a union bound, we can show that this event occurs with probability at least $1 - \frac{k}{d}$.

We can bound the first term (*I*) using that the confidences concentrate around the true parameter at a rate of $\sqrt{d/n}$ according to Lemma A.1 and considering the worst-case coverage of the optimal policy over all labeler's data. For some constant $c > 0$, this yields

$$\begin{aligned}
\langle \theta_{\text{avg}}^*, \mathbf{z}_{\pi^*} \rangle - \min_{\theta \in \mathcal{A}} \langle \theta, \mathbf{z}_{\pi^*} \rangle &= \max_{\theta \in \mathcal{A}} \langle \theta_{\text{avg}}^* - \theta, \mathbf{z}_{\pi^*} \rangle \\
&= \frac{1}{k} \max_{\theta_1 \in C_1} \dots \max_{\theta_k \in C_k} \sum_{i=1}^k \langle \theta_i^* - \theta_i, \mathbf{z}_{\pi^*} \rangle \\
&= \frac{1}{k} \sum_{i=1}^k \max_{\theta_i \in C_i} \langle \theta_i^* - \theta_i, \mathbf{z}_{\pi^*} \rangle \\
&\leq \frac{1}{k} \sum_{i=1}^k \max_{\theta_i \in C_i} \|\theta_i^* - \theta_i\|_{\Sigma_{\mathcal{D}_i}} \|\mathbf{z}_{\pi^*}\|_{\Sigma_{\mathcal{D}_i}^{-1}} \\
&\leq c \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}} \cdot \frac{1}{k} \sum_{i=1}^k \|\mathbf{z}_{\pi^*}\|_{\Sigma_{\mathcal{D}_i}^{-1}} \\
&\leq c \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}} \cdot \max_{i \in [k]} \|\mathbf{z}_{\pi^*}\|_{\Sigma_{\mathcal{D}_i}^{-1}}.
\end{aligned}$$

Bounding the second term (*II*) is more involved as the policy $\hat{\pi}$ is not maximizing the average but the pessimistic median. We further decompose the second term into four parts as follows:

$$\begin{aligned}
\min_{\theta \in \mathcal{A}} \langle \theta, \mathbf{z}_{\pi^*} \rangle - \langle \theta_{\text{avg}}^*, \mathbf{z}_{\hat{\pi}} \rangle &= \left(\min_{\theta \in \mathcal{A}} \langle \theta, \mathbf{z}_{\pi^*} \rangle - \min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\pi^*} \rangle \right) + \left(\min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\pi^*} \rangle - \min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\hat{\pi}} \rangle \right) \\
&\quad + \left(\min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\hat{\pi}} \rangle - \langle \theta_{\text{med}}^*, \mathbf{z}_{\hat{\pi}} \rangle \right) + \left(\langle \theta_{\text{med}}^*, \mathbf{z}_{\hat{\pi}} \rangle - \langle \theta_{\text{avg}}^*, \mathbf{z}_{\hat{\pi}} \rangle \right).
\end{aligned}$$

We first show that the second and third term are less or equal to zero. We have

$$\min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\pi^*} \rangle \leq \min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\hat{\pi}} \rangle,$$

since $\hat{\pi}$ maximizes $\min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\pi} \rangle$ by definition of the Pessimistic Median of MLEs. Moreover, we see that

$$\min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\hat{\pi}} \rangle \leq \langle \theta_{\text{med}}^*, \mathbf{z}_{\hat{\pi}} \rangle,$$

as the true median is contained in the confidence set \mathcal{C} on the good event when $\theta_i^* \in C_i$. Hence, both the second and third term can be bounded from above by zero.

To bound the first term, we once again decompose the expression as follows:

$$\min_{\theta \in \mathcal{A}} \langle \theta, \mathbf{z}_{\pi^*} \rangle - \min_{\theta \in \mathcal{C}} \langle \theta, \mathbf{z}_{\pi^*} \rangle = \underbrace{\min_{\theta \in \mathcal{A}} \langle \theta - \theta_{\text{avg}}^*, \mathbf{z}_{\pi^*} \rangle}_{(a)} + \underbrace{\langle \theta_{\text{avg}}^* - \theta_{\text{med}}^*, \mathbf{z}_{\pi^*} \rangle}_{(b)} + \underbrace{\max_{\theta \in \mathcal{C}} \langle \theta_{\text{med}}^* - \theta, \mathbf{z}_{\pi^*} \rangle}_{(c)}. \quad (3)$$

Similarly to before, using Lemma A.1, we bound (a) as

$$\min_{\theta \in \mathcal{A}} \langle \theta - \theta_{\text{avg}}^*, \mathbf{z}_{\pi^*} \rangle \leq c \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}} \cdot \max_{i \in [k]} \|\mathbf{z}_{\pi^*}\|_{\Sigma_{\mathcal{D}_i}^{-1}}.$$

For (b), it follows from Cauchy-Schwarz and Lemma A.2 that

$$\langle \theta_{\text{avg}}^* - \theta_{\text{med}}^*, \mathbf{z}_{\pi^*} \rangle \leq \|\theta_{\text{avg}}^* - \theta_{\text{med}}^*\|_2 \|\mathbf{z}_{\pi^*}\|_2 \leq c \sqrt{\frac{d \log(1/\delta)}{k}} \cdot \|\mathbf{z}_{\pi^*}\|_2. \quad (4)$$

For (c), first note that we can write the difference between two medians as the telescoping sum

$$\begin{aligned} \text{med}(\theta_1^*, \dots, \theta_k^*) - \text{med}(\theta_1, \dots, \theta_k) \\ = \sum_{i=1}^k \text{med}(\theta_1^*, \dots, \theta_i^*, \theta_{i+1}, \dots, \theta_k) - \text{med}(\theta_1^*, \dots, \theta_{i-1}^*, \theta_i, \dots, \theta_k). \end{aligned}$$

By definition of the median, each difference on the right hand side can be bounded in terms of the difference $\theta_i^* - \theta_i$. Using Lemma A.1 and the fact that $\theta_i \in C_i$ for all $i \in [k]$, we obtain

$$\begin{aligned} \max_{\theta \in \mathcal{C}} \langle \theta_{\text{med}}^* - \theta, \mathbf{z}_{\pi^*} \rangle &\leq \sum_{i=1}^k \|\theta_i^* - \theta_i\|_{\Sigma_{D_i}} \|\mathbf{z}_{\pi^*}\|_{\Sigma_{D_i}^{-1}} \\ &\leq ck \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}} \cdot \|\mathbf{z}_{\pi^*}\|_{\Sigma_{D_i}^{-1}}. \end{aligned}$$

The proof is complete by combining these bounds. \square

A.7 Proof of Proposition 4.3

Proposition 4.3. *When the labelers report their preferences according to any weakly dominant strategy under Pessimistic Median of MLEs, with probability at least $1 - \delta$, the output $\hat{\pi}$ satisfies:*

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot \sqrt{\frac{d \log(k/\delta)}{k}} + \text{const} \cdot k \sqrt{\frac{d + \log(k/\delta)}{n}} \max_{i \in [k]} \|\mathbb{E}_{s \sim \rho} [\phi(s, \pi^*(s))] \|_{\Sigma_{D_i}^{-1}}.$$

Proof. Let $\theta_i \in \mathbb{R}^d$ be the reward parameter according to which labeler $i \in [k]$ samples its preferences under a weakly dominant strategy. The intuition for the result is fairly straightforward so that we describe it here first. First of all, we have seen in the proof of Theorem 4.1 that due to the median rule a labeler cannot achieve an individually better outcome by misreporting the sign of its reward parameter (see Lemma A.3 and Lemma A.4). As a result, θ_i will have identical signs to θ_i^* but potentially exaggerate its magnitude. Crucially, such exaggeration cannot worsen the suboptimality as it only helps to prevent flipped signs as we are taking the worst-case over confidence sets. Here, it is also worth noting that the primary reasons why Pessimistic Median of MLEs is approximately strategyproof are the estimation errors and the pessimism selection of the median over potentially large confidence sets.

Any weakly dominant strategy must preserve signs. Assume for contradiction that there exists some coordinate j such that $\theta_{i,j}^* > 0$ but the labeler's chosen reward parameter is such that $\theta_{i,j} < 0$ (or similarly $\theta_{i,j}^* < 0$ but $\theta_{i,j} > 0$). By the hyperrectangular assumption, the Pessimistic Median of MLEs algorithm outputs a policy that maximizes each dimension of the feature space independently. Specifically, $\mathbf{z}_{\hat{\pi},j} = \mathbb{E}_{s \sim \rho} [\phi(s, \hat{\pi}(s))]_j$ will be positive if the considered median is positive and vice versa. Hence, by nature of the median, flipping the sign of coordinate j can only have an adverse effect for labeler i (see Lemma A.3) and no such strategy can be weakly dominant.

By the same argument, if $\theta_{i,j}^* < 0$ but $\theta_{i,j} > 0$, then labeler i would risk pushing the aggregator's dimension j to be positive, contrary to its true negative preference, and thus risk reducing its true utility in that dimension. Hence it cannot be a weakly dominant strategy to flip signs in that scenario either. Consequently, in every dimension j , a weakly dominant report $\theta_{i,j}$ must preserve $\text{sign}(\theta_{i,j}) = \text{sign}(\theta_{i,j}^*)$.

Exaggeration benefits Pessimistic Median of MLEs. By the hyperrectangular (“sign-based”) structure, the decision in each coordinate j of the learned policy depends essentially on whether

the aggregated median is positive or negative. Pessimistic Median of MLEs aggregates each labeler i 's confidence set C_i by taking a coordinate-wise median over a selection $\theta_i \in C_i$. Thus, to form the median, it chooses exactly one $\theta_{i,j}$ from each C_i and then takes the median value among these k numbers. Assume that the labeler i 's original (w.l.o.g.) positive coordinate is $\theta_{i,j}^*$, whereas its inflated coordinate is $\theta_{i,j} > \theta_{i,j}^*$. Under the inflated reported reward parameter, the labeler's MLE and confidence set for dimension j shift toward strictly larger positive values (note that the covariance matrix $\Sigma_{\mathcal{D}_i}$ is positive definite). Consequently, the set of considered medians μ_j for $\mu \in \mathcal{C}$ (i.e. all possible ways to pick $\theta_{i,j} \in C_i$ for $i = 1, \dots, k$ and take their coordinate-wise median) does not move down: it can only stay the same or shift to more positive values. Intuitively, replacing one of the i entries by a strictly larger positive number cannot decrease the median.

Hence, when labeler i is misreporting $\theta_{i,j}$ such that $\theta_{i,j} > \theta_{i,j}^*$ (while keeping the same sign), this cannot worsen the suboptimality of the final policy, but only, in some special cases, strictly lower suboptimality by “protecting” the sign within the confidence set. Since this argument holds for any dimension j , it follows that an entire sign-preserving inflation by labeler i cannot yield a higher suboptimality than the truthful report would. \square

A.8 Proof of Corollary 4.5

Corollary 4.5. *When there is only a single labeler, with probability at least $1 - \delta$:*

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot \sqrt{\frac{d + \log(k/\delta)}{n}} \|\mathbb{E}_{s \sim \rho}[\phi(s, \pi^*(s))]\|_{\Sigma_{\mathcal{D}_1^{-1}}}.$$

When all k labelers have the same reward function, with probability at least $1 - \delta$:

$$\text{SubOpt}(\hat{\pi}) \leq \text{const} \cdot k \sqrt{\frac{d + \log(k/\delta)}{n}} \max_{i \in [k]} \|\mathbb{E}_{s \sim \rho}[\phi(s, \pi^*(s))]\|_{\Sigma_{\mathcal{D}_i^{-1}}}.$$

Proof. When $k = 1$ the claimed result follows directly from setting $k = 1$ in our previous suboptimality bounds (see Theorem 4.2).

Next, suppose that all $k \geq 1$ labelers have the same reward parameter $\theta^* = \theta_1^* = \dots = \theta_k^*$. As a result, the true average and median coincide and we have $\theta^* = \theta_{\text{avg}}^* = \theta_{\text{med}}^*$. To bound the suboptimality of the Pessimistic Median of MLEs algorithm in this special case we take the same steps as in the proof of Theorem 4.2 in Section A.6 with the difference that the expression (b) in equation (3) is zero since $\theta^* = \theta_{\text{avg}}^* = \theta_{\text{med}}^*$. This yields the claimed upper bound. \square

A.9 Proof of Corollary 4.6

Corollary 4.6. *Suppose $\mathcal{W}(\pi^*) > 0$ is constant. When the number of samples is sufficiently large and provide sufficient coverage of the optimal policy, with probability at least $1 - \delta$, the approximation ratio of the Pessimistic Median of MLEs algorithm is given by $\alpha(\rho, \hat{\pi}) \geq 1 - \mathcal{O}(\sqrt{d \log(k/\delta)/k})$.*

Proof. For n sufficiently large and sufficient coverage of the optimal policy, Theorem 4.2 implies that with probability at least $1 - \delta$:

$$\text{SubOpt}(\hat{\pi}) := \mathcal{W}(\pi^*) - \mathcal{W}(\hat{\pi}) \leq c \sqrt{\frac{d \log(k/\delta)}{n}}$$

for some constant $c > 0$. As a result, the approximation ratio is upper bounded as

$$\alpha(\rho, \hat{\pi}) := \frac{\mathcal{W}(\hat{\pi})}{\mathcal{W}(\pi^*)} = 1 - \frac{\mathcal{W}(\pi^*) - \mathcal{W}(\hat{\pi})}{\mathcal{W}(\pi^*)} \geq 1 - c \sqrt{\frac{d \log(k/\delta)}{n}},$$

where we used that $\mathcal{W}(\pi^*) > 0$ is constant by assumption. \square

A.10 Proof of Theorem 5.1 and Theorem 5.2

Proof. We can prove Theorem 5.1 and Theorem 5.2 in a similar way we proved the analogous results in the contextual bandit problem. We refrain from reiterating and restating all necessary steps to prove these results as they are almost identical to before. Most importantly, a similar MLE concentration bound holds for MDPs as for contextual bandits.

Lemma A.6 (MLE Concentration Bound for MDPs). *With probability at least $1 - \delta$,*

$$\|\hat{\theta}_i^{\text{MLE}} - \theta_i^*\|_{\Sigma_{\mathcal{D}_i}} \leq \text{const} \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n}},$$

where $\gamma := 1/(2 + \exp(-HLB) + \exp(HLB))$. The covariance matrix $\Sigma_{\mathcal{D}_i}$ is given by $\Sigma_{\mathcal{D}_i} = \sum_{j=1}^n x^{i,j}(x^{i,j})^\top$ where $x^{i,j} = \sum_{h=1}^H (\phi(s_h^{i,j}, a_h^{i,j}) - \phi(\bar{s}_h^{i,j}, \bar{a}_h^{i,j}))$ with $s_1^{i,j} = \bar{s}_1^{i,j} = s^{i,j}$.

Swapping the initial state distribution ρ (i.e., context distribution) for the state occupancy q_π as defined in Section 5, we can follow the same line of argument as in Section A.5 to prove Theorem 5.1. \square