Near-Equivalence Between Bounded Regret and Delay Robustness in Interactive Decision Making

Enoch Hyunwook Kang EHWKANG@UW.EDU
Foster School of Business
University of Washington

P. R. Kumar PRK@TAMU.EDU
Department of Electrical & Computer Engineering
Texas A&M University

Abstract

Interactive decision making, encompassing bandits, contextual bandits, and reinforcement learning, has recently been of interest to theoretical studies of experimentation design and recommender system algorithm research. Recently, it has been shown that the well-known Graves-Lai constant being zero is a necessary and sufficient condition for achieving bounded (or constant) regret in interactive decision making. As this condition may be a strong requirement for many applications, the practical usefulness of pursuing bounded regret has been questioned. In this paper, we show that the condition of the Graves-Lai constant being zero is also necessary to achieve delay model robustness when reward delays are unknown (i.e., when feedbacks are anonymous). Here, model robustness is measured in terms of ϵ-robustness, one of the most widely used and one of the least adversarial robustness concepts in the robust statistics literature. In particular, we show that ϵ-robustness cannot be achieved for a consistent (i.e., uniformly sub-polynomial regret) algorithm however small the nonzero ϵ value is when the Grave-Lai constant is not zero. While this is a strongly negative result, we also provide a positive result for linear rewards models (Linear contextual bandits, Reinforcement learning with linear MDP) that the Grave-Lai constant being zero is also sufficient for achieving bounded regret without any knowledge of delay models, i.e., the best of both the efficiency world and the delay robustness world.

Keywords: Bandits, Reinforcement learning, Bounded regret, Delay robustness

1 Introduction

We consider the cost of addressing stochastic and unknown reward delays in *Decision-Making with Structured Observations (DMSO)* (Wagenmaker and Foster, 2023; Dong and Ma, 2023), which generalizes interactive decision-making problems such as structured bandit, contextual bandit, and reinforcement learning (see Section 2.1 and Appendix A). In many real-life applications of interactive decision-making problems, stochastic and unknown delays in reward makes it hard to attribute the sequence of observed outcomes to

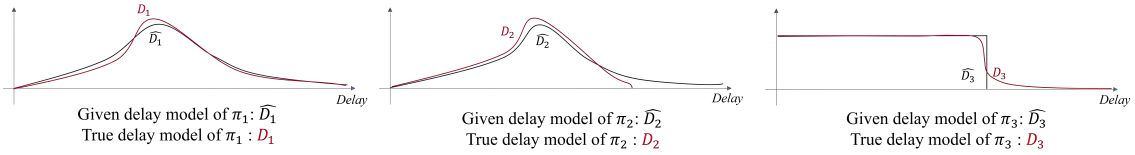

Figure 1: Examples of misspecification of the reward delay model of decisions (π_1, π_2, π_3)

the sequence of our decisions. In medical treatments, for example, a doctor cannot easily be sure whether a medical outcome is due to the effect of current treatment or due to some other previously taken treatment's delayed effect. This type of reward delays in decisions is called an 'unknown reward delays' (Li et al., 2019) or 'delayed anonymous feedback' (Pike-Burke et al., 2018; Cesa-Bianchi et al., 2018)[1].

Knowledge of the probabilistic distribution of each decision's reward delay, combined with the careful design of algorithms, may help to resolve this reward attribution problem under stochastic and anonymous reward delays (Pike-Burke et al., 2018). However, those delay models themselves may be misspecified (Wang et al., 2021). Therefore, whether we can design an algorithm that is robust to model misspecification becomes a main concern in the problems with stochastic, delayed, and anonymous rewards.

One of the most widely used concepts of model misspecification in the robust statistics literature is ϵ-*robustness* (Huber, 2004). Given a parameter $\epsilon > 0$ and true distribution D, a model distribution $\widehat{D}$ is called an ϵ-(general) contamination of D if $d_{TV}(D, \widehat{D}) \leq \epsilon$, where d_{TV} denotes the total variation distance function[2]. Figure 1 illustrates some examples of ϵ-contamination of the delay models. As ϵ-robustness is also one of the weakest (i.e., least adversarial), most elementary notion of robustness (Diakonikolas and Kane, 2023), the first question on an algorithm's delay robustness will be "up to which ϵ the algorithm's properties are robust to ϵ-contamination of delay model misspecification?".

In this paper, we prove that no consistent (i.e., uniformly sub-polynomial regret) algorithm for DMSO can be designed to be robust to ϵ-contamination of delay model misspecification unless DMSO's Graves-Lai constant (Graves and Lai, 1997; Dong and Ma, 2023; Wagenmaker and Foster, 2023) being zero. While this is a strong negative result, we also provide a positive result for linear DMSO problems (linear contextual bandit, reinforcement learning with linear MDPs) that the Graves-Lai constant (Graves and Lai, 1997) being zero is sufficient for achieving bounded regret without any knowledge of delay models. As the Graves-Lai constant being zero holds if and only if we can achieve bounded regret (Dong and Ma, 2023; Wagenmaker and Foster, 2023), the results in this paper strongly motivate the practical usefulness of achieving bounded regret.

[1]Most research in this literature focuses on the setting of delayed, anonymous, and *aggregated* (DAAF) feedback, where we only observe sum of the rewards arriving at each episode. Here, we consider impossibility results for the strictly easier case, where we observe each anonymous delayed reward separately.

[2]The total variation distance $d_{\mathrm{TV}}(\nu, \upsilon)$ is defined as $\frac{1}{2}\|\nu - \upsilon\|_1 = \sup_{E \in \Sigma} |\nu(E) - \upsilon(E)|$, where Σ stands for the measurable sets on which two distributions ν and υ are defined.

2 Preliminaries

2.1 Decision-Making with Structured Observations (DMSO)

Here, we focus on an interactive decision-making problem framework called *Decision-Making with Structured Observations (DMSO)* (Dong and Ma, 2023; Wagenmaker and Foster, 2023) which generalizes bandit, contextual bandit, and episodic reinforcement learning problems (see Appendix A). Like other interactive decision-making problems, it is characterized by the environment and its learning protocol. The environment of a DMSO problem framework is specified by a decision space Π, a reward space $\mathcal{R}$, an observation space $\mathcal{O}$ and a model class $\mathcal{F} = \prod_{\pi \in \Pi} \mathcal{F}_\pi$, where $\mathcal{F}_\pi \subseteq \triangle_{\mathcal{R} \times \mathcal{O}}$ (Here, $\triangle_E$ denotes the collection of probability distributions over a set E). A ground-truth model $f^\star \in \mathcal{F}$ governs the rewards and the observations based on the decisions made in the rounds. While $f^\star$ is unknown to the learner, it is typically assumed that a set $\mathcal{F}$ that includes $f^\star$ is known to the learner.

The learning protocol of DMSO problem consists of n rounds. In round $k \leq n$,

1. The learner makes a decision $\pi_k \in \Pi$.
2. A reward $r_k \in \mathcal{R}$ and an observation $o_k \in \mathcal{O}$ are generated, where $(r_k, o_k) \sim f^\star_{\pi_k} \in \mathcal{F}_{\pi_k}$
3. Learner observes o_k. The learner also observes R_k, the set of rewards that arrive at the round k. R_k is equivalent to r_k when there are no reward delays.

2.2 Learning Algorithm and its Regret for DMSO

Given that we characterized the DMSO problem framework, we can now describe a learning algorithm. Let h_k be the history up to round k, i.e., $h_k = (\pi_1, R_1, o_1), \ldots, (\pi_{k-1}, R_{k-1}, o_{k-1})$, and $\mathcal{H}$ be the set of all possible histories of rounds for $k \geq 1$. A learning algorithm A is defined as an element of $\mathcal{A} \subseteq (\mathcal{H} \mapsto \triangle_\Pi)$, which is a subset of the set of all possible mappings from the history space $\mathcal{H}$ to the set of all possible distributions over Π. At each round k, given the history $h_k \in \mathcal{H}$, learning algorithm $A \in \mathcal{A}$ chooses $p_k = A(h_k) \in \triangle_\Pi$. The decision at round k, π_k, is sampled from p_k. Note that $f \in \mathcal{F}$, $A \in \mathcal{A}$ and the round n completely determine the stochastic behavior of the learning protocol up to round n, i.e., they induce a probability distribution we call $P_{f,n,A}[\cdot]$ over the set of all histories up to round n. We also denote the respective expectation by $\mathbb{E}_{f,n,A}[\cdot]$. When the meaning is clear from the context, we use $P_{f,n}[\cdot]$ and $\mathbb{E}_{f,n}[\cdot]$ instead of $P_{f,n,A}[\cdot]$ and $\mathbb{E}_{f,n,A}[\cdot]$.

Given $(r, o) \sim f_\pi$, we denote $\mu_{f_\pi} := \mathbb{E}_{f_\pi}[r]$; when the meaning is clear from context, we use $\mu_{f_\pi} = \mu_\pi$ and $\mu_{f^\star{}_\pi} = \mu^\star_\pi$. Furthermore, let $\pi_f \in \arg\max_{\pi \in \Pi} \mu_{f_\pi}$ denote an optimal decision for the model f; when the meaning is clear from context, we use $\pi^\star := \pi_{f^\star}$. We denote $\{g \in \mathcal{F} \mid \pi_g = \pi_f\}$ by $\mathcal{F}(f)$. The suboptimality gap of the decision π for model f is defined as $\mathbf{d}_f(\pi) := \mu_{f_{\pi_f}} - \mu_{f_\pi}$.

When the ground truth model is f, choosing π_f every round until round n yields the largest total reward until round n. Therefore, we can measure the optimality of an algorithm A until round n in terms of regret, which is defined by $\mathrm{Reg}_{f,A}(n) := \mathbb{E}_{A,f,n}\left[\sum_{k=1}^n \mathbf{d}_f(\pi_k)\right]$.

2.3 Consistent Learning Algorithm and Asymptotic Regret Lower Bound

It is natural to exclude non-uniformly good algorithms, i.e., algorithms that are specifically designed to achieve small regret for some instances and suffer polynomially increasing regret in some other instances. The notion of *consistent* algorithm formalizes this idea.

Definition 1 (Consistent learning algorithm (Graves and Lai, 1997))**.** *A learning algorithm A is called consistent if $\mathrm{Reg}_{f,A}(n) = o(n^p)$ holds for every $p > 0$ and $f \in \mathcal{F}$.*

For DMSO problems, it has been recently shown that any consistent algorithm's instance-dependent regret must satisfy the following asymptotic regret lower bound.

Theorem 1 (Dong and Ma (2022) Dong and Ma (2023))**.** *For every instance $f \in \mathcal{F}$, the expected regret of any consistent algorithm A satisfies $\limsup_{n\to\infty} \frac{\mathrm{Reg}_{f,A}(n)}{\ln n} \geq \mathcal{C}(f) = \lim_{n\to\infty} \mathcal{C}(f,n)$, where $\mathcal{C}(f,n)$ is the solution to the optimization equation*

$$\begin{aligned} \mathcal{C}(f,n) \triangleq \min_{w\in\mathbb{R}_+^{|\Pi|}} & \sum_{\pi\in\Pi} w_\pi \mathbf{d}_f(\pi) \\ s.t. & \sum_{\pi\in\Pi} w_\pi D_{\mathrm{KL}}(f_\pi \| g_\pi) \geq 1 \ , \ \forall g \in \mathcal{F}(f)^c \\ & \|w\|_\infty \leq n, \end{aligned} \tag{1}$$

where D_{KL} is the KL divergence and $\mathcal{F}(f) := \{g \in \mathcal{F} \mid \pi_g = \pi_f\}$.

Corollary 1. *A consistent algorithm achieves sub-logarithmic regret only if $\mathcal{C}(f^\star) = 0$.*

According to Corollary 1, the design of a learning algorithm that a priori assures sub-logarithmic regret requires the condition $\mathcal{C}(f) = 0$ for $f \in \mathcal{F}$ to hold.

2.4 Main Model: Unknown Reward Delays and Robustness

Denote the true reward delay distributions of each decision $\pi \in \Pi$ by D_π. Every time π_k is determined at each round k, $d_k \sim D_{\pi_k}$ is generated along with the generation of $(r_k, o_k) \sim f_\pi$. While o_k is observed immediately at round k, r_k is scheduled to arrive at round $d_k + k$. As discussed earlier, we consider *unknown reward delays* (also called anonymous feedback) throughout the paper, as in Pike-Burke et al. (2018)[3].

As also discussed earlier, we specifically model delay robustness against ϵ-contamination. That is, we address the case when $\hat{D}_\pi$ is a result of ϵ-contamination of the delay distribution model D_π, i.e., $d_{TV}(D_\pi, \hat{D}_\pi) \leq \epsilon$. The formal definition of robustness in terms of delay distribution knowledge is as follows.

Definition 2. *We say that a consistent algorithm has ϵ-delay robustness if it is consistent when the given delay distributions are ϵ-contaminatations of the true delay distributions.*

[3]While Pike-Burke et al. (2018) considers aggregated anonymous reward arrivals, we show impossibility result that holds even for the weaker setting, where reward arrivals are still anonymous but not aggregated.

3 Main Results for DMSO

3.1 Main Result 1: Delay Robustness Requires $\mathcal{C}(f^\star) = 0$

Given the definition of ϵ-delay robustness provided in Section 2.4, the main question is when a consistent, ϵ-delay robust algorithm exists. The answer is quite negative: unless $\mathcal{C}(f^\star) = 0$ holds, no consistent algorithm can be ϵ-delay robust, however small $\epsilon > 0$ is. That is, *delay robustness can be achieved only if bounded regret can be achieved.*

This theoretical result can be intuitively understood as follows. When the reward delay model is precisely known, i.e., when the reward delay model is not contaminated, we might be able to address this challenge by designing a good algorithm that makes the probability of confusion in reward attribution as small as we want. However, in the case of ϵ-contamination of delay models, under minor technical assumptions (Appendix B), we can provide a delay model contamination that makes the precision of any consistent algorithm's reward attribution no better than $1 - \delta$ for some $\delta > 0$. This leads to reward distribution suffering δ-contamination. Theorem 2 formally states this result.

Theorem 2. *Under minor assumptions (see Appendix B), regardless of how small $\epsilon > 0$ is, a consistent learning algorithm is ϵ-delay robust only if $\mathcal{C}(f^\star) = 0$.*

The proof of Theorem 2 can be found in Appendix C.1. Theorem 2 implies that the concept of consistent algorithm fails even with a very small misspecification of the delay model unless $\mathcal{C}(f^\star) = 0$. Since we want to design a learning system with existence of a consistent algorithm that works for all instances of $f \in \mathcal{F}$, we need $\mathcal{C}(f) = 0$ for $f \in \mathcal{F}$.

3.2 Main Result 2: Design of Robust, Bounded Regret Algorithm

While the result shown above a negative result, the results we provide here are positive results on the condition required to assure achieving the best-of both worlds, i.e., achieving bounded regret and robustness to any delay-model misspecification at the same time. Before answering this question, we need the following Lemma 2 to see what $C(f) = 0$ for $f \in \mathcal{F}$ implies. Let us denote the algorithm that always chooses decision $\pi \in \Pi$ as $\overline{\pi}$.

Lemma 2. *Suppose that $f \in \mathcal{F}$ is the ground-truth model instance. Then $C(f) = 0$ implies that $D_{\mathrm{KL}}\left(P_{f,n,\overline{\pi_f}} \| P_{g,n,\overline{\pi_f}}\right) = \Omega(n)$ holds for $g \in \mathcal{F}(f)^c$.*

Lemma 2 shows how informative π_f is when the true hypothesis is f. When the true hypothesis is not f, π_f can be arbitrarily uninformative. The natural question that arises is how much cross-informativeness (informativeness of $\pi_h \in \Pi$ for the ground truth $f \in \mathcal{F}$ when $h \neq f$) is sufficient for us to achieve bounded regret.

Assumption 1 (Cross-informativeness)**.** *Suppose that $f \in \mathcal{F}$ is the ground-truth model. Then for any $h \in \mathcal{F}$, $D_{\mathrm{KL}}\left(P_{f,n,\overline{\pi_h}} \| P_{g,n,\overline{\pi_h}}\right) = \omega(\ln n)$ holds for $g \in \mathcal{F}(f)^c$.*

We also need a technical assumption 2, which excludes trivially informative cases where hypotheses f and g are almost immediately distinguished given the observation $o \in \mathcal{O}$.

Assumption 2. *For all $f, g \in \mathcal{F}$, $\pi \in \Pi$, and $o \in \mathcal{O}$, $D_{KL}(f_\pi(\cdot \mid o) \| g_\pi(\cdot \mid o)) < \infty$.*

Note that we can well-define $\beta = (\sup_{g \in \Pi, \pi \in \Pi, E \in \mathcal{E}_\pi} \frac{\mathrm{d}f_\pi(\cdot|o)}{\mathrm{d}g_\pi(\cdot|o)}(E))^{-1}$ (where $\mathcal{E}_\pi$ denote the collection of measurable sets for $f_\pi(\cdot \mid o)$ and $g_\pi(\cdot \mid o)$), as Assumption 2 holds if and only if the log-likelihood ratio $\ln \frac{f_\pi(\cdot|o)}{g_\pi(\cdot|o)}$ is well-defined on the support of g_π and is finite *a.e.*.

We now provide a simple algorithm, which is shown to achieve bounded regret without any knowledge of the delay model under Assumption 1 and 2. For this, we need to define the concept of max-contamination $\delta_\pi^{\max}(k)$. For details, see Appendix D and E.

Algorithm 1: Simply-Test-to-Commit (ST2C) Algorithm

for $n = 0$ **do**
  Choose any $h \in \mathcal{F}$, set $\widehat{f} = h$
for $n = 1, 2, \ldots$ **do**
  Choose $\pi_n = \pi_{\widehat{f}}$
  Observe o_n and R_n, newly compute $\delta^{\max}_{\pi_{\widehat{f}}}(n)$
  for $g \in \mathcal{F} \setminus \widehat{f}$ **do**
    if $\{\sum_{k=1}^n \ln \frac{g^c_{\pi_{\widehat{f}}}(k)}{\widehat{f}^c_{\pi_{\widehat{f}}}(k)} \geq 2 \ln n + \sum_{k=1}^n \frac{2}{\sqrt{\beta}} \delta^{\max}_{\pi_{\widehat{f}}}(k)\}$ **then**
      Set $\hat{f} = g$

Theorem 3. *Under Assumptions 1 and 2, the algorithm ST2C (Algorithm 1). which does not require any knowledge of the delay distribution model, achieves bounded regret.*

The question remains as to how strong Assumption 1 is. Theorem 4 and Theorem 5 (Appendix F, G) show that $\mathcal{C}(f) = 0$ for $f \in \mathcal{F}$ makes linear systems satisfy the condition of Assumption 1, completing the equivalence between bounded regret and delay robustness.

Theorem 4. *Under the linear contextual bandit setting in Hao et al. (2020) and almost the same condition as $\mathcal{C}(\theta) = 0$ for $\theta \in \Theta$, $D_{\mathrm{KL}}\left(P_{\theta^\star, n, \overline{\pi_\theta}} \| P_{\theta', n, \overline{\pi_\theta}}\right) = \Omega(n)$ holds for $\theta' \in \mathcal{F}(f)^c$.*

Theorem 5. *Under linear MDP setting with horizon H in Papini et al. (2021a) and close to $\mathcal{C}(\theta) = 0$ for $\theta \in \Theta$, $D_{\mathrm{KL}}\left(P_{\theta^\star_h, n, \overline{\pi_\theta}} \| P_{\theta'_h, n, \overline{\pi_\theta}}\right) = \Omega(n)$ holds for $\theta'_h \in \mathcal{F}(f)^c, h \in [H]$.*

4 Conclusion

We show that no consistent algorithm for DMSO can achieve any level of delay model robustness, unless the previously known condition for bounded regret holds. Given the condition for bounded regret, on the other hand, you can achieve bounded regret without any knowledge of delay models in case of linear systems. Achieving such best-of-both-worlds result for general non-linear systems will be an interesting future study.

References

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. In Conference On Learning Theory, pages 750–773. PMLR, 2018.

Thomas M Cover. Elements of information theory. John Wiley & Sons, 1999.

Ilias Diakonikolas and Daniel M Kane. Algorithmic high-dimensional robust statistics. Cambridge University Press, 2023.

Kefan Dong and Tengyu Ma. Asymptotic instance-optimal algorithms for interactive decision making, 2023.

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. SIAM journal on control and optimization, 35(3):715–743, 1997.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In International Conference on Artificial Intelligence and Statistics, pages 3536–3545. PMLR, 2020.

Peter J Huber. Robust statistics, volume 523. John Wiley & Sons, 2004.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory, pages 2137–2143. PMLR, 2020.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

Bingcong Li, Tianyi Chen, and Georgios B Giannakis. Bandit online learning with unknown delays. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 993–1002. PMLR, 2019.

Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. Advances in Neural Information Processing Systems, 34:16371–16383, 2021a.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In International Conference on Machine Learning, pages 8371–8380. PMLR, 2021b.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In International Conference on Machine Learning, pages 4105–4113. PMLR, 2018.

Sergio Verdú. Total variation distance and the distribution of relative information. In 2014 Information Theory and Applications Workshop (ITA), pages 1–3. IEEE, 2014.

Andrew Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. arXiv preprint arXiv:2304.12466, 2023.

Siwei Wang, Haoyun Wang, and Longbo Huang. Adaptive algorithms for multi-armed bandit with composite and anonymous feedback. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 10210–10217, 2021.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In International Conference on Machine Learning, pages 10978–10989. PMLR, 2020.

Appendix A. Special Cases of DMSO (Wagenmaker and Foster, 2023)

- In finite-armed bandit problems, each round is an arm pull. Π is the arm space, and $\mathcal{R}$ is the space of rewards from arms. Since there is no observation space $\mathcal{O}$, the model class degenerates to $\mathcal{F} \subseteq (\Pi \mapsto \Delta_{\mathcal{R}})$.

- In contextual bandit problems, each round is an arm pull. Π is the set $(\mathcal{X} \mapsto \mathcal{A})$ of all policies, where $\mathcal{X}$ is the context space and the $\mathcal{A}$ is the arm space. The reward space $\mathcal{R}$ is the space of rewards from arms. The observation space $\mathcal{O}$ is $\mathcal{X}$, where the kth round's observation $o_k \in \mathcal{O}$ (which results from π_k) is the $k+1$th round's context. Since the future context arrival is not affected by previous decisions, the model class degenerates to $\mathcal{F} \subseteq (\Pi \mapsto \Delta_{\mathcal{R}})$.

- In episodic reinforcement learning problems, each round is an episode. Π is the set $(\mathcal{S} \mapsto \mathcal{A})$ of all policies, where $\mathcal{S}$ is the space of all possible states and $\mathcal{A}$ is the action space. The reward space $\mathcal{R}$ is the space of value functions at each initial state, and the observation space $\mathcal{O}$ is the set of all possible sequences of action choices, state transitions, and received rewards in one episode. The model class $\mathcal{F}$ is characterized jointly by the initial state distribution and the transition kernel, which are shared across all the episodes.

Appendix B. Technical assumptions for Theorem 2 of Section 3.1

Assumption 3. *For the family of distributions $\mathcal{D}_{r|o} := \{f_\pi(\cdot \mid o) \mid f \in \mathcal{F}, \pi \in \Pi, o \in \mathcal{O}\}$, there exists a function $q(\delta)$ s.t. for $D_1, D_2 \in \mathcal{D}_{r|o}$, $|\mathbb{E}[D_1] - \mathbb{E}[D_2]| \leq q(\delta)$ implies $d_{TV}(D_1, D_2) \leq \delta$.*

For some special families of reward distributions, such q is known Diakonikolas and Kane (2023). (Let k be a constant in what follows)

- For the family of Gaussian distributions with standard deviation 1, $q(\delta) = k\delta$.
- For the family of log-concave distributions with standard deviation 1, $q(\delta) = k\delta \log(1/\delta)$.
- For the family of distributions with kth moment bounded by 1 for $k \geq 2$, $q(\delta) = k\delta^{1-1/k}$.

Assumption 4 expresses the conditional unimodality in likelihood functions in terms of rewards.

Assumption 4. *Given ground truth $f \in \mathcal{F}$ and $g^1, g^2 \in \mathcal{F}$, for every $\pi \in \Pi$, $\mathbb{E}_{g^1_\pi}[r|o] \leq \mathbb{E}_{g^2_\pi}[r|o] \leq \mathbb{E}_{f_\pi}[r|o]$ or $\mathbb{E}_{g^1_\pi}[r|o] \geq \mathbb{E}_{g^2_\pi}[r|o] \geq \mathbb{E}_{f_\pi}[r|o]$ implies $D_{KL}(f_\pi(\cdot \mid o), g^2_\pi(\cdot \mid o)) \leq D_{KL}(f_\pi(\cdot \mid o), g^1_\pi(\cdot \mid o))$ almost everywhere (a.e.).*

Assumptions 5 and 6 exclude trivial cases where the reward information is not at all needed for the inference of the ground-truth model f.

Assumption 5 (Density of $\mathcal{F}_\pi$ for every $\pi \in \Pi$)**.** *For every* $\pi \in \Pi$, $\{g_\pi \in \mathcal{F}_\pi \mid \mu_{g_\pi} \geq \mu_{f_{\pi_f}}\} \cap \{g_\pi \in \mathcal{F}_\pi \mid |\mathbb{E}_{f_\pi}(r \mid o) - \mathbb{E}_{g_\pi}(r \mid o)| \leq q(\delta)\ a.e.\}$ *is nonempty given* $\delta > 0$.

Intuitively, $\{g_\pi \in \mathcal{F}_\pi \mid \mu_{g_\pi} \geq \mu_{f_{\pi_f}}\}$ is the set of hypotheses in $\mathcal{F}_\pi$ we need to reject, and $\{g_\pi \in \mathcal{F}_\pi \mid |\mathbb{E}_{f_\pi}(r \mid o) - \mathbb{E}_{g_\pi}(r \mid o)| \leq q(\delta)\ a.e.\}$ is the set of hypothesis we cannot reject under contamination of outcomes from decision π. Note that $\{g_\pi \in \mathcal{F}_\pi \mid |\mathbb{E}_{f_\pi}(r \mid o) - \mathbb{E}_{g_\pi}(r \mid o)| \leq q(\delta)\ a.e.\} \subseteq \{g_\pi \in \mathcal{F}_\pi \mid |\mu_{g_\pi} - \mu_{f_\pi}| \leq q(\delta)\}$.

Assumption 6. *Let* g_π^o *be the marginal distribution of the observatiion for of* $g_\pi \in \mathcal{F}_\pi$. *There exists* $r_o > 0$ *such that for every* $\pi \in \Pi$, $|\mu_{f_\pi} - \mu_{g_\pi}| \leq r_o$ *implies* $f_\pi^o = g_\pi^o$ *a.e..*

Note that $f_\pi^o = g_\pi^o$ *a.e.* if and only if $D_{KL}(f_\pi^o \| g_\pi^o) = 0$ holds. If $D_{KL}(f_\pi^o \| g_\pi^o) > 0$, no information on the rewards will be required to reject g_π under f_π, the true hypothesis for the decision π. On the other hand, in the reinforcement learning problems where reward functions are parametrized independent of the transition model parameters, r_0 in the Assumption 6 is $+\infty$.

Appendix C. Proof of Theorem 2

C.1 Proof of Theorem 2

Recall that $P_{f,n,A}[\cdot]$ denotes the distribution of outcomes of algorithm A on the true model instance f by the round n. We further denote the marginal distribution of $P_{f,n,A}[\cdot]$ in terms of decision π's rewards and outcomes by $P^\pi_{f,n,A}[\cdot]$.

Lemma 3. *Suppose that the ground-truth model is* $f \in \mathcal{F}$. *Then a consistent algorithm must satisfy* $(1 + o(1)) \ln n \leq \sum_{\pi \in \Pi} D_{\mathrm{KL}} \left(P^\pi_{f,n,A} \| P^\pi_{g,n,A} \right)$ *for* $g \in \mathcal{F}(f)^c$.

Proof According to Dong and Ma (2022) Dong and Ma (2023), any consistent algorithm A must satisfy $(1 + o(1)) \ln n \leq D_{\mathrm{KL}} \left(P_{f,n,A} \| P_{g,n,A} \right)$ for $g \in \mathcal{F}(f)^c$. Since the terms involving A (the algorithm used to collect the data) cancel out and the outcomes of decisions are independent of each other, $\frac{P_{f,n,A}}{P_{g,n,A}} = \prod_{\pi \in \Pi} \frac{P^\pi_{f,n,A}}{P^\pi_{g,n,A}}$ holds. Therefore, the condition of Dong and Ma (2022) becomes $(1 + o(1)) \ln n \leq \sum_{\pi \in \Pi} D_{\mathrm{KL}} \left(P^\pi_{f,n,A} \| P^\pi_{g,n,A} \right)$ for $g \in \mathcal{F}(f)^c$.

Lemma 4. *For any* $\epsilon > 0$, ϵ*-contamination in the delay model of* $\pi^\star$ *makes the rewards of decisions* $\pi \neq \pi^\star$ *suffer* δ*-contamination for some* $\delta > 0$ *under consistency.*

The proof of Lemma 4 is deferred to Section C.3. Lemma 5 shows that, under δ-reward contaminations in reward distributions of all $\pi \neq \pi^\star$, choosing optimal decision $\pi^\star$ alone must be enough to satisfy the condition described in Lemma 3 and otherwise, we cannot satisfy it.

Lemma 5. *If the rewards of decisions* $\pi \in \Pi \setminus \pi_f$ *suffer* δ*-contamination for some* $\delta > 0$, *consistency of the algorithm requires* $(1 + o(1)) \ln n \leq D_{\mathrm{KL}} \left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n} \right)$ *for all* $g \in \mathcal{F}(f)^c$.

The proof of Lemma 5 is deferred to Section C.3.

The rest of the proof of Theorem 2 is immediate from the derivation of Dong and Ma (2023)'s Theorem 1, which is as follows: from the chain rule of divergence, $D_{\mathrm{KL}}\left(P^{\pi}_{f,n} \| P^{\pi}_{g,n}\right) = \mathbb{E}_{f,n}\left[N_\pi\right] D_{\mathrm{KL}}(f_\pi \| g_\pi)$ holds for $\pi \in \Pi$. Defining $w_\pi := \mathbb{E}_f\left[N_\pi\right] / ((1+o(1)) \ln n)$, Lemma 5 implies that $\mathcal{C}(f) = 0$ by the definition of $\mathcal{C}(f)$ in the equation (1). Since we don't know the ground truth f a priori, designing a learning system that assures the existence of a robust algorithm requires $\mathcal{C}(f) = 0$ for all $f \in \mathcal{F}$.

C.2 Proof of Lemma 4

Denote by $N_\pi^{[a,b]}$ the random variable that counts the number of decisions of π between rounds a and b. For consistency, for any small enough $p > 0$, for any $r > 0$, for some m, there must exist a constant $n_{r,p}$ such that for all intervals $[a, b]$ with $b - a \geq n_{r,p}$ and $a, b > m$, $E[N_{\pi^\star}^{[a,b]}] \geq (b-a) - (b-a)^p r$ holds. Recall that we denote by D_π the true delay model for the decision $\pi \in \Pi$, and by $\hat{D}_\pi$ the given model for D_π. Note that ϵ-contamination means we can arbitrarily choose $\{D_\pi\}_{\pi \in \Pi}$ as long as $d_{TV}(D_\pi, \hat{D}_\pi) \leq \epsilon$. Consider the case when $D_{\pi^\star} = \hat{D_{\pi^\star}} + \epsilon D_a - \epsilon D_c$ where $P(D_a = k) = \frac{1}{n_{r,p}}$ for $0 \leq k \leq n_{r,p} - 1$ and 0 for elsewhere, and D_c is an arbitrary distribution. For $\pi \in \Pi \setminus \pi^\star$, consider $D_\pi = \hat{D}_\pi$. Then for $k \geq \max(n_{r,p}, m)$,

$$
\begin{aligned}
&P(\{\text{A reward arrival at } k \text{ is not from } \pi^\star\}) \\
&\quad = \frac{\sum_{i=1}^{k} P(\{\pi_i\text{'s reward arrives at } k \text{ and } \pi_i \neq \pi^\star\})}{\sum_{i=1}^{k} P(\{\pi_i\text{'s reward arrives at } k\})} \\
&\quad \leq \frac{|\Pi| - 1}{|\Pi| - 1 + \sum_{i=1}^{k} P(\{\pi_i\text{'s reward arrives at } k \text{ and } \pi_i = \pi^\star\})} \qquad (2) \\
&\quad \leq \frac{|\Pi| - 1}{|\Pi| - 1 + \frac{\epsilon}{n_{r,p}} \sum_{i=k-n_{r,p}}^{k} \mathbb{E}[1_{\pi_i = \pi^\star}]} \\
&\quad \leq \frac{|\Pi| - 1}{|\Pi| - 1 + \frac{\epsilon}{n_{r,p}}(n_{r,p} - n_{r,p}^{p} r)} = \frac{|\Pi| - 1}{|\Pi| - 1 + \epsilon(1 - n_{r,p}^{p-1} r)}
\end{aligned}
$$

where the inequality in the equation (2) follows from the fact that the delay distribution of each decision sums to one. Therefore, $P(\{\text{A reward arrival at } k \text{ is from } \pi^\star\}) > \delta := \frac{\epsilon(1 - n_{r,p}^{p-1} r)}{|\Pi| - 1 + \epsilon(1 - n_{r,p}^{p-1} r)}$.

Denote the rewards distributions associated with $D_{\pi^\star}, \hat{D_{\pi^\star}}, D_a$, and D_c as $R_{\pi^\star}, \hat{R_{\pi^\star}}, R_a$, and R_c each. Then $R_{\pi^\star} = \hat{R_{\pi^\star}} + \epsilon R_a - \epsilon R_c$ must hold, where the mean of $R_{\pi^\star}$ and $R_{\pi^\star}$ are supposed to be the same. Since the choice of R_a can be arbitrary by choosing R_c accordingly, we can conclude that the reward distributions of decisions $\pi \neq \pi^\star$ indeed suffer δ-contamination.

C.3 Proof of Lemma 5

Recall that we use f_π to refer to an element of $\mathcal{F}_\pi$, while f_π also denotes the π-coordinate of some $f \in \mathcal{F}$. Suppose that the ground-truth model is $f \in \mathcal{F}$. For $g \in \mathcal{F}(f)^c$, a consistent algorithm must satisfy

$$(1+o(1))\ln n \le D_{\mathrm{KL}}\left(P_{f,n} \| P_{g,n}\right) = \sum_{\pi \in \Pi} D_{\mathrm{KL}}\left(P^\pi_{f,n} \| P^\pi_{g,n}\right). \tag{3}$$

$$= D_{\mathrm{KL}}\left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n}\right) + \sum_{\pi \in \Pi \setminus \pi_f} \mathbb{E}_{f,n}\left[N_\pi\right] D_{\mathrm{KL}}(f_\pi(r,o) \| g_\pi(r,o)) \tag{4}$$

$$= D_{\mathrm{KL}}\left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n}\right) + \sum_{\pi \in \Pi \setminus \pi_f} \mathbb{E}_{f,n}\left[N_\pi\right] \left(D_{\mathrm{KL}}(f^o_\pi(o) \| g^o_\pi(o)) + \mathbb{E}_{f_\pi}\left[\log \frac{f_\pi(r|o)}{g_\pi(r|o)}\right]\right) \tag{5}$$

Above,

- The inequality in the equation (3) is from Dong and Ma (2022) Dong and Ma (2023).
- The equality in the equation (3) follows from the fact that algorithm-related terms cancel out.
- The inequality in the equation (4) is from the Divergence decomposition Lemma Lattimore and Szepesvári (2020)
- The inequality in the equation (5) follows from the chain rule of KL divergence Cover (1999).

Let $\min(q(\delta), r_o) = q'(\delta)$, where r_o is defined as in Assumption 6. By Assumption 5, for every $\pi \neq \pi_f$, there exists a non-empty set $\mathcal{E}_\pi := \{l_\pi \in \mathcal{F}_\pi \mid |E_{f_\pi}[r|o] - E_{l_\pi}[r|o]| \le q'(\delta) \ a.e.\} \cap \{\mu_{l_\pi} \ge \mu_{f_{\pi_f}}\}$. Hense we can construct $\mathcal{E}(\pi) := \{l \in \mathcal{F} \mid l_\pi \in \mathcal{E}_\pi, l_{\pi'} = f_{\pi'} \text{ for } \pi' \neq \pi\}$. Note that $\mathcal{E}(\pi) \subseteq \mathcal{F}(f)^c := \{g \in \mathcal{F} \mid \pi_g \neq \pi_f\} = \{g \in \mathcal{F} \mid \exists \pi \in \Pi \ s.t. \ \mu_{g_\pi} \ge \mu_{f_{\pi_f}}\}$. Therefore, for every $\pi \neq \pi_f$,

$$(1+o(1))\ln n \le D_{\mathrm{KL}}\left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n}\right) + \sum_{\pi \in \Pi \setminus \pi_f} \mathbb{E}_{f,n}\left[N_\pi(n)\right] \mathbb{E}_{f_\pi}\left[\log \frac{f_\pi(r|o)}{g_\pi(r|o)}\right] \text{ for } g \in \mathcal{E}(\pi) \tag{6}$$

$$(\Rightarrow)\ (1+o(1))\ln n \le D_{\mathrm{KL}}\left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n}\right) + \mathbb{E}_{f,n}\left[N_\pi(n)\right] \mathbb{E}_{f_\pi}\left[\log \frac{f_\pi(r|o)}{g_\pi(r|o)}\right] \text{ for } g \in \mathcal{E}(\pi) \tag{7}$$

$$(\Rightarrow)\ (1+o(1))\ln n \le D_{\mathrm{KL}}\left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n}\right) \text{ for } g \in \mathcal{E}(\pi) \tag{8}$$

$$(\Rightarrow)\ (1+o(1))\ln n \le D_{\mathrm{KL}}\left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n}\right) \text{ for } g \in \mathcal{E}'(\pi) \tag{9}$$

$$(\Rightarrow)\ (1+o(1))\ln n \le D_{\mathrm{KL}}\left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n}\right) \text{ for } g \in \{g \in \mathcal{F} \mid \mu_{g_\pi} \ge \mu_{f_{\pi_f}}\}. \tag{10}$$

where $\mathcal{E}'_\pi := \{l_\pi \in \mathcal{F}_\pi \mid \mu_{f_{\pi_f}} \le \mu_{l_\pi}\}$ and $\mathcal{E}'(\pi) := \{l \in \mathcal{F} \mid l_\pi \in \mathcal{E}'_\pi, l_{\pi'} = f_{\pi'} \text{ for } \pi' \neq \pi\}$. Above,

- Equation (6) follows from Assumption 6 and equation (5).
- The logical implication in equation (7) follows from the definition of $\mathcal{E}(\pi)$.
- The logical implication in equation (8) follows from Assumptions 3, 4 and 5: $|E_{f_\pi}[r|o] - E_{g_\pi}[r|o]| \leq q'(\delta) \leq q(\delta)$ *a.e.* implies $d_{TV}(f_\pi(\cdot \mid o), g_\pi(\cdot \mid o)) < \delta$ *a.e.* from Assumption 3; therefore, under some δ-contamination of $f_\pi(\cdot \mid o)$, the contaminated $E_{f_\pi}[r|o]$ can be farther from the true $E_{f_\pi}[r|o]$ than $E_{g_\pi}[r|o]$. Therefore, $D_{\mathrm{KL}}\left(f_\pi(\cdot \mid o) \parallel g_\pi(\cdot \mid o)\right) \leq 0$ *a.e.* due to Assumption 4, and so $\mathbb{E}_{f_\pi}\left[\log \frac{f_\pi(r|o)}{g_\pi(r|o)}\right] = \mathbb{E}_{f_\pi^o}\left[\mathbb{E}_{f_\pi(r|o)}\left[\log \frac{f_\pi(r|o)}{g_\pi(r|o)}\right]\right] = \mathbb{E}_{f_\pi^o}\left[D_{\mathrm{KL}}\left(f_\pi(\cdot \mid o) \parallel g_\pi(\cdot \mid o)\right)\right] \leq 0$.
- The logical implication in equation (9) follows from Assumption 4: Define $\hat{\mathcal{E}}_\pi := \{|\mu_{f_\pi} - \mu_{g_\pi}| \leq q'(\delta)\} \cap \{\mu_{g_\pi} \geq \mu_{f_{\mu_f}}\}$ and $\hat{\mathcal{E}}(\pi) := \{l \in \mathcal{F} \mid l_\pi \in \hat{\mathcal{E}}_\pi, l_{\pi'} = f_{\pi'} \text{ for } \pi' \neq \pi\}$. Note that $\hat{\mathcal{E}}(\pi) \subseteq \mathcal{E}(\pi)$. Then for $g' \in \mathcal{E}'(\pi) \setminus \hat{\mathcal{E}}(\pi)$ and $g \in \hat{\mathcal{E}}(\pi)$ with $(\mu_{g_\pi} - \mu_{f_\pi})(\mu_{g'_\pi} - \mu_{f_\pi}) \geq 0$, $D_{\mathrm{KL}}\left(P^\pi_{f,n} \| P^\pi_{g,n}\right) \leq D_{\mathrm{KL}}\left(P^\pi_{f,n} \| P^\pi_{g',n}\right)$ due to the monotonicity assumption of Assumption 4.
- The logical implication in equation (10) follows from the fact that any element in $\{g \in \mathcal{F} \mid \mu_{g_\pi} \geq \mu_{f_{\pi_f}}\}$ has an element in $\mathcal{E}'$ that is strictly closer to f.

Since $\mathcal{F}(f)^c := \{g \in \mathcal{F} \mid \pi_g \neq \pi_f\} = \{g \in \mathcal{F} \mid \exists \pi \in \Pi \; s.t. \; \mu_{g_\pi} \geq \mu_{f_{\pi_f}}\}$, we immediately get $(1 + o(1)) \ln n \leq D_{\mathrm{KL}}\left(P^{\pi_f}_{f,n} \| P^{\pi_f}_{g,n}\right)$ for $g \in \mathcal{F}(f)^c$.

Appendix D. Max-contamination

Here we define the concept of max-contamination. Whatever true delay distribution the reward delays follow, the maximum number of reward arrivals from π by the round k is $N_\pi(k)$, the total number of π decisions by the round k. Then the max-contamination of the decision $\pi' \in \Pi$ at round k is defined as $\delta^{\max}_{\pi'}(k) := \min(\frac{\sum_{\pi \in \Pi \setminus \pi'} N_\pi(k)}{\widetilde{N}(k)}, 1)$, where $\widetilde{N}(k)$ stands for the total number of reward arrivals by round k. Note that the contamination of reward arrival at k is bounded by the max-contamination $\delta^{\max}_{\pi'}(k)$, as the delay distributions of decisions are stationary, i.e., they do not change over time.

Let $P^c_{g,k,\overline{\pi}}$ indicates the likelihood of $g \in \mathcal{F}$ that is computed as if all reward arrivals by the round k are from the decision $\overline{\pi}$. (We add the superscript c since this is not true, as we allow decision transitions in Algorithm 1.) Note that $\ln \frac{P^c_{f,k,\overline{\pi}}}{P^c_{g,k,\overline{\pi}}} = \sum_{k=1}^n \ln \frac{f^c_\pi(k)}{g^c_\pi(k)}$, where $f^c_\pi(k)$ and $g^c_\pi(k)$ are likelihood of each data assuming that the data is from π.

Appendix E. Proof of Theorem 3

Lemma 6 shows that it takes finite time in expectation to transition to correct instances.

Lemma 6. *After Algorithm 1's $\widehat{f}$ transitions to $g \in \mathcal{F}(f^\star)^c$, it converges within a finite expected number of rounds to $\widehat{f} = f$.*

Proof Let f be the ground truth model. After each transition to $g \in \mathcal{F}(f)^c$,

$$
\begin{aligned}
&P(\{\sum_{k=1}^{n} \ln \frac{f^c_{\pi_g}(k)}{g^c_{\pi_g}(k)} \leq 2\ln n + \sum_{k=1}^{n} \frac{2}{\sqrt{\beta}} \delta^{\max}_{\pi_g}(k)\}) \\
&\quad = P(\{\sum_{k=1}^{n} (\ln \frac{f_{\pi_g}}{g_{\pi_g}} + \ln \frac{f^c_{\pi_g}(k)}{f_{\pi_g}} + \ln \frac{g_{\pi_g}}{g^c_{\pi_g}(k)}) \leq 2\ln n + \sum_{k=1}^{n} \frac{2}{\sqrt{\beta}} \delta^{\max}_{\pi_g}(k)\}) \\
&\quad \leq P(\{\sum_{k=1}^{n} (\ln \frac{f_{\pi_g}}{g_{\pi_g}} + \ln \frac{f^c_{\pi_g}(k)}{f_{\pi_g}} + \ln \frac{g_{\pi_g}}{g^c_{\pi_g}(k)}) \leq 2\ln n + \frac{2C}{\sqrt{\beta}} \ln n\}) \text{ for some } C \qquad (11) \\
&\quad \leq P(\{\sum_{k=1}^{n} \left(\ln \frac{f_{\pi_g}}{g_{\pi_g}} - D_{KL}(f_{\pi_g}, g_{\pi_g})\right) + \sum_{k=1}^{n} \left(\ln \frac{f^c_{\pi_g}(k)}{f_{\pi_g}}\right) + \sum_{k=1}^{n} \left(\ln \frac{g_{\pi_g}}{g^c_{\pi_g}(k)}\right) \\
&\qquad \leq -\ln n\}) \text{ for } n \geq n_0 \text{ for some } n_0 < \infty \qquad (12) \\
&\quad \leq P(\{\sum_{k=1}^{n} \left(\ln \frac{f_{\pi_g}}{g_{\pi_g}} - D_{KL}(f_{\pi_g}, g_{\pi_g})\right) + \sum_{k=1}^{n} \left(\ln \frac{f^c_{\pi_g}(k)}{f_{\pi_g}} - D_{KL}(f_{\pi_g}, f^c_{\pi_g}(k))\right) \\
&\qquad + \sum_{k=1}^{n} \left(\ln \frac{g_{\pi_g}}{g^c_{\pi_g}(k)} - D_{KL}(g_{\pi_g}, g^c_{\pi_g}(k))\right) \leq -\ln n\}) \text{ for } n \geq n_0 \qquad (13) \\
&\quad \leq P(\{\sum_{k=1}^{n} \left(\ln \frac{f_{\pi_g}}{g_{\pi_g}} - D_{KL}(f_{\pi_g}, g_{\pi_g})\right) \geq -\ln n, \sum_{k=1}^{n} \left(\ln \frac{f^c_{\pi_g}(k)}{f_{\pi_g}} - D_{KL}(f_{\pi_g}, f^c_{\pi_g}(k))\right) \\
&\qquad \geq -\ln n, \sum_{k=1}^{n} \left(\ln \frac{g_{\pi_g}}{g^c_{\pi_g}(k)} - D_{KL}(g_{\pi_g}, g^c_{\pi_g}(k))\right) \geq -\ln n\}^{\mathbf{c}}) \text{ for } n \geq n_0 \\
&\quad = P(\{\sum_{k=1}^{n} \left(\ln \frac{f_{\pi_g}}{g_{\pi_g}} - D_{KL}(f_{\pi_g}, g_{\pi_g})\right) \leq -\ln n\} \cup \{\sum_{k=1}^{n} \left(\ln \frac{f^c_{\pi_g}(k)}{f_{\pi_g}} - D_{KL}(f_{\pi_g}, f^c_{\pi_g}(k))\right) \\
&\qquad \leq -\ln n\} \cup \{\sum_{k=1}^{n} \left(\ln \frac{g_{\pi_g}}{g^c_{\pi_g}(k)} - D_{KL}(g_{\pi_g}, g^c_{\pi_g}(k))\right) \leq -\ln n\}) \text{ for } n \geq n_0 \\
&\quad \leq 3O(\frac{1}{n^2}) = O(\frac{1}{n^2}). \qquad (14)
\end{aligned}
$$

Above,

- Equation 11 follows from the fact that $\delta^{\max}_{\pi_g}(k)$ decreases with the rate $1/n$.
- Equation 12 follows from the fact that $nD_{KL}(f_{\pi_g}, g_{\pi_g}) = D_{KL}(P_{f,n,\overline{\pi_g}} \| P_{g,n,\overline{\pi_g}}) = \omega(\ln n)$ from the Assumption 1.

- Equation 13 follows from the fact that substracting positive value on the left does not change the inequality.
- Equation 14 follows from the fact that the log-likelihood ratios are bounded due to Assumption 2, and thus sub-gaussian random variables.

Therefore, after each time a bad transition to $g \in \mathcal{F}(f)^c$ happens, the event $\{\sum_{k=1}^n \ln \frac{f^c_{\pi_g}(k)}{g^c_{\pi_g}(k)} \leq 2\ln n + \sum_{k=1}^n \frac{2}{\sqrt{\beta}}\delta^{\max}_{\pi_g}(k)\}$ happens only finite many times in expectation by the Borel-Cantelli lemma, which implies that the inference will arrive at the correct instance within finite expected time.

Lemma 7 shows that the total number of wrong transitions from the correct inferences is finite in expectation.

Lemma 7. *Under Algorithm 1, the number of rounds at which event* $\{\widehat{f} = f^\star\} \cap \{\exists g \in \mathcal{F}(f^\star)^c \text{ s.t. } \sum_{k=1}^n \ln \frac{g^c_{\pi_{\hat{f}}}(k)}{\hat{f}^c_{\pi_{\hat{f}}}(k)} \geq 2\ln k + \sum_{k=1}^n \frac{2}{\sqrt{\beta}}\delta^{\max}_{\pi_{\hat{f}}}(k)\}$ *holds is finite in expectation.*

Proof For any $g \in \mathcal{F}(f)^c$, when f is the ground truth model,

$$
\begin{aligned}
&P(\{\sum_{k=1}^n \ln \frac{g^c_{\pi_g}(k)}{f^c_{\pi_g}(k)} \geq 2\ln n + \sum_{k=1}^n \frac{2}{\sqrt{\beta}}\delta^{\max}_{\pi_g}(k)\}) \\
&\quad = P(\{\sum_{k=1}^n (\ln \frac{g_{\pi_g}}{f_{\pi_g}} + \ln \frac{f_{\pi_g}}{f^c_{\pi_g}(k)} + \ln \frac{g^c_{\pi_g}(k)}{g_{\pi_g}}) \geq 2\ln n + \sum_{k=1}^n \frac{2}{\sqrt{\beta}}\delta^{\max}_{\pi_g}(k)\}) \\
&\quad \leq P(\{\sum_{k=1}^n \left(\ln \frac{g_{\pi_g}}{f_{\pi_g}}\right) + \sum_{k=1}^n \left(\ln \frac{f_{\pi_g}}{f^c_{\pi_g}(k)} - D_{KL}(f_{\pi_g}, f^c_{\pi_g}(k))\right) \\
&\qquad + \sum_{k=1}^n \left(\ln \frac{g^c_{\pi_g}(k)}{g_{\pi_g}} - D_{KL}(g^c_{\pi_g}(k), g_{\pi_g})\right) \geq 2\ln n\}) \qquad (15) \\
&\quad \leq P(\{\sum_{k=1}^n \left(\ln \frac{g_{\pi_g}}{f_{\pi_g}}\right) \leq 2\ln n, \sum_{k=1}^n \left(\ln \frac{f_{\pi_g}}{f^c_{\pi_g}(k)} - D_{KL}(f_{\pi_g}, f^c_{\pi_g}(k))\right) \leq 2\ln n, \\
&\qquad , \sum_{k=1}^n \left(\ln \frac{g^c_{\pi_g}(k)}{g_{\pi_g}} - D_{KL}(g^c_{\pi_g}(k), g_{\pi_g})\right) \leq 2\ln n\}^{\mathbf{c}}) \\
&\quad \leq P(\{\sum_{k=1}^n \left(\ln \frac{g_{\pi_g}}{f_{\pi_g}}\right) \geq 2\ln n\} \cup \{\sum_{k=1}^n \left(\ln \frac{f_{\pi_g}}{f^c_{\pi_g}(k)} - D_{KL}(f_{\pi_g}, f^c_{\pi_g}(k))\right) \geq 2\ln n\} \\
&\qquad , \cup \{\sum_{k=1}^n \left(\ln \frac{g^c_{\pi_g}(k)}{g_{\pi_g}} - D_{KL}(g^c_{\pi_g}(k), g_{\pi_g})\right) \geq 2\ln n\})
\end{aligned}
$$

$$\leq 3O(\frac{1}{n^2}) = O(\frac{1}{n^2}). \tag{16}$$

Above,

- Equation 15 follows from the reverse Pinsker's inequality (Verdú, 2014) (total variation distance smaller than δ implies KL divergence smaller than $\frac{1}{\sqrt{\beta}}\delta$)
- Equation 16 follows from Lemma 4.3 of Dong and Ma (2022) Dong and Ma (2023), which says that $P_Q\left(\left\{\sum_{i=1}^{m} \ln \frac{P_i}{Q_i} \geq c\right\}\right) \leq \exp(-c)$, and the fact that the log-likelihood ratios are bounded due to Assumption 2, and thus sub-gaussian random variables.

Therefore, the event holds in total only for finite rounds of k in expectation by the Borel-Cantelli lemma.

Combining Lemmas 6 and 7, we can conclude that $\widehat{f} \notin \mathcal{F}(f^\star)$ holds only for a finite number of rounds in expectation. That is, regret is bounded in expectation.

Appendix F. Details of Theorem 4 (Linear contextual bandit case)

F.1 Setting of Hao et al. (2020)

Hao et al. (2020) was the first to characterize the condition for achieving bounded regret in linear contextual bandit problem. Consider the stochastic K-armed contextual linear bandit problem with a horizon of n rounds and a finite A-size set of k-dimensional possible contexts $\mathcal{X} = \{\mathbf{x}_j\}_{j\in[A]}$. At each round, a context is chosen according to the unknown distribution p over $\mathcal{X}$. When the sampled context is $\mathbf{x}_j$ and its chosen arm is m, we receive $\phi_m(\mathbf{x}_j)'\theta + \epsilon$, where $\{\phi_m : \mathbb{R}^k \mapsto \mathbb{R}^d\}_{m\in[M]}$ are linear representation functions that are assumed to be *precisely known*, θ is a parameter vector of dimension d that is shared across the arms, and ϵ is an i.i.d. random noise that follows a sub-Gaussian distribution with variance proxy σ^2.

Theorem 6 (Hao et al. (2020); Papini et al. (2021b))**.** *Let $m_{j\theta}$ be an optimal arm for context $j \in [A]$, when the true parameter is θ, i.e., $m_{j\theta} \in \operatorname{argmax}_{m\in[M]} \phi_m(\mathbf{x}_j)'\theta$. Given linear contextual bandit setting described above, bounded regret can be achieved if and only if $\left\{\phi_{m_{j\theta}}(x_j) \mid j \in A\right\}$ spans $\mathbb{R}^d$.*

F.2 Proof of Theorem 4

Let Θ be the set of all parameters, and let $\theta^\star \in \Theta$ be the unknown true parameter. Suppose that $\mathcal{C}(\theta) = 0$ for $\theta \in \Theta$. By Theorem 1 and Theorem 6, $\left\{\phi_{m_{j\theta}}(x_j) \mid j \in A\right\}$ spans $\mathbb{R}^d$ for $\theta \in \Theta$. Denote $T_{\mathbf{x}}(n)$ be the number of arrivals of context $\mathbf{x} \in \mathcal{X}$.

Then for any $\tilde{\theta} \in \Theta \setminus \{\theta^\star\}$,

$$D_{\mathrm{KL}}\left(P_{\theta^\star,n,\overline{\pi_\theta}} \| P_{\widetilde{\theta},n,\overline{\pi_\theta}}\right) = \frac{1}{2}\sum_{x\in\mathcal{A}} \mathbb{E}\left[T_x(n)\right]\langle x, \theta^\star - \widetilde{\theta}\rangle^2 \tag{17}$$

$$= \frac{1}{2}(\theta^\star - \widetilde{\theta})^\top \mathbb{E}\left[\sum_{x\in\mathcal{A}} T_x(n) x x^\top\right](\theta^\star - \widetilde{\theta})$$

$$= \frac{1}{2}(\theta^\star - \widetilde{\theta})^\top n\mathbb{E}\left[\sum_{x\in\mathcal{A}} \frac{T_x(n)}{n} x x^\top\right](\theta^\star - \widetilde{\theta})$$

$$\geq \frac{1}{2}\|\theta^\star - \widetilde{\theta}\|^2 n\lambda_{\min} \tag{18}$$

$$= \Omega(n) \tag{19}$$

Above,

- The equality in equation (17) is from the divergence decomposition lemma Lattimore and Szepesvári (2020)
- $\lambda_{\min}$ of equation (18) denotes the smallest eigenvalue for $\mathbb{E}_{\mathbf{x}_j\sim p}\left[\phi_{m_j\theta}(\mathbf{x}_j)\phi_{m_j\theta}(\mathbf{x}_j)^\top\right]$
- The inequality of equation (18) is from the fact that $\frac{x^T A x}{x^T x}$ is larger than the smallest eigenvalue of A.
- The equality of equation (19) comes from the fact that $\lambda_{\min} > 0$ is equivalent to $\left\{\phi_{m_j\theta}(x_j) \mid j \in A\right\}$ spanning $\mathbb{R}^d$ (Papini et al., 2021b).

Appendix G. Details of Theorem 5 (RL with Linear MDP case)

G.1 Setting of Papini et al. (2021a)

Papini et al. (2021a) characterized the condition for achieving bounded regret in reinforcement learning with Linear MDP. Consider a Linear MDP with a horizon of H, while we are given total n episodes. It has finite A-size state space with k-dimensional covariates $\mathcal{S} = \{\mathbf{s}_j\}_{j\in[A]}$. All states share K-size action space. At each episode $l \leq n$, a policy π_l is chosen. When state is $\mathbf{x}_j$ and its chosen action at $h \in [H]$ is m, The state-action function $Q_h(s,a)$ equals $\phi_h(\mathbf{x}_j, m)'\theta_h$, where $\{\phi_h : \mathbb{R}^k \mapsto \mathbb{R}^d\}_{h\in[H]}$ are linear representation functions that are assumed to be *precisely known*, θ_h is a parameter vector of dimension d that is shared across different episodes.

Theorem 7 (Papini et al. (2021a)). *Suppose that the MDP satisfies Bellman closure (Zanette et al., 2020) or Low-rank MDP assumption (Jin et al., 2020). Define the optimal policy as $\pi^\star$ and $\phi_h^\star(s) := \phi_h\left(s, \pi^\star(s,h)\right)$. Then, the condition that*
$\operatorname{span}\left\{\phi_h^\star(s) \mid \forall s, \textit{optimal policy visits } s \textit{ at } h \textit{ with positive probability}\right\} = \mathbb{R}^d$ *is sufficient for achieving bounded regret.*

G.2 Proof of Theorem 5

It is straightforward that the proof of Theorem 5 is almost equivalent to the proof of Theorem 4, except that we infer θ_h for each $h \in [H]$.