Proceedings of Machine Learning Research vol 167:1–54, 2022 33rd International Conference on Algorithmic Learning Theory

A Model Selection Approach for Corruption Robust Reinforcement Learning

Chen-Yu Wei* CHENYU.WEI@USC.EDU
University of Southern California

Christoph Dann CDANN@CDANN.NET
Google Research

Julian Zimmert ZIMMERT@GOOGLE.COM
Google Research

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

We develop a model selection approach to tackle reinforcement learning with adversarial corruption in both transition and reward. For finite-horizon tabular MDPs, without prior knowledge on the total amount of corruption, our algorithm achieves a regret bound of $\widetilde{\mathcal{O}}\big(\min\{\frac{1}{\Delta}, \sqrt{T}\} + C\big)$ where T is the number of episodes, C is the total amount of corruption, and Δ is the reward gap between the best and the second-best policy. This is the first worst-case optimal bound achieved without knowledge of C, improving previous results of Lykouris et al. (2021); Chen et al. (2021b); Wu et al. (2021). For finite-horizon linear MDPs, we develop a computationally efficient algorithm with a regret bound of $\widetilde{\mathcal{O}}(\sqrt{(1+C)T})$, and another computationally inefficient one with $\widetilde{\mathcal{O}}(\sqrt{T}+C)$, improving the result of Lykouris et al. (2021) and answering an open question by Zhang et al. (2021b). Finally, our model selection framework can be easily applied to other settings including linear bandits, linear contextual bandits, and MDPs with general function approximation, leading to several improved or new results.

1. Introduction

Reinforcement learning (RL) studies how an agent learns to behave in an unknown environment with reward feedback. The environment is often modeled as a Markov decision process (MDP). In the standard setting, the MDP is assumed to be static, i.e., the state transition kernel and the instantaneous reward function remain fixed over time. Under this assumption, numerous computationally and statistically efficient algorithms with strong theoretical guarantees have been developed (Jaksch et al., 2010; Lattimore and Hutter, 2012; Dann and Brunskill, 2015; Azar et al., 2017; Jin et al., 2018, 2020b). However, these guarantees might break completely if the transition or the reward is corrupted by an adversary, even if the corruption is limited to a small fraction of rounds.

To model adversarial corruptions in MDPs, a framework called *adversarial MDP* has been extensively studied. In adversarial MDPs, the adversary is allowed to choose the reward function arbitrarily in every round, while keeping the transition kernel fixed (Neu et al., 2010b,a; Dick et al., 2014; Rosenberg and Mansour, 2019, 2021; Jin et al., 2020a; Neu and Olkhovskaya, 2020; Lee et al., 2020; Chen and Luo, 2021; He et al., 2021; Luo et al., 2021). Under this framework, strong sublinear regret bounds can be established, which almost match the bounds for the fixed reward case. Notably, Jin and Luo (2020); Jin et al. (2021b) developed algorithms that achieve near-minimax

* Research conducted when the author was an intern at Google Research.

regret bound in the adversarial reward case, while preserving refined instance-dependent bounds in the static case, showing that adversarial reward can be handled almost without price.

The situation becomes very different when the transition kernel can also be corrupted. It is first shown by Abbasi-Yadkori et al. (2013) that achieving sub-linear regret in this setting is *computationally hard*, and recently enhanced by Tian et al. (2021) showing that it is even *information-theoretically hard*. To establish meaningful guarantees, previous work aims to achieve a regret bound that smoothly degrades with the amount of corruption, and thus the learner can still behave well when only a small fraction of data is corrupted. When the total amount corruption is given as a prior knowledge to the learner, Wu et al. (2021) designed an algorithm with a regret upper bound that scales optimally with the amount of corruption. However, this kind of prior knowledge is rarely available in practice. When the total amount of corruption is unknown, efforts were made by Lykouris et al. (2021); Chen et al. (2021b); Cheung et al. (2020); Wei and Luo (2021); Zhang et al. (2021b) to obtain similar guarantees. Unfortunately, all their bounds scale sub-optimally in the amount of corruption. Therefore, the following question remains open: *When both reward and transition can be corrupted, how can the learner achieve a regret bound that has optimal dependence on the unknown amount of corruption?* We address this open problem by designing an efficient algorithm with the desired worst-case optimal bound. Specifically, in tabular MDPs, our regret bound scales with $\sqrt{T} + C$ where T is the number of rounds and C is the total amount of corruption (omitting dependencies on other quantities). This matches the lower bound by Wu et al. (2021). In contrast, the bounds obtained by Lykouris et al. (2021) and Chen et al. (2021b) are $(1 + C)\sqrt{T} + C^2$ and $\sqrt{T} + C^2$ respectively, which are non-vacuous only when $C \leq \sqrt{T}$, a rather limited case. The bounds obtained by Cheung et al. (2020); Wei and Luo (2021); Zhang et al. (2021a) are $(1 + C)^{1/4}T^{3/4}$, $\sqrt{T} + C^{1/3}T^{2/3}$, and $\sqrt{(1 + C)T}$ respectively. Although these bounds are meaningful for all $C \leq T$, the dependence on C is *multiplicative* to T, which is undesirable.

For tabular MDPs, we further show that the bound can be improved to $\min\{\frac{1}{\Delta}, \sqrt{T}\} + C$, where Δ is the gap between the expected reward of the best and the second-best policies. This kind of refined instance-dependent regret bound is also established by Lykouris et al. (2021) and Chen et al. (2021b). The bound of Lykouris et al. (2021) is $(1 + C)\min\{\mathrm{G}, \sqrt{T}\} + C^2$ for some gap-complexity $\mathrm{G} \leq \frac{1}{\Delta}$, while Chen et al. (2021b) obtained $\min\{\frac{1}{\Delta}, \sqrt{T}\} + C^2$. It is left as an open question whether the best-of-all-world bound $\min\{\mathrm{G}, \sqrt{T}\} + C$ is achievable.

Our method is based on the framework of *model selection* (Agarwal et al., 2017; Foster et al., 2019; Arora et al., 2021; Abbasi-Yadkori et al., 2020; Pacchiano et al., 2020a,b). In model selection problems, the learner is given a set of *base algorithms*, each with an underlying model or assumption for the world. However, the learner does not know in advance which model fits the real world the best. The goal of the learner is to be comparable to the best base algorithm in hindsight. In our case, a model of the world corresponds to a hypothetical amount of corruption C; for a given C, there are algorithms with near-optimal bounds (e.g., Wu et al. (2021)) that can serve as base algorithms. Therefore, the problem of handling unknown corruption can be cast as a model selection problem. To get the bound of $\sqrt{T} + C$, we adopt the idea of *regret balancing* similar to those of Abbasi-Yadkori et al. (2020); Pacchiano et al. (2020a), while to get $\min\{\frac{1}{\Delta}, \sqrt{T}\} + C$, we develop another novel *two-model selection* algorithm to achieve the goal (see Section 5).

Extensions to linear and general function approximation Our model selection framework can be readily extended to the cases of linear contextual bandits and linear MDPs. However, in Appendix A, we demonstrate that even with at most C corrupted rounds, a straightforward extension

Table 1: * indicates computationally inefficient algorithms. G is the GapComplexity defined in (Simchowitz and Jamieson, 2019); Δ is the gap between the expected reward of the best and second-best policy. It holds that $\mathrm{G} \leq \frac{1}{\Delta}$. $C^{\mathsf{a}} = \sum_t c_t$ and $C^{\mathsf{r}} = \sqrt{T\sum_t c_t^2}$, where c_t is the amount of corruption in round t. By definition, $C^{\mathsf{a}} \leq C^{\mathsf{r}} \leq \min\{\sqrt{C^{\mathsf{a}}T}, T\max_t c_t\}$. C^{a} is the standard notion of corruption in the literature. A refined version of this table with other dependencies explicitly written out is in Table 2 in Appendix G.
†: The bound reported in (Jin et al., 2021b) is $\min\{\mathrm{G} + \sqrt{\mathrm{G}C^{\mathsf{a}}}, \sqrt{T}\}$ under a different definition of regret.
$^\sharp$: Linearized corruption restricts that the corruption on action a equals to $c^\top a$ for some vector c shared among all actions.

Setting	Algorithm	$\mathrm{Reg}(T)$ in $\widetilde{\mathcal{O}}(\cdot)$	Restrictions
Tabular MDP	(Lykouris et al., 2021)	$(1 + C^{\mathsf{a}})\min\{\mathrm{G}, \sqrt{T}\} + (C^{\mathsf{a}})^2$	
	(Chen et al., 2021b)	$\min\{\frac{1}{\Delta}, \sqrt{T}\} + (C^{\mathsf{a}})^2$	$*$
	(Jin et al., 2021b)†	$\min\{\mathrm{G}, \sqrt{T}\} + C^{\mathsf{a}}$	only for corruption in reward
	G-COBE + UCBVI	$\min\{\frac{1}{\Delta}, \sqrt{T}\} + C^{\mathsf{a}}$	
Linear bandit	(Li et al., 2019)	$\frac{1}{\Delta^2} + \frac{C^{\mathsf{a}}}{\Delta}$	
	(Bogunovic et al., 2020)	$(1 + C^{\mathsf{a}})\sqrt{T}$	
	(Bogunovic et al., 2021)	$\sqrt{T} + (C^{\mathsf{a}})^2$	
	(Lee et al., 2021)	$\min\{\frac{1}{\Delta}, \sqrt{T}\} + C^{\mathsf{a}}$	only for linearized corruption$^\sharp$
	G-COBE + PE	$\min\{\frac{1}{\Delta}, \sqrt{T}\} + C^{\mathsf{a}}$	
Linear contextual bandit	(Foster et al., 2020)	$\sqrt{T} + C^{\mathsf{r}}$	
	COBE + OFUL	$\sqrt{T} + C^{\mathsf{r}}$	
	COBE + VOFUL	$\sqrt{T} + C^{\mathsf{a}}$	$*$
Linear MDP	(Lykouris et al., 2021)	$\sqrt{T} + (C^{\mathsf{a}})^2\sqrt{T}$	
	(Wei and Luo, 2021)	$\sqrt{T} + (C^{\mathsf{a}})^{1/3}T^{2/3}$	
	COBE + LSVI-UCB	$\sqrt{T} + C^{\mathsf{r}}$	
	COBE + VARLin	$\sqrt{T} + C^{\mathsf{a}}$	$*$
Low BE-dimension	COBE + GOLF	$\sqrt{T} + C^{\mathsf{r}}$	$*$

of standard algorithms (i.e., OFUL (Abbasi-Yadkori et al., 2011), LSVI-UCB (Jin et al., 2020b)) results in an overall regret of $\Omega(\sqrt{CT})$, a sharp contrast with the $\mathcal{O}(\sqrt{T} + C)$ bound in tabular MDPs. We find that the $\mathcal{O}(\sqrt{T}+C)$ bound can indeed be achieved efficiently in the non-contextual case (i.e., linear bandits with a fixed action set) by using the Phased Elimination (PE) approach developed by Lattimore et al. (2020); Bogunovic et al. (2021). We achieve the same bounds for linear contextual bandits and linear MDPs, but we resort to the idea of Zhang et al. (2021c), who deal with linear models through a sophisticated and computationally inefficient clipping technique. Note that the original purpose of Zhang et al. (2021c) is to get a variance-reduced bound for linear contextual bandits and a horizon-free bound for linear mixture MDPs, which are very different from our goal here. Their idea being applicable to improve robustness against corruption is surprising and of independent interest. The fact that additive dependence on C is possible under linear settings (though computationally inefficient) partially answers an open question by Zhang et al. (2021b).

We further extend our framework to general function approximation settings. We consider the class of MDPs that have low Bellman-eluder dimension (Jin et al., 2021a), and derive a corruption-robust version of their algorithm (GOLF). The algorithm achieves a regret bound of $\mathcal{O}(\sqrt{(1+C)T})$. Whether the bound of $\mathcal{O}(\sqrt{T}+C)$ is possible is left as an open problem.

In Table 1, we compare our bounds with those in previous works (omitting dependencies other than C and T). Note that C^{r} is best interpreted as $\sqrt{CT}$ in the notation of prior work. More precise bounds (including dependencies other than C and T) are provided in Table 2 in Appendix G.

2. Related Work

Corruption-robust bandit/RL have been studied under various setting, and have many other closely related topics, as we discuss in this section.

Corrupted multi-armed bandits and tabular MDPs Corruption-robust multi-armed bandits have been studied by Lykouris et al. (2018); Gupta et al. (2019); Zimmert and Seldin (2019) through three representative approaches. Interestingly, these three approaches have all been extended to the tabular MDP case by Lykouris et al. (2021); Chen et al. (2021b); Jin et al. (2021b) respectively. However, the extensions by Lykouris et al. (2021); Chen et al. (2021b) produce a new $+C^2$ term in the regret bound, largely limiting the use case of their algorithms. Besides, the computational complexity of Chen et al. (2021b)'s algorithm scales with the number of policies, which is exponentially high. On the other hand, Jin et al. (2021b) successfully achieves a near-optimal bound, but requires that the transition remains uncorrupted.

Corrupted linear bandits Li et al. (2019) and Bogunovic et al. (2020) extend the ideas of Gupta et al. (2019) and Lykouris et al. (2018) to linear bandits and Gaussian bandits respectively. Their regret bounds both have multiplicative dependence on C. Bogunovic et al. (2021) considers a stronger corruption model where the adversary can observe the action in the current round. Their bound $\sqrt{T}+C^2$ additively depends on C, but can only tolerate $C \leq \sqrt{T}$.[1] Recently, Lee et al. (2021) established the first upper bound that has optimal dependence on the amount of corruption as well as a refined gap-dependent bound. However, their algorithm only handles a restricted form of corruption – the corruption injected to action a must be in the form of $a^\top c$ for some vector c shared among all actions. In our work, we are able to get a similar bound but without this strong assumption.

Corrupted MDPs with linear function approximation Lykouris et al. (2021) studies corrupted linear MDPs and gets a bound of order $C^2\sqrt{T}$, which only tolerates $C \leq T^{1/4}$. Zhang et al. (2021b) leverages the intrinsic robustness of policy gradient and tools in robust statistics to achieve an improved bound $\sqrt{(1+C)T}$ when the feature space has a bounded *relative condition number*. Zhang et al. (2021a) further studies offline RL in linear MDPs, showing that if the offline data has wide coverage, then there is an algorithm that can output a $\mathcal{O}(\sqrt{1/T + C/T})$-optimal policy after seeing T samples with C of them corrupted. Although this result indicates that $+C$ penalty in regret might be possible, their result heavily relies on the coverage assumption and does not apply to our setting.

Robust statistics The goal of robust statistics is to design estimators of some unknown quantity that are robust to data corruption. In several recent works, computationally efficient and highly robust estimators for linear regression that tolerate a constant fraction of data corruption have been designed (Bhatia et al., 2017; Diakonikolas et al., 2019; Chen et al., 2021a). However, such strong

1. In the stronger adversary setting consider by Bogunovic et al. (2021), however, the C^2 dependence is unavoidable.

guarantees usually require additional assumptions on the data generation process or the corruption process. Robust statistics has been used in corruption-robust RL under special cases. For example, Zhang et al. (2021b,a) achieve robustness in MDPs with certain exploratory properties, and Awasthi et al. (2020) handles the case where the corrupted rounds are i.i.d. generated.

Model mis-specification The notion of corruption we consider subsumes the notion of model mis-specification studied in many previous works (Jiang et al., 2017; Du et al., 2020; Jin et al., 2020b; Zanette et al., 2020; Lattimore et al., 2020; Wang et al., 2020). These works assume that the model class can only approximate the true world up to an order of $\mathcal{O}(\epsilon)$, and they establish regret bounds that have an additive $\mathcal{O}(\epsilon T)$ penalty. Clearly, one can also view the difference between the model and the true world as corruption, and as shown in Table 1, all our bounds (both $\sqrt{T} + C^{\mathsf{a}}$ and $\sqrt{T} + C^{\mathsf{r}}$) recover the $\mathcal{O}(\sqrt{T} + \epsilon T)$ bound in the mis-specification case. While many previous works assume a known ϵ, there are also works dealing with the case of unknown ϵ (Takemura et al., 2021; Foster et al., 2020; Pacchiano et al., 2020a).

Best-of-both-world bounds The best-of-both-world problem was studied by Bubeck and Slivkins (2012); Seldin and Slivkins (2014); Auer and Chiang (2016); Seldin and Lugosi (2017); Wei and Luo (2018); Zimmert and Seldin (2019); Zimmert et al. (2019); Jin and Luo (2020); Ito (2021); Jin et al. (2021b); Lee et al. (2021) for various settings including multi-armed bandits, combinatorial semi-bandits, linear bandits, and tabular MDPs. The goal of this line of work is to have a single algorithm that achieves a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound when the reward is adversarial and $\mathcal{O}(\log T)$ when the reward is stochastic, without knowing the type of reward in advance. Compared to our setting, their regret bound is always sub-linear in T against a fixed policy, while ours is linear in the amount of corruption. However, their results usually rely on stronger structural assumptions than the corrupted setting we consider (e.g., the fixed transition assumption for MDPs or the linearized corruption assumption for linear bandits).

Non-stationary RL Non-stationary RL is another line of research that deals with non-static reward and transition (Cheung et al., 2020; Wei and Luo, 2021). In non-stationary RL, the difficulty of the problem is usually quantified by the number of times the reward or transition changes, or their fine-grained amount of variation. The corruption setting can be viewed as a special case of it, so existing algorithms for the latter can be readily applied. However, since non-stationary RL is more general, this reduction only leads to sub-optimal regret bounds. For example, the tight bound $\sqrt{T} + B^{1/3}T^{2/3}$ obtained in Wei and Luo (2021), where B is the overall variation, only translates to a sub-optimal bound $\sqrt{T} + C^{1/3}T^{2/3}$ in the corruption setting.

Model selection Our approach is closely related to the *regret balancing* technique developed by Abbasi-Yadkori et al. (2020); Pacchiano et al. (2020a); Cutkosky et al. (2021). Pacchiano et al. (2020a); Cutkosky et al. (2021) have applied regret balancing to tackle model mis-specification, but it remains unclear whether it also handles the more general corruption setting, where the adversary chooses which rounds to corrupt. Existing techniques which choose base learners deterministically can only use a regret bound that includes the total corruption budget θ which leads to loose guarantees. Instead, our randomized choice allows us to scale regret bounds of base learners as $\alpha_i\theta$, where α_i is the probability of being selected. While Pacchiano et al. (2020a); Cutkosky et al. (2021) also provide a version of their algorithm with a randomized learner choice for the special case of linear stochastic bandits with adversarial contexts, they resort to a weaker elimination test that requires additional information from base learners. Our work shows that this is indeed unnecessary by pairing a randomized learner selection with a simple elimination test. This may be of interest beyond the corruption setting.

In our work, we also develop a special model selection algorithm that achieves a gap-dependent bound that is better than $\sqrt{T}$. This kind of better-than-$\sqrt{T}$ bound is rare in the literature of model selection, and even proven to be impossible for general cases (Pacchiano et al., 2020b). To our best knowledge, the only work on model selection that breaks the $\sqrt{T}$ barrier is Arora et al. (2021), who considers a stationary multi-armed bandit setting where every base algorithm learns over a subset of arms, and the best arm is only controlled by one of the base algorithms. However, their stochastic bandit setting is less challenging than our adversarial/corrupted RL setting, so their techniques cannot be directly applied. We hope that our technique can also hint about how to achieve better-than-$\sqrt{T}$ bounds in more general model selection problems.

3. Problem Setting

We consider a general decision making framework that covers a wide range of problems. We first describe the *uncorrupted* setting. The learner is given a policy set Π and a context set $\mathcal{X}$. Ahead of time, the environment decides a context-to-expected-reward mapping $\mu^\pi : \mathcal{X} \to [0,1]$ for all $\pi \in \Pi$, which are hidden from the learner. In each round $t = 1, \ldots, T$, the environment first arbitrarily generates a context $x_t \in \mathcal{X}$, and generates a noisy reward $r_t^\pi \in [0,1]$ for all π such that $\mathbb{E}[r_t^\pi] = \mu^\pi(x_t)$. The context x_t is revealed to the learner. Then the learner chooses a policy π_t, and receives $r_t \triangleq r_t^{\pi_t}$. The goal of the learner is to minimize the regret defined as

$$\mathrm{Reg}(T) = \max_{\pi\in\Pi} \sum_{t=1}^{T} \big(\mu^\pi(x_t) - \mu^{\pi_t}(x_t)\big). \tag{1}$$

In the *corrupted* setting, the protocol is similar, but in each round t, an adversary can change the context-to-expected-reward mapping from $\mu^\pi(\cdot)$ to $\mu_t^\pi(\cdot)$. Then r_t^π is drawn such that $\mathbb{E}[r_t^\pi] = \mu_t^\pi(x_t)$. We assume that $\mu_t^\pi(x_t)$ and r_t^π still lie in $[0,1]$. As before, the learner observes x_t, chooses π_t, and receives $r_t = r_t^{\pi_t}$. The goal of the learner remains to minimize the regret defined in Eq. (1) (notice that it is defined through the uncorrupted μ). The adversary we consider falls into the category of an *adaptive adversary*, an adversary that can decide the corruption in round t based on the history up to round $t-1$.

We consider the realizable setting, where the following assumption holds:

Assumption 1 *There exists a policy $\pi^\star \in \Pi$ such that $\mu^{\pi^\star}(x_t) \geq \mu^\pi(x_t)$ for all t and all $\pi \in \Pi$.*

Below, we instantiate this framework to linear (contextual) bandits and episodic MDPs. For each setting, we define a suitable quantity c_t to measure the amount of corruption in round t. It always holds that $\max_{\pi\in\Pi} |\mu^\pi(x_t) - \mu_t^\pi(x_t)| \leq c_t$, but c_t might be strictly larger than $\max_{\pi\in\Pi} |\mu^\pi(x_t) - \mu_t^\pi(x_t)|$ in some cases.

Linear contextual bandits In linear contextual bandits, the policy set can be identified as a bounded parameter set $\mathcal{W} \subset \mathbb{R}^d$ in which the true underlying parameter $w^\star$ lies, and the "context" can be identified as the "action set" in each round. In the uncorrupted setting, in each round t, the environment first generates an action set $\mathcal{A}_t \subset \mathbb{R}^d$ (the context). Then each policy $w \in \mathcal{W}$ is associated with an action $a^w(\mathcal{A}_t) = \mathrm{argmax}_{a\in\mathcal{A}_t} \langle w, a\rangle$. The expected reward of policy w under context $\mathcal{A}_t$ is then given by $\mu^w(\mathcal{A}_t) = \langle w^\star, a^w(\mathcal{A}_t)\rangle \in [0,1]$ for all w. Clearly, $w^\star$ is a policy that satisfies Assumption 1.

In the corrupted setting, for any t and w, the adversary can choose $\mu_t^w(\cdot)$ to be an arbitrary mapping from $\mathcal{A}_t$ to $[0,1]$, and r_t^w is generated such that $\mathbb{E}[r_t^w] = \mu^w(\mathcal{A}_t)$ and $r_t^w \in [0,1]$. We define $c_t = \max_{w\in\mathcal{W}} |\mu^w(\mathcal{A}_t) - \mu_t^w(\mathcal{A}_t)|$.

Linear bandits Linear bandits can be viewed as a special case of linear contextual bandits with a fixed action set $\mathcal{A}_t = \mathcal{A}$ for all t. In this case, since every policy chooses the same action in every round, a more direct formulation is to identify the policy set as the action set $\mathcal{A}$ and ignore the context. The expect reward of policy/action $a \in \mathcal{A}$ is given by $\mu^a = \langle w^\star, a\rangle \in [0,1]$ for some unknown $w^\star$. Note that the action $\mathrm{argmax}_a \langle w^\star, a\rangle$ satisfies Assumption 1. In the corrupted setting, μ_t^a can be set to an arbitrary value in $[0,1]$, and r_t^a is generated such that $\mathbb{E}[r_t^a] = \mu_t^a$ and $r_t^a \in [0,1]$. We define $c_t = \max_{a\in\mathcal{A}} |\mu^a - \mu_t^a|$.

Episodic Markov decision processes An episodic MDP is associated with a state space $\mathcal{S}$, an action space $\mathcal{A}$, a number of layers H, a transition kernel $p : \mathcal{S}\times\mathcal{A} \to \Delta_{\mathcal{S}}$, and a reward function $\sigma : \mathcal{S}\times\mathcal{A} \to [0, \frac{1}{H}]$. A policy $\pi = \{\pi_h : \mathcal{S} \to \mathcal{A},\ h = 1, 2, \ldots, H\}$ consists of mappings $\mathcal{S} \to \mathcal{A}$ for each layer that specifies which action it takes in each state in that layer. The context is identified as the "initial state."

In the uncorrupted setting, in each round t, the environment arbitrarily generates an initial state $s_{t,1} \in \mathcal{S}$ (the context). Then the learner decides a policy $\pi_t = \{\pi_{t,h}\}_{h=1}^H$, and interacts with the environment for H steps starting from $s_{t,1}$. On the h-th step, she chooses an action $a_{t,h} = \pi_{t,h}(s_{t,h})$, observes a noisy reward $\sigma_{t,h} \in [0, \frac{1}{H}]$ with $\mathbb{E}[\sigma_{t,h}] = \sigma(s_{t,h}, a_{t,h})$, and transitions to the next state $s_{t,h+1} \sim p(\cdot|s_{t,h}, a_{t,h})$.[2] The round ends right after the learner transitions to state $s_{t,H+1}$. With this procedure, the expected reward of policy π given the initial state s can be represented as

$$\mu^\pi(s) = \mathbb{E}\left[\sum_{h=1}^H \sigma(s_h, a_h)\ \middle|\ s_1 = s,\ a_h = \pi_h(s_h),\ s_{h+1} \sim p(\cdot|s_h, a_h),\ \forall h = 1, \ldots, H\right], \quad (2)$$

and r_t^π is a realized reward of policy π in round t, which satisfies $\mathbb{E}[r_t^\pi] = \mu^\pi(s_{t,1})$.

In the corrupted setting, in round t, we allow the adversary to change p and σ to p_t and σ_t respectively. The corrupted expected value $\mu_t^\pi(s)$ is defined similarly to Eq. (2) but with p and σ replaced by p_t and σ_t. We assume that after corruption, $\sigma_{t,h}$ (whose expectation is $\sigma_t(s_{t,h}, a_{t,h})$) still lies in $[0, \frac{1}{H}]$.

To measure the amount of corruption in round t, we define the Bellman operators $\mathcal{T} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ under the uncorrupted MDP and $\mathcal{T}_t$ under the corrupted MDP as $(\mathcal{T}V)(s,a) \triangleq \sigma(s,a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[V(s')]$ and $(\mathcal{T}_t V)(s,a) \triangleq \sigma_t(s,a) + \mathbb{E}_{s'\sim p_t(\cdot|s,a)}[V(s')]$ for $V \in \mathbb{R}^{\mathcal{S}}$. Then the amount of corruption in round t is defined as $c_t \triangleq H \cdot \sup_{s,a} \sup_{V\in[0,1]^{\mathcal{S}}} |(\mathcal{T}V - \mathcal{T}_t V)(s,a)|$.

Additional note on corruption In all the above settings, we assume that the corruption is bounded. It holds that $c_t \le c_{\max}$ for some $c_{\max}$ in all t. For linear (contextual) bandits, we can set $c_{\max} = 1$, while for episodic MDPs, we can set $c_{\max} = 2H$. While we assume bounded corruption to keep the exposition clean, our algorithm can actually handle more general scenarios. For example, for linear contextual bandits, we can handle the case where $\mu^\pi(\cdot) \in [0,1]$, but $\mu_t^\pi(\cdot)$ can be arbitrary, and

2. Our setting is a scaled version of the standard episodic MDP setting where all rewards are scaled by $1/H$. This scaling does not affect the difficulty of the problem but allows us to unify the presentation with our other settings.

$r_t^\pi - \mu_t^\pi(x_t)$ is zero-mean and 1-sub-Gaussian. For this case, since $r_t^\pi - \mu_t^\pi(x_t)$ is bounded between $\pm c' \log(1/\delta)$ for some absolute constant c' with high probability (i.e., with probability $1 - \mathcal{O}(\delta)$), if we receive a reward r_t that is outside $[-c' \log(1/\delta), 1 + c' \log(1/\delta)]$, then with high probability it is caused by corruption. The learner only needs to project the reward back to this range. This essentially reduces the problem to bounded corruption case.

Other notations We use $[u, v]$ to denote $\{u, u+1, \ldots, v\}$, and $[u]$ to denote $\{1, 2, \ldots, u\}$. The notations $\widetilde{\mathcal{O}}(\cdot)$, $\widetilde{\Theta}(\cdot)$ hide poly-logarithmic factors.

3.1. Two Ways to Compute Aggregated Corruption

In previous works of corruption-robust RL, the total corruption is defined as $C = \sum_{t=1}^T c_t$, where c_t is the per-round corruption defined above. In Section 1 and Section 2, we also adopt this definition when comparing with previous works. However, to unify the analysis under different settings, we introduce another notion of total regret defined as $\sqrt{T \sum_{t=1}^T c_t^2}$. To distinguish them, we denote $C^{\mathsf{a}} = \sum_{t=1}^T c_t$ and $C^{\mathsf{r}} = \sqrt{T \sum_{t=1}^T c_t^2}$, for that C^{a} is T times the *arithmetic mean* of c_t's, while C^{r} is T times the *root mean square* of c_t's. By defining C^{r}, we are able to recover the bounds in the "model misspecification" literature, in which the regret bound is often expressed through $T \max_t c_t$, which is an upper bound of C^{r} (see Table 1 and Section 2 for more details). We further define $C_t^{\mathsf{a}} \triangleq \sum_{\tau=1}^t c_\tau$ and $C_t^{\mathsf{r}} \triangleq \sqrt{t \sum_{\tau=1}^t c_\tau^2}$.

4. Gap-Independent Bounds via Model Selection

In this section, we develop a general corruption-robust algorithm based on model selection. The regret bound we achieve is of order $\widetilde{\mathcal{O}}(\sqrt{T} + C)$, where C is either C^{a} or C^{r} (see Table 1 for the choices in different settings). Model selection approaches rely on a meta algorithm learning over a set of base algorithms. We first specify the properties that each base algorithm should satisfy:

Assumption 2 (base algorithm, with either $C_t \triangleq C_t^{\mathsf{a}}$ or $C_t \triangleq C_t^{\mathsf{r}}$) ALG *is an algorithm that takes as input a time horizon T, a confidence level δ, and a hypothetical corruption level θ.* ALG *ensures the following: with probability at least $1 - \delta$, for all $t \le T$ such that $C_t \le \theta$, it holds that*

$$\sum_{\tau=1}^{t} (r_\tau^{\pi^\star} - r_\tau) \le \mathcal{R}(t, \theta)$$

for some function $\mathcal{R}(t, \theta)$. Without loss of generality, we assume that $\mathcal{R}(t, \theta)$ is non-decreasing in both t and θ, and that $\mathcal{R}(t, \theta) \ge \theta$.

If a base algorithm satisfies Assumption 2 with $C_t \triangleq C_t^{\mathsf{a}}$, we call it a type-a base algorithm, while if $C_t \triangleq C_t^{\mathsf{r}}$, we call it a type-r. Base algorithms are essentially corruption-robust algorithms that require the prior knowledge of the total corruption. Therefore, the algorithms developed by Lykouris et al. (2021, Appendix B) or Wu et al. (2021) can be readily used as our base algorithms. For example, for tabular MDPs, a variant of the UCBVI algorithm (Azar et al., 2017) satisfies Assumption 2 with $C_t \triangleq C_t^{\mathsf{a}}$ and $\mathcal{R}(t, \theta) = \text{poly}(H, \log(SAT/\delta))(\sqrt{SAt} + S^2A + SA\theta)$; for linear MDPs, a variant of the LSVI-UCB algorithm (Jin et al., 2020b) satisfies Assumption 2 with $C_t \triangleq C_t^{\mathsf{r}}$ and $\mathcal{R}(t, \theta) = \text{poly}(H, \log(dT/\delta))(\sqrt{d^3 t} + d\theta)$. More examples are provided in Appendix G.

A base algorithm with a higher hypothetical corruption level θ is more *robust*, but incurs more regret overhead. In contrast, base algorithms with lower hypothetical corruption level introduce less overhead, but have higher possibility of mis-specifying the amount of corruption. When the true total corruption is unknown, just running a single base algorithm with a fixed θ is risky either way.

The idea of our algorithm is to simultaneously run multiple base algorithms (in each round, sample one of the base algorithms and execute it), each with a different hypothesis on the total amount of corruption. This idea is also used by Lykouris et al. (2021). Intuitively, if two base algorithms have a valid hypothesis for the total corruption (i.e., their hypotheses upper bound the true total corruption), then the one with smaller hypothesis should learn faster than the larger one because its hypothesis is closer to the true value, and incurs less overhead. Therefore, if at some point we find that the average performance of a base algorithm with a smaller hypothesis is significantly worse than that of a larger one, it is an evidence that the former has mis-specified the amount of corruption. If this happens, we simply stop running this base algorithm.

There are two key questions to be answered. First, what distribution should we use to select among the base algorithms? Second, given this distribution, how should we detect mis-specification of the amount of corruption by comparing the performance of base algorithms? In Section 4.1, we answer the second question. The first question will be addressed in Section 4.2 and Section 5 slightly differently depending on our target regret bound.

4.1. Single Epoch Algorithm

In this section, we analyze BASIC (Algorithm 1), a building block of our final algorithms. In BASIC, the distribution over base algorithms is fixed and given as an input (α in Algorithm 1). Other inputs include: a length parameter L that specifies the maximum number of rounds (the algorithm might terminate before finishing all L rounds though) and an index $k \in [k_{\max}]$ ($k_{\max}$ is defined in Algorithm 1) that specifies the smallest index of base algorithms (the base algorithms are indexed by $k, k+1, \ldots, k_{\max}$).

Below, we sometimes unify the statements for the two definitions of total corruption (see Section 3.1). The notations (C, C_t) refer to $(C^{\mathsf{a}}, C_t^{\mathsf{a}})$ if the base algorithm is type-a, and refer to $(C^{\mathsf{r}}, C_t^{\mathsf{r}})$ if it is type-r. We will explicitly write the superscripts if we have to distinguish them.

The base algorithm with index $i \in [k, k_{\max}]$ (denoted as ALG_i) hypothesizes that the total corruption C is upper bounded by 2^i. We say ALG_i is *well-specified* at round t if $C_t \le 2^i$; otherwise we say it is *mis-specified* at round t. Naively, we might want to set the θ parameter of ALG_i to 2^i. However, we can actually set it to be smaller to reduce the overhead, as explained below. Since each base algorithm is sub-sampled according to the distribution α, the total corruption experienced by ALG_i in $[1, t]$ is only roughly $\sum_{\tau \le t} \alpha_i c_\tau \le \alpha_i C^{\mathsf{a}}$ or $\sqrt{(\alpha_i t) \sum_{\tau \le t} \alpha_i c_\tau^2} \le \alpha_i C^{\mathsf{r}}$ (for type-a and type-r base algorithms respectively). This means that ALG_i, which hypothesizes a total corruption of 2^i, only needs to set the θ parameter in Assumption 2 to roughly $\alpha_i 2^i$, instead of 2^i. Our choice of θ_i in Eq. (3) is slightly larger than $\alpha_i 2^i$ to accommodate the randomness in the sampling procedure.

Besides performing sampling over base algorithms, BASIC also compares the performance any two base algorithms using Eq. (4). If all base algorithms hypothesize large enough corruption, then all of them enjoy the regret bound specified in Assumption 2 in the subset of rounds they are executed. In this case, we can show that with high probability, the termination condition Eq. (4) will not hold. This is formalized in Lemma 1.

Algorithm 1 **B**ase **A**lgorithms run **S**imultaneously with m**I**s-specification **C**heck (BASIC)

input: base algorithm ALG satisfying Assumption 2, $L \in [T]$, $k \in [k_{\max}]$ where $k_{\max} \triangleq \lceil \log_2(c_{\max} L) \rceil$, $\delta \in (0,1)$, and a distribution $\alpha = (\alpha_k, \alpha_{k+1}, \dots, \alpha_{k_{\max}})$ satisfying:

$$\alpha_k \geq \alpha_{k+1} \geq \cdots \geq \alpha_{k_{\max}} > 0 \qquad \text{and} \qquad \sum_{i=k}^{k_{\max}} \alpha_i = 1.$$

for $i = k, \dots, k_{\max}$ **do**

Initiate an instance of ALG with inputs T, δ, and θ chosen as below:

$$\theta_i \triangleq \begin{cases} 1.25 \cdot \alpha_i 2^i + 21 c_{\max} \log(T/\delta) & \text{if ALG is type-a} \\ 1.25 \cdot \alpha_i 2^i + 8 c_{\max} \sqrt{\alpha_i L \log(T/\delta)} + 21 c_{\max} \log(T/\delta) & \text{if ALG is type-r} \end{cases} \tag{3}$$

(We call this instance ALG_i.)

end

for $t = 1, \dots, L$ **do**

Random pick an sub-algorithm $i_t \sim \alpha$, receive the context x_t, and use ALG_{i_t} to output π_t.

Execute π_t, receive feedback, and perform update on ALG_{i_t}.

Define $N_{t,i} \triangleq \sum_{\tau=1}^{t} \mathbf{1}[i_\tau = i]$, $R_{t,i} \triangleq \sum_{\tau=1}^{t} \mathbf{1}[i_\tau = i] r_\tau$.

if $\exists i, j \in [k, k_{\max}]$, $i < j$, such that

$$\frac{R_{t,i}}{\alpha_i} + \frac{\mathcal{R}(N_{t,i}, \theta_i)}{\alpha_i} < \frac{R_{t,j}}{\alpha_j} - 8 \left(\sqrt{\frac{t \log(T/\delta)}{\alpha_j}} + \frac{\log(T/\delta) + \theta_j}{\alpha_j} \right), \tag{4}$$

return false.

end

end

return true.

Lemma 1 *With probability at least $1 - \mathcal{O}(k_{\max}\delta)$, the termination condition Eq. (4) of the* BASIC *algorithm, does not hold in any round t, such that $C_t \leq 2^k$.*

In other words, Eq. (4) is triggered only when $C_t > 2^k$, i.e., ALG_k is mis-specified at round t. Once this happens, the BASIC algorithm terminates. Checking condition Eq. (4) essentially ensures that the quantity $\frac{R_{t,i}}{\alpha_i}$ of all base algorithms remain close. Notice that at all t, there is always a well-specified base algorithm $i^\star$ with $C_t \leq 2^{i^\star}$ which enjoys the regret guarantee of Assumption 2. Therefore, $\frac{R_{t,i^\star}}{\alpha_{i^\star}}$ is not too low, and thus, testing condition Eq. (4) prevents $\frac{R_{t,i}}{\alpha_i}$ of any i from falling too low. This directly controls the performance of every base algorithms before termination. The following lemma bounds the learner's cumulative regret at termination.

Lemma 2 *Let $L_0 \leq L$ be the round at which* BASIC *terminates, and let $i^\star$ be the smallest $i \in [k, k_{\max}]$ such that $C_{L_0} \leq 2^i$. Then with probability at least $1 - \mathcal{O}(k_{\max}\delta)$,*

$$\sum_{t=1}^{L_0} (r_t^{\pi^\star} - r_t) \leq \sum_{i=k}^{k_{\max}} \mathcal{R}(N_{L_0,i}, \theta_i) + \widetilde{\mathcal{O}}\left(\mathbf{1}\left[i^\star > k\right] \left(\sqrt{\frac{L_0}{\alpha_{i^\star}}} + \frac{\mathcal{R}(N_{L_0,i^\star}, \theta_{i^\star})}{\alpha_{i^\star}} \right) \right).$$

Algorithm 2 COrruption-robustness through Balancing and Elimination (COBE)

input: base algorithm ALG satisfying Assumption 2 with the form specified in Eq. (5).
define: $Z \triangleq c_{\max}$ if ALG is type-a, and $Z \triangleq c_{\max}\sqrt{T}$ if ALG is type-r.
$k_{\text{init}} \triangleq \max\left\{\left\lceil \log_2 \frac{\sqrt{\beta_1 T}+\beta_2 Z+\beta_3}{\beta_2}\right\rceil,\ 0\right\}$ with $\beta_1, \beta_2, \beta_3$ defined in Eq. (5).
for $k = k_{init}, \ldots$ **do**
  Run BASIC with input k and $L = T$, and $\{\alpha_i\}_{i=k}^{k_{\max}}$ specified in Eq. (6), until it terminates or the total number of rounds reaches T.
end

where $N_{L_0,i}$ is the total number of rounds ALG_i was played.

4.2. Corruption-robust Algorithms with $\sqrt{T} + C^{\mathsf{a}}$ or $\sqrt{T} + C^{\mathsf{r}}$ Bounds

Next, we use BASIC to build a corruption-robust algorithm with a regret bound of either $\sqrt{T} + C^{\mathsf{a}}$ or $\sqrt{T} + C^{\mathsf{r}}$ without prior knowledge of C^{a} or C^{r}. The algorithm is called COBE and presented in Algorithm 2. We consider base algorithms with the following concrete form of $\mathcal{R}(t, \theta)$:

$$\mathcal{R}(t, \theta) = \sqrt{\beta_1 t} + \beta_2 \theta + \beta_3 \tag{5}$$

for some $\beta_1, \beta_2, \beta_3 \geq 1$. COBE starts with $k = k_{\text{init}}$ (defined in Algorithm 2) and runs BASIC with inputs k and $L = T$ and the following choice of $\{\alpha_i\}_{i=k}^{k_{\max}}$:

$$\alpha_i = \begin{cases} 2^{k-i-1} & \text{for } i > k, \\ 1 - \sum_{i=k+1}^{k_{\max}} \alpha_i & \text{for } i = k. \end{cases} \tag{6}$$

Whenever the subroutine BASIC terminates before T, we eliminate the ALG_k and start a new instance of BASIC with k increased by 1 (see the for-loop in COBE). This is because as indicated by Lemma 1, early termination implies that ALG_k mis-specifies the amount of corruption.

Notice that $2^{k_{\text{init}}}$ is roughly of order $\sqrt{T}$, i.e., we start from assuming that the total amount of corruption is $\sqrt{T}$. This is because we only target the worst-case regret rate of $\sqrt{T} + C$ here, so refinements for smaller corruption levels $C \leq \sqrt{T}$ do not improve the asymptotic bound. Our choice of α_i makes $\alpha_i 2^i \approx 2^k$ for all i, and this further keeps the magnitudes of $\mathcal{R}(N_{t,i}, \theta_i)$ of all i's roughly the same. This conforms with the regret balancing principle by Abbasi-Yadkori et al. (2020); Pacchiano et al. (2020a), as well as the sub-sampling idea of Lykouris et al. (2021). This makes the bound of the model selection algorithm only worse than the best base algorithm by a factor of $\mathcal{O}(k_{\max}) = \widetilde{\mathcal{O}}(1)$ if all base algorithms are well-specified.

In the following theorem, we show guarantees of COBE for both $C \triangleq C^{\mathsf{a}}$ and $C \triangleq C^{\mathsf{r}}$. The proof essentially plugs the choices of parameters into Lemma 2, and sum the regret over epochs.

Theorem 3 *If ALG satisfies Assumption 2 and $\mathcal{R}(t, \theta)$ in the form of Eq. (5), then with α_i's specified in Eq. (6), COBE guarantees with probability at least $1 - \mathcal{O}(k_{\max}\delta)$ that*

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 T} + \beta_2 (C^{\mathsf{a}} + Z) + \beta_3\right),$$

where $Z = c_{\max}$ if $C \triangleq C^{\mathsf{a}}$ and $Z = c_{\max}\sqrt{T}$ if $C \triangleq C^{\mathsf{r}}$.

Algorithm 3 Gap-bound enhanced COBE (G-COBE)

1 $k_{\text{init}} = \max\left\{\left\lceil \log_2 \frac{\sqrt{\beta_1}+\beta_2 c_{\max}+\beta_3}{\beta_2}\right\rceil,\ 0\right\},\ \beta_4 = 10^4\left(2\beta_1 + 42\beta_2 c_{\max}\log(T/\delta) + 2\beta_3\right).$

2 **for** $k = k_{init}, \ldots$ **do**

3   `// Phase 1`

4   Let L be the smallest integer such that $\sqrt{\beta_4 L} \geq \beta_2 2^k$.

5   **if** $L > T$ **then break**;

6   Run BASIC with input k and L, and $\{\alpha_i\}_{i=k}^{k_{\max}}$ specified in Eq. (6) until it terminates or the number of rounds reaches T. Let o be its output, and let $\widehat{\pi}$ be the policy that is executed the most number of times by the base algorithm ALG_k.

7   **if** $o =$ ***true*** **then**

8   `// Phase 2`

9   Run TwoModelSelect with $L, \widehat{\pi}, \mathcal{B}_{\widehat{\pi}}$, until it terminates or the number of rounds reaches T.

    end

end

10 `// Phase 3`

11 Run COBE in the remaining rounds.

5. Gap-Dependent Bounds

In this section, the goal is to get instance-dependent bounds similar to those in Lykouris et al. (2021); Chen et al. (2021b). There are extra assumptions to be made in this section. First, we only deal with the case without contexts, i.e., the following assumption holds:

Assumption 3 *Assume that* $\mu^\pi(x_t) = \mu^\pi$.

This covers linear bandits and MDPs with a fixed initial state. In fact, our approach can handle a slightly more general case where the context is i.i.d. generated in the uncorrupted case, and the non-iid-ness of the context distribution is considered as corruption (in contrast, in Section 4, the non-iid-ness of contexts is not considered as corruption). Besides, our bound depends on the sub-optimality gap defined in the following:

Assumption 4 *There exists a policy* $\pi^\star \in \Pi$ *such that for all* $\pi \in \Pi\backslash\{\pi^\star\}$, $\mu^\pi \leq \mu^{\pi^\star} - \Delta$.

This gap assumption is in fact stronger than that made by Chen et al. (2021b). In (Chen et al., 2021b), $\Delta := \min_{\pi:\, \Delta_\pi > 0} \Delta_\pi$ where $\Delta_\pi = \mu^{\pi^\star} - \mu^\pi$. Their definition keeps $\Delta > 0$ when there are multiple optimal policies, while our Assumption 4 forces $\Delta = 0$ if there are two optimal policies with the same expected reward. This kind of stronger gap assumption is similar to those in (Lee et al., 2021; Jin et al., 2021b). Finally, we only focus on the case with $C = C^{\mathsf{a}}$ throughout this section. [3]

5.1. Algorithm Overview

Our algorithm G-COBE (Algorithm 3) consists of three phases where the first two phases are executed interleavingly. In Phase 1 (Line 4-Line 6 in G-COBE), we run BASIC with a type-a base

3. When $C = C^{\mathsf{r}}$, our approach produces a regret term of $c_{\max}\sqrt{T}$ as in Theorem 3, spoiling the gap-dependent bound.

algorithm that satisfies Assumption 2 with the following gap-dependent bound:

$$\mathcal{R}(t,\theta) = \min\left\{\sqrt{\beta_1 t},\ \frac{\beta_1}{\Delta}\right\} + \beta_2\theta + \beta_3 \tag{7}$$

for some $\beta_1, \beta_2, \beta_3$ satisfying $\beta_1 \geq 16\log(T/\delta)$, $\beta_2 \geq 1$, $\beta_3 \geq 10\sqrt{\beta_1 \log(T/\delta)}$.

In every for-loop of k, if BASIC in Phase 1 returns true, the algorithm proceeds to Phase 2 (Line 9 in G-COBE). In Phase 2 we execute TwoModelSelect (Algorithm 4). TwoModelSelect is a specially designed two-model selection algorithm that dynamically chooses between two instances. One of the them is $\widehat{\pi}$, a candidate optimal policy identified in Phase 1 (defined in Line 6 of G-COBE); the other is an algorithm with $\widehat{\pi}$ as input (we call this algorithm $\mathcal{B}_{\widehat{\pi}}$). We assume that $\mathcal{B}_{\widehat{\pi}}$ has the following property:

Assumption 5 *$\mathcal{B}_{\widehat{\pi}}$ is a corruption-robust algorithm over the policy set $\Pi\backslash\{\widehat{\pi}\}$ without the prior knowledge of total corruption. In other words, when running alone, in every round t, it chooses a policy $\pi_t \in \Pi\backslash\{\widehat{\pi}\}$ and receives r_t with $\mathbb{E}[r_t] = \mu_t^{\pi_t}$. It ensures the following for all t with probability at least $1-\delta$:*

$$\max_{\pi\in\Pi\backslash\{\widehat{\pi}\}} \sum_{\tau=1}^{t} (r_\tau^\pi - r_\tau) \leq \mathcal{R}_{\mathcal{B}}(t, C_t) \triangleq \sqrt{\beta_1 t} + \beta_2 C_t + \beta_3. \tag{8}$$

Notice that in Section 4.2 we have already developed a corruption-robust algorithm COBE, whose guarantee is already in the form of Eq. (8), albeit over the original policy set Π (see Theorem 3). In Appendix F, we describe how to implement $\mathcal{B}_{\widehat{\pi}}$ through running COBE on a modified MDP.

The TwoModelSelect in Phase 2 might end earlier than time T. This happens only when $\frac{1}{\Delta} + C$ is larger than the order of 2^k. In this case, the algorithm goes back to Phase 1 with k increased by 1. When 2^k grows to the order of $\sqrt{T}$ (implying that $\sqrt{T} \gtrsim \frac{1}{\Delta} + C$), we instead proceed to Phase 3 and simply run COBE in the remaining rounds (Line 5 of G-COBE).

The regret guarantee of G-COBE is summarized by the following theorem.

Theorem 4 G-COBE *ensures that (with β_4 defined in Algorithm 3)*

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\min\left\{\sqrt{\beta_4 T},\ \frac{\beta_4}{\Delta}\right\} + \beta_2 C + \beta_4\right).$$

Theorem 4 gives the first $\min\{\frac{1}{\Delta}, \sqrt{T}\} + C$ bound in the literature of corrupted MDPs without the knowledge of C. To show Theorem 4, we establish some key lemmas for Phase 1 and Phase 2 in Section 5.2 and Section 5.3 respectively. The complete proof of Theorem 4 is given in Appendix E. Note that within the sub-routines BASIC and TwoModelSelect, we re-index the time so that they both start from $t = 1$ for convenience.

5.2. Phase 1 of G-COBE

In Phase 1 we run BASIC with base algorithms that achieve gap-dependent bounds Eq. (7). The regret bound of BASIC under general choices of $\mathcal{R}(t,\theta)$ and α_i is already derived in Lemma 2. Here, we apply it with the new form of $\mathcal{R}(t,\theta)$ in Eq. (7), and the new choice of α_i as below:

$$\alpha_i = \begin{cases} \min\left\{\frac{\sqrt{\beta_1 L}/\beta_2 + 2^k}{2^i},\ \frac{1}{2(k_{\max}-k)}\right\} & \text{for } i > k \\ 1 - \sum_{i=k+1}^{k_{\max}} \alpha_i & \text{for } i = k \end{cases} \tag{9}$$

The regret bound of BASIC under such choices of parameters is summarized as the following:

Lemma 5 *Let $L_0 \le L$ be the round at which* BASIC *terminates. If $\mathcal{R}(t,\theta)$ is in the form of Eq. (7), and α_i's follow Eq. (9), then with high probability,* BASIC *guarantees*

$$\sum_{t=1}^{L_0}(r_t^{\pi^\star} - r_t) = \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 L_0} + \beta_2 C_{L_0} + \beta_2 c_{\max} + \beta_3\right).$$

We see that even though our base algorithms achieve a gap-dependent bound (Eq. (7)), the advantage is not reflected on the final bound of BASIC (as can be seen in Lemma 5, we still do not achieve a gap-dependent bound). This is due to the fundamental limitation of general model selection problems (Pacchiano et al., 2020b). Therefore, Lemma 5 does not seem to give any advantage over Theorem 3. However, the hidden advantage of using base algorithms with gap-dependent bounds is that if a base algorithm well-specifies the total corruption, it will quicker concentrate on the best policy. This enables the learner to *identify the best policy* faster. This is formalized in Lemma 6.

Lemma 6 *Suppose that we run* BASIC *with base algorithms satisfying Eq. (7). Let $L_0 \le L$ be the round at which* BASIC *terminates. If $32\left(\frac{\beta_4}{\Delta} + \beta_2 C_{L_0}\right) \le \beta_2 2^k \le \sqrt{\beta_4 L}$, then with probability at least $1 - \mathcal{O}(\delta)$, $L_0 = L$, and the following holds:*

$$\sum_{t=1}^{L} \mathbf{1}[i_t = k]\mathbf{1}[\pi_t = \pi^\star] > \frac{1}{2}\sum_{t=1}^{L}\mathbf{1}[i_t = k]. \tag{10}$$

Lemma 6 ensures that if $2^k \gtrsim \frac{1}{\Delta} + C$, by looking at which policy is most frequently executed by ALG_k, the learner can correctly identify the best policy $\widehat{\pi} = \pi^\star$ with high probability (by Eq. (10) and the definition of $\widehat{\pi}$ in G-COBE).

5.3. Phase 2 of G-COBE

In Phase 2, we execute TwoModelSelect, which is a model selection algorithm between $\widehat{\pi}$ and $\mathcal{B}_{\widehat{\pi}}$. The high-level goal is to make the learner concentrate on executing $\widehat{\pi}$ until the end of T rounds if $\widehat{\pi} = \pi^\star$ and $\frac{1}{\Delta} + C$ is relatively small, and otherwise terminate the algorithm quickly before incurring too much regret. It proceeds in epochs of varying length, indexed with j. The quantity $\widehat{\Delta}_j$ is an estimator of the gap between the average performance of $\widehat{\pi}$ and $\mathcal{B}_{\widehat{\pi}}$ at the beginning of epoch j; M_j is the maximum possible length of epoch j, and p_j is the probability that the learner chooses $\mathcal{B}_{\widehat{\pi}}$ in epoch j. The learner constantly monitors the difference between the average performance of $\widehat{\pi}$ and $\mathcal{B}_{\widehat{\pi}}$ (Line 9-Line 13 in TwoModelSelect). Whenever she finds that their performance gap is actually much smaller or larger than $\widehat{\Delta}_j$ (i.e., if Eq. (11) or Eq. (12) holds), she updates $\widehat{\Delta}_j, M_j$, and p_j, and restarts a new epoch. If at any time $\widehat{\Delta}_j$ becomes smaller than $\widehat{\Delta}_1$, or j grows larger than $3\log^2 T$, she terminates TwoModelSelect. We establish the following two key lemmas.

Lemma 7 *Let T_0 be the last round of* TwoModelSelect*, then with probability at least $1 - \mathcal{O}(\delta)$,*

$$\sum_{t=1}^{T_0}(r_t^{\pi^\star} - r_t) = \widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L} + \beta_2 C_{T_0} + \beta_4\right).$$

Lemma 8 *Let T_0 be last round of* TwoModelSelect*. If $\widehat{\pi} = \pi^*$ and $\sqrt{\beta_4 L} \ge 16\left(\frac{\beta_4}{\Delta} + \beta_2 C_{T_0}\right)$, then with probability at least $1 - \mathcal{O}(\delta)$, it is terminated because the number of rounds reaches T.*

We combine Lemma 5-Lemma 8 to prove Theorem 4 in Appendix E.

Algorithm 4 TwoModelSelect $(L, \widehat{\pi}, \mathcal{B}_{\widehat{\pi}})$

initialization: $\widehat{\Delta}_1 \leftarrow \min\left\{\sqrt{\frac{\beta_4}{L}}, 1\right\}, \ M_1 \leftarrow \frac{\beta_4}{\widehat{\Delta}_1^2}, \ t \leftarrow 1.$ (β_4 defined in Algorithm 3)

for $j = 1, 2, \ldots, (3\log^2 T)$ **do**

  $t_j \leftarrow t, \ \ p_j \leftarrow \frac{\beta_4}{2M_j\widehat{\Delta}_j^2}$, and re-initialize $\mathcal{B}_{\widehat{\pi}}$.

  while $t \le t_j + M_j - 1$ **do**

    $Y_t \leftarrow \text{Bernoulli}\,(p_j)$.

    if $Y_t = 1$ **then** Execute $\mathcal{B}_{\widehat{\pi}}$ for one round and update $\mathcal{B}_{\widehat{\pi}}$;

    else Execute $\widehat{\pi}$ for one round ;

    $t \leftarrow t + 1$

    Let $\widehat{R}_0 = \frac{1}{1-p_j}\sum_{\tau=t_j}^{t-1} r_\tau \mathbf{1}[Y_\tau = 0], \ \ \widehat{R}_1 = \frac{1}{p_j}\sum_{\tau=t_j}^{t-1} r_\tau \mathbf{1}[Y_\tau = 1].$

    if $\widehat{R}_0 \le \widehat{R}_1 + \frac{1}{2}(t-t_j)\widehat{\Delta}_j - \frac{5}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j(t-t_j), \frac{p_j\sqrt{\beta_1 L}}{\beta_2}\right)$ (11)

    then $\widehat{\Delta}_{j+1} \leftarrow \frac{1}{1.25}\widehat{\Delta}_j$ **and break**

    if $\widehat{R}_0 \ge \widehat{R}_1 + 3M_j\widehat{\Delta}_j + 8\sqrt{\beta_1 L}$ (12)

    then $\widehat{\Delta}_{j+1} \leftarrow 1.25\widehat{\Delta}_j$ **and break**

  end

  if $\widehat{\Delta}_{j+1} < \widehat{\Delta}_1$ **then return**;

  $M_{j+1} \leftarrow 2(t-t_j) + \frac{\beta_4}{\widehat{\Delta}_{j+1}^2}$ (13)

end

6. Applications to Different Settings

In Appendix G, we give examples of the base algorithms whose regret bound is of the form Eq. (5) or Eq. (7). For tabular MDPs, we directly use the Robust UCBVI algorithm by Lykouris et al. (2021) as our base algorithm (Appendix G.1). For linear bandit, we adopt the Robust Phased Elimination algorithm developed by Bogunovic et al. (2021), and additionally prove a gap-dependent bound for it (Appendix G.2). For linear contextual bandits and linear MDPs, we modify the OFUL/LSVI-UCB algorithm to make them robust to corruption (Appendix G.3). Then we extend the VOFUL/VARLin algorithms by Zhang et al. (2021c), further improving the dependence on C over the OFUL/LSVI-UCB approach (Appendix G.4). Finally, we derive a corruption-robust variant of the GOLF algorithm by Jin et al. (2021a) for the general function approximation setting (Appendix G.5).

7. Conclusions and Future Work

In this work, we develop a general model selection framework to deal with corruption in bandits and reinforcement learning. In the tabular MDP setting, without knowing the total corruption, our result is the first to achieve a worst-case optimal bound. This resolves open problems raised by Lykouris et al. (2021); Chen et al. (2021b); Wu et al. (2021). A general framework to obtain refined gap-dependent bounds is also developed. In linear bandits, linear contextual bandits, and linear MDPs, our bounds also improve those of previous works in various ways.

However, our result is not the end of the story. There are many remaining open problems to be investigated in the future:

- For the tabular setting, our gap complexity measure is larger than those in (Simchowitz and Jamieson, 2019; Lykouris et al., 2021; Jin et al., 2021b). It is an important future direction to further improve our gap-dependent bound without sacrificing the worst-case dependence on T or C.
- For linear contextual bandits and linear MDPs, a regret bound with additive dependence on C^{a} is only achieved through computationally inefficient algorithms (i.e., the variants of VOFUL and VARLin). These algorithms also have a bad dependence on the feature dimension d. Can we address these computational and statistical issues?
- In the model mis-specification literature, Agarwal et al. (2020a); Zanette et al. (2021) defines a new notion of *local* model mis-specification for the state aggregation scenario. It is much smaller and more favorable than the notion of model mis-specification defined in Jin et al. (2020b); Zanette et al. (2020). Is there any counterpart for the corruption setting? If there is, how can we achieve robustness under such notion without prior knowledge?

Acknowledgments

The authors would like to thank Liyu Chen and Thodoris Lykouris for helpful discussions.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Conference on Neural Information Processing Systems*, 2011.

Yasin Abbasi-Yadkori, Peter L Bartlett, and Csaba Szepesvári. Online learning in markov decision processes with adversarially chosen transition probability distributions. In *Conference on Neural Information Processing Systems*, 2013.

Yasin Abbasi-Yadkori, Aldo Pacchiano, and My Phan. Regret balancing for bandit and rl model selection. In *Conference on Neural Information Processing Systems*, 2020.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, 2017.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. In *Conference on Neural Information Processing Systems*, 2020a.

Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. `https://rltheorybook.github.io/rltheorybook_AJKS.pdf`, 2020b.

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, 2016.

Pranjal Awasthi, Sreenivas Gollapudi, Kostas Kollias, and Apaar Sadhwani. Online learning under adversarial corruptions. https://openreview.net/forum?id=gBpYGXH9J7F, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In *Conference on Neural Information Processing Systems*, 2017.

Ilija Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, 2012.

Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, 2021.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Online and distribution-free robustness: Regression and contextual bandits with huber contamination. In *Symposium on Foundations of Computer Science*, 2021a.

Yifang Chen, Simon S. Du, and Kevin Jamieson. Improved corruption robust algorithms for episodic reinforcement learning. In *International Conference on Machine Learning*, 2021b.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, 2020.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, 2021.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Conference on Neural Information Processing Systems*, 2015.

Christoph Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, and Karthik Sridharan. Reinforcement learning with feedback graphs. In *Conference on Neural Information Processing Systems*, 2020.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. *ACM-SIAM Symposium on Discrete Algorithms*, 2019.

Travis Dick, Andras Gyorgy, and Csaba Szepesvari. Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, 2014.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *Conference on Neural Information Processing Systems*, 2019.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. In *Conference on Neural Information Processing Systems*, 2020.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, 2019.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly optimal regret for learning adversarial mdps with linear function approximation. *arXiv preprint arXiv:2102.08940*, 2021.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, 2021.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Conference on Neural Information Processing Systems*, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020b.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. In *Conference on Neural Information Processing Systems*, 2021a.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. In *Conference on Neural Information Processing Systems*, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. In *Conference on Neural Information Processing Systems*, 2021b.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, 2012.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. In *Conference on Neural Information Processing Systems*, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, 2021.

Yingkai Li, Edmund Y Lou, and Liren Shan. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. In *Conference on Neural Information Processing Systems*, 2021.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *ACM SIGACT Symposium on Theory of Computing*, 2018.

Thodoris Lykouris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Corruption robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, 2021.

Gergely Neu and Julia Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. *arXiv preprint arXiv:2007.01612*, 2020.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Conference on Neural Information Processing Systems*, 2010a.

Gergely Neu, András György, Csaba Szepesvári, et al. The online loop-free stochastic shortest-path problem. In *Conference on Learning Theory*, 2010b.

Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020a.

Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. In *Conference on Neural Information Processing Systems*, 2020b.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, 2019.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. In *AAAI Conference on Artificial Intelligence*, 2021.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, 2014.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Conference on Neural Information Processing Systems*, 2019.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In *International Conference on Machine Learning*, 2021.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Conference on Neural Information Processing Systems*, 2020.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, 2018.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on Learning Theory*, 2021.

Tianhao Wu, Yunchang Yang, Simon Du, and Liwei Wang. On reinforcement learning with adversarial corruption and its application to block mdp. In *International Conference on Machine Learning*, 2021.

Yihong Wu. Lecture 14: Packing, covering, and consequences on minimax risk. `http://www.stat.yale.edu/~yw562/teaching/598/lec14.pdf`, 2016.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *Internation Conference on Machine Learning*, 2020.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, 2021.

Xuezhou Zhang, Yiding Chen, Jerry Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. *arXiv preprint arXiv:2106.06630*, 2021a.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Robust policy gradient against strong data corruption. In *International Conference on Machine Learning*, 2021b.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S. Du. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. In *Conference on Neural Information Processing Systems*, 2021c.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, 2019.

Appendix A. The Non-robustness of Least Square Regression

In this section, we show that for linear contextual bandits with non-i.i.d. contexts, the most natural extension from the standard OFUL algorithm to a corruption-robust version results in a regret bound of $\Omega(\sqrt{C^{\mathsf{a}}T})$, even if C^{a} is known. See Section 3 for the definition of the linear contextual bandit framework that we consider. Recall that in the standard OFUL algorithm (Abbasi-Yadkori et al., 2011), the learner constructs a confidence set for the underlying parameter:

$$\mathcal{W}_t = \left\{ w \;:\; \|w - \widehat{w}_t\|_{\Lambda_t}^2 \le \iota_t \right\} \tag{14}$$

for some $\iota_t > 0$, where

$$\Lambda_t = \lambda I + \sum_{\tau=1}^{t-1} a_\tau a_\tau^\top, \qquad \widehat{w}_t = \Lambda_t^{-1} \left(\sum_{\tau=1}^{t-1} a_\tau r_\tau \right)$$

for some hyper-parameter $\lambda > 0$ (a_τ is the action taken at round τ, and r_τ is the reward received at round t). The action chosen at round t is

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}_t} \max_{w \in \mathcal{W}_t} a^\top w. \tag{15}$$

To make this algorithm robust to corruption, a natural modification is to widen the confidence set Eq. (14). That is, the confidence set is changed to

$$\mathcal{W}_t = \left\{ w \;:\; \|w - \widehat{w}_t\|_{\Lambda_t}^2 \le \iota_t' \right\} \tag{16}$$

for some $\iota_t' > \iota_t$. The definition of ι_t' may involve the knowledge of C^{a}. Below we show a regret lower bound for this class of algorithms.

We consider the following example for $d = 1$. The action set in each round is the following:

$$\mathcal{A}_t = \begin{cases} \{-1, 1\} & \text{if } 1 \le t \le C \\ \{-\epsilon, \epsilon\} & \text{if } t > C \end{cases}$$

for some $C \in \mathbb{N}$. The true underlying parameter is $w^\star = 1$, but in rounds $1, 2, \ldots, C$, the rewards are generated using $w' = -1$. We assume that there is no noise, i.e., $r_t = a_t w'$ for $t \in [1, C]$ and $r_t = a_t w^\star$ for $t > C$.

In this case, the confidence set Eq. (16) can be written as

$$\mathcal{W}_t = \left\{ w \;:\; (w - \widehat{w}_t)^2 \leq \frac{\iota'_t}{\Lambda_t} \right\} \tag{17}$$

where $\Lambda_t = \lambda + \sum_{\tau=1}^{t-1} a_\tau^2$ and

$$\widehat{w}_t = \frac{1}{\Lambda_t} \sum_{\tau=1}^{t-1} a_\tau r_\tau.$$

By the reward generation process and the definition of action sets, we have that

$$\begin{aligned} \sum_{\tau=1}^{t-1} a_\tau r_\tau &= \sum_{\tau=1}^{t-1} a_\tau \left(\mathbf{1}[\tau \leq C] a_\tau w' + \mathbf{1}[\tau > C] a_\tau w^\star \right) \\ &= \begin{cases} (t-1)w' = -t+1 & \text{if } t \leq C \\ Cw' + (t-1-C)\epsilon^2 w^\star = -C + (t-1-C)\epsilon^2 & \text{if } t > C \end{cases} \end{aligned}$$

Therefore, $\widehat{w}_t < 0$ for $2 \leq t \leq C\left(1 + \frac{1}{\epsilon^2}\right)$. Since the confidence set $\mathcal{W}_t$ (Eq. (17)) is symmetric around $\widehat{w}_t$, by the action selection rule Eq. (15), when $\widehat{w}_t < 0$, the learner will choose action -1 if $t \leq C$, and $-\epsilon$ if $t > C$.

Therefore, the learner will choose sub-optimal actions in $2 \leq t \leq C\left(1 + \frac{1}{\epsilon^2}\right)$, and the regret is of order

$$\begin{aligned} &\sum_{t=2}^{C} w^\star(1 - (-1)) + \sum_{t=C+1}^{\min\{C(1+1/\epsilon^2), T\}} w^\star(\epsilon - (-\epsilon)) \\ &= (C-1) \times 2 + \min\left\{ \frac{C}{\epsilon^2}, T - C \right\} \times 2\epsilon = \Theta\left(C + \min\left\{ \frac{C}{\epsilon}, \epsilon T \right\} \right). \end{aligned}$$

By picking $\epsilon = \sqrt{\frac{C}{T}}$, we see that the regret is at least of order $\sqrt{CT}$. Finally, notice that in the example we construct $C^{\mathsf{a}} = \Theta(C)$, hence proving our claim.

Appendix B. Concentration Inequalities

Lemma 9 (Freedman's inequality, Theorem 1 of (Beygelzimer et al., 2011)) *Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ be real random variables such that X_i is $\mathcal{F}_i$-measurable, $\mathbb{E}[X_i | \mathcal{F}_{i-1}] = 0$, $|X_i| \leq b$, and $\sum_{i=1}^n \mathbb{E}[X_i^2 | \mathcal{F}_{i-1}] \leq V$ for some fixed $b \geq 0$ and $V \geq 0$. Then with probability at least $1 - \delta$,*

$$\sum_{i=1}^n X_i \leq 2\sqrt{V \log(1/\delta)} + b \log(1/\delta).$$

Lemma 10 (Freedman's inequality, Lemma 4.4 of (Bubeck and Slivkins, 2012)) *Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ be real random variables such that X_i is $\mathcal{F}_i$-measurable, $\mathbb{E}[X_i | \mathcal{F}_{i-1}] = 0$, $|X_i| \leq b$, for some fixed $b \geq 0$. Let $V_n = \sum_{i=1}^n \mathbb{E}[X_i^2 | \mathcal{F}_{i-1}]$. Then with probability at least $1 - \delta$,*

$$\sum_{i=1}^n X_i \leq 2\sqrt{V_n \log(n/\delta)} + 3b \log(n/\delta).$$

Lemma 11 *Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_T$ be a filtration, and $X_1, \ldots, X_T$ be real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$, $|X_t| \le b$, for some fixed $b \ge 0$. Let $z_t \sim$ Bernoulli(α) be an i.i.d. random variable independent of all other variables, and let $0 \le y_t \le b$ be a deterministic scalar given $\mathcal{F}_{t-1}$. Then with probability at least $1-\delta$, the following holds for all $\mathcal{I} = [t_1, t_2] \subseteq [1, T]$:*

$$\left|\sum_{t\in\mathcal{I}} y_t(z_t - \alpha)\right| \le \min\left\{4\sqrt{\alpha\sum_{t\in\mathcal{I}} y_t^2 \log(T/\delta)} + 9b\log(T/\delta),\quad \frac{1}{4}\alpha\sum_{t\in\mathcal{I}} y_t + 21b\log(T/\delta)\right\}.$$

Proof Fixing an interval $\mathcal{I} \in [1, T]$, we apply Lemma 10 with $X_t = y_t(z_t - \alpha)$. Then we get that with probability at least $1 - 2\delta'$,

$$\begin{aligned}
\left|\sum_{t\in\mathcal{I}} y_t(z_t - \alpha)\right| &\le 2\sqrt{\sum_{t\in\mathcal{I}} y_t^2 \mathbb{E}_t[(z_t-\alpha)^2]\log(T/\delta')} + 3b\log(T/\delta') \quad (\text{define } \mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}]) \\
&\le 2\sqrt{\alpha\sum_{t\in\mathcal{I}} y_t^2 \log(T/\delta')} + 3b\log(T/\delta') \\
&\qquad (\mathbb{E}_t[(z_t-\alpha)^2] = \alpha(1-\alpha)^2 + (1-\alpha)\alpha^2 \le \alpha) \\
&\le 2\sqrt{b\log(T/\delta')}\sqrt{\alpha\sum_{t\in\mathcal{I}} y_t} + 3b\log(T/\delta') \qquad (|y_t| \le b) \\
&\le \frac{1}{4}\alpha\sum_{t\in\mathcal{I}} y_t + 7b\log(T/\delta') \qquad (\text{AM-GM})
\end{aligned}$$

Notice that there are $\frac{T(T-1)}{2}$ different $\mathcal{I}$'s, so we pick $\delta' = \frac{\delta}{T(T-1)}$, and take an union bound over $\mathcal{I}$'s. This gives the desired bound.

■

Appendix C. Omitted Proofs in Section 4

We start with some extra notations to be used in this section.

Definition 12 *For any time t, base algorithm i, and policy π, define $C^{\mathsf{a}}_{t,i} \triangleq \sum_{\tau=1}^{t} \mathbf{1}[i_\tau = i]c_\tau$ and $C^{\mathsf{r}}_{t,i} \triangleq \sqrt{\left(\sum_{\tau=1}^{t}\mathbf{1}[i_\tau = i]\right)\left(\sum_{\tau=1}^{t}\mathbf{1}[i_\tau = i]c_\tau^2\right)}$. Similarly, when we write $C_{t,i}$ to indicate either $C^{\mathsf{a}}_{t,i}$ or $C^{\mathsf{r}}_{t,i}$, depending on the type of base algorithms we use.*

Definition 13 *For any time t, base algorithm i, and policy π, define $R^{\pi}_{t,i} = \sum_{\tau=1}^{t}\mathbf{1}[i_\tau = i]r^{\pi}_\tau$.*

Next, we prove some lemmas to be used in the later analysis.

Lemma 14 *In* BASIC *(Algorithm 1), for any fixed i, with probability at least $1 - 2\delta$, the following holds for all t:*

$$\frac{3}{4}\alpha_i t - 21\log(T/\delta) \le N_{t,i} \le \frac{5}{4}\alpha_i t + 21\log(T/\delta).$$

Proof This is by directly applying Lemma 11 with $y_t = 1$ and $\alpha = \alpha_i$ ■

Lemma 15 *For any fixed i, with probability at least $1 - 3\delta$, the following holds for all t:*

$$C^{\mathsf{a}}_{t,i} \le 1.25\alpha_i C^{\mathsf{a}}_t + 21c_{\max}\log(T/\delta),$$
$$C^{\mathsf{r}}_{t,i} \le 1.25\alpha_i C^{\mathsf{r}}_t + 8c_{\max}\sqrt{\alpha_i t\log(T/\delta)} + 21c_{\max}\log(T/\delta).$$

Proof We prove the lemma for $(C_{t,i}, C) = (C^{\mathsf{a}}_{t,i}, C^{\mathsf{a}})$ and $(C_{t,i}, C) = (C^{\mathsf{r}}_{t,i}, C^{\mathsf{r}})$ cases separately.

Case 1. $(C_{t,i}, C) = (C^{\mathsf{a}}_{t,i}, C^{\mathsf{a}})$.

$$C_{t,i} = \sum_{\tau=1}^{t} c_\tau \mathbf{1}[i_\tau = i] \le \frac{5}{4}\alpha_i C + 21c_{\max}\log(T/\delta).$$
(holds w.p. $\ge 1-\delta$ by Lemma 11 with $y_\tau = c_\tau$, $z_\tau = \mathbf{1}[i_\tau = i]$)

Case 2. $(C_{t,i}, C) = (C^{\mathsf{r}}_{t,i}, C^{\mathsf{r}})$.

$$\begin{aligned}
C_{t,i} &= \sqrt{\left(\sum_{\tau=1}^{t}\mathbf{1}[i_\tau = i]\right)\left(\sum_{\tau=1}^{t}\mathbf{1}[i_\tau = i]c_\tau^2\right)} \\
&\le \sqrt{\left(\frac{5}{4}\alpha_i t + 21\log(T/\delta)\right)\left(\frac{5}{4}\alpha_i\sum_{\tau=1}^{t}c_\tau^2 + 21c_{\max}^2\log(T/\delta)\right)} \\
&\qquad\qquad \text{(holds w.p. } \ge 1-2\delta \text{ by Lemma 11 with } y_\tau = 1 \text{ and } y_\tau = c_\tau^2) \\
&\le \sqrt{\frac{25}{16}\alpha_i^2 t\sum_{\tau=1}^{t}c_\tau^2 + 52.5\alpha_i t c_{\max}^2\log(T/\delta) + 21^2 c_{\max}^2\log^2(T/\delta)} \\
&\le \frac{5}{4}\alpha_i\sqrt{t\sum_{\tau=1}^{t}c_\tau^2} + 8c_{\max}\sqrt{\alpha_i t\log(T/\delta)} + 21c_{\max}\log(T/\delta) \\
&\qquad\qquad (\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}) \\
&= \frac{5}{4}\alpha_i C + 8c_{\max}\sqrt{\alpha_i t\log(T/\delta)} + 21c_{\max}\log(T/\delta).
\end{aligned}$$

■

Lemma 16 *For any i, with probability at least $1 - \mathcal{O}(\delta)$, the following holds for all t such that $C_t \le 2^i$:*

$$R^{\pi^\star}_{t,i} - R_{t,i} \le \mathcal{R}(N_{t,i}, \theta_i).$$

Proof The total amount of corruption experienced by ALG_i up to round t is $C_{t,i}$, whose upper bound is given in Lemma 15 for both types of base algorithms. Comparing the upper bounds of $C_{t,i}$ with our choice of θ_i in Eq. (3), we see that for a fixed i, under the condition $C_t \le 2^i$, we have $C_{t,i} \le \theta_i$

with probability $1-\mathcal{O}(\delta)$. In other words, the condition specified in Assumption 2 is satisfied for ALG_i in the rounds that it is executed. Therefore, by the regret bound in Assumption 2, we have

$$R_{t,i}^{\pi^\star} - R_{t,i} = \sum_{\tau=1}^{t}(r_\tau^{\pi^\star} - r_\tau)\mathbf{1}[i_\tau = i] \le \mathcal{R}(N_{t,i}, \theta_i).$$

■

Lemma 17 *For any fixed i, with probability at least $1-\delta$, the following holds for all t:*

$$\left|\frac{1}{\alpha_i}R_{t,i}^{\pi^\star} - \sum_{\tau=1}^{t} r_{t,i}^{\pi^\star}\right| \le 2\sqrt{\frac{t\log(T/\delta)}{\alpha_i}} + \frac{\log(T/\delta)}{\alpha_i}.$$

Proof By Lemma 9, for a fixed i, with probability $1-\delta$, for all t,

$$\left|\frac{1}{\alpha_i}R_{t,i}^{\pi^\star} - \sum_{\tau=1}^{t} r_{t,i}^{\pi^\star}\right| = \left|\sum_{\tau=1}^{t}\left(\frac{\mathbf{1}[i_\tau = i]}{\alpha_i} - 1\right) r_\tau^{\pi^\star}\right| \le 2\sqrt{\frac{t\log(T/\delta)}{\alpha_i}} + \frac{\log(T/\delta)}{\alpha_i}.$$

■

Proof of Lemma 1. Notice that $C_t \le 2^k$ implies that $C_t \le 2^i$ for all $i \in [k, k_{\max}]$. Notice that

$$\begin{aligned}
R_{t,i} - R_{t,i}^{\pi^\star} &= \sum_{\tau=1}^{t}\mathbf{1}[i_\tau = i](r_\tau - r_\tau^{\pi^\star}) \\
&= \sum_{\tau=1}^{t}\mathbf{1}[i_\tau = i]\left(r_\tau - \mu^{\pi_\tau}(x_\tau) + \underbrace{\mu^{\pi_\tau}(x_\tau) - \mu^{\pi^\star}(x_\tau)}_{\le 0 \text{ (by Assumption 1)}} + \mu^{\pi^\star}(x_\tau) - r_\tau^{\pi^\star}\right) \\
&\le \sum_{\tau=1}^{t}\mathbf{1}[i_\tau = i]\left(r_\tau - \mu_\tau^{\pi_\tau}(x_\tau) + \mu_\tau^{\pi^\star}(x_\tau) - r_\tau^{\pi^\star}\right) + 2C_{t,i}^{\mathsf{a}} \\
&\qquad\qquad (|\mu_\tau^{\pi}(x_\tau) - \mu^{\pi}(x_\tau)| \le c_\tau \text{ for all } \pi \text{ and } \tau) \\
&\le 2\sqrt{2\alpha_i t\log(T/\delta)} + 6\log(T/\delta) + 2C_{t,i} \\
&\qquad\qquad \text{(by Lemma 10 with an union bound over } t\text{, and that } C_{t,i}^{\mathsf{a}} \le C_{t,i}^{\mathsf{r}}) \\
&\le 2\sqrt{2\alpha_i t\log(T/\delta)} + 6\log(T/\delta) + 2\theta_i. \\
&\qquad\qquad (C_{t,i} \le \theta_i \text{ with high probability by Lemma 15})
\end{aligned} \tag{18}$$

Combining Lemma 16 and Lemma 17, we see that the performance of ALG_i admits the following lower bound with probability at least $1-\mathcal{O}(\delta)$:

$$\frac{R_{t,i}}{\alpha_i} \ge \frac{R_{t,i}^{\pi^\star} - \mathcal{R}(N_{t,i},\theta_i)}{\alpha_i} \ge \sum_{\tau=1}^{t} r_\tau^{\pi^\star} - \frac{\mathcal{R}(N_{t,i},\theta_i)}{\alpha_i} - 2\sqrt{\frac{t\log(T/\delta)}{\alpha_i}} - \frac{\log(T/\delta)}{\alpha_i}. \tag{19}$$

Combining Eq. (18) and Lemma 17, we also have the following with probability at least $1 - \mathcal{O}(\delta)$:

$$\begin{aligned}\frac{R_{t,i}}{\alpha_i} &\leq \frac{R^{\pi^\star}_{t,i}}{\alpha_i} + 3\sqrt{\frac{t\log(T/\delta)}{\alpha_i}} + \frac{6\log(T/\delta) + 2\theta_i}{\alpha_i} \\ &\leq \sum_{\tau=1}^{t} r^{\pi^\star}_{\tau} + 5\sqrt{\frac{t\log(T/\delta)}{\alpha_i}} + \frac{7\log(T/\delta) + 2\theta_i}{\alpha_i}. \end{aligned} \tag{20}$$

The bounds Eq. (19) and Eq. (20) together with an union bound over i's indicate that the following holds for all $i, j \in [k, k_{\max}]$ with probability $1 - \mathcal{O}(k_{\max}\delta)$:

$$\frac{R_{t,i}}{\alpha_i} + \frac{\mathcal{R}(N_{t,i}, \theta_i)}{\alpha_i} + 2\sqrt{\frac{t\log(T/\delta)}{\alpha_i}} + \frac{\log(T/\delta)}{\alpha_i} \geq \sum_{\tau=1}^{t} r^{\pi^\star}_{\tau} \geq \frac{R_{t,j}}{\alpha_j} - 5\sqrt{\frac{t\log(T/\delta)}{\alpha_j}} - \frac{7\log(T/\delta) + 2\theta_j}{\alpha_j}.$$

Further combined with the fact that $\alpha_i \geq \alpha_j$ since $i \leq j$, the last inequality implies that the termination condition Eq. (4) will not hold.

■

Proof of Lemma 2.

$$\sum_{t=1}^{L_0}\left(r^{\pi^\star}_t - r_t\right) \leq 1 + \sum_{i=k}^{k_{\max}}\sum_{t=1}^{L_0-1}\left(r^{\pi^\star}_t - r_t\right)\mathbf{1}[i_t = i] = 1 + \sum_{i=k}^{k_{\max}}\left(R^{\pi^\star}_{L_0-1,i} - R_{L_0-1,i}\right). \tag{21}$$

For $i \geq i^\star$, since the corruption level is well-specified, by Lemma 16, with probability at least $1 - \mathcal{O}(\delta)$,

$$R^{\pi^\star}_{L_0-1,i} - R_{L_0-1,i} \leq \mathcal{R}(N_{L_0-1,i}, \theta_i). \tag{22}$$

For $i < i^\star$, with probability $1 - \mathcal{O}(\delta)$,

$$\begin{aligned}\frac{R_{L_0-1,i}}{\alpha_i} &\geq \frac{R_{L_0-1,i^\star}}{\alpha_{i^\star}} - \frac{\mathcal{R}(N_{L_0-1,i}, \theta_i)}{\alpha_i} - \mathcal{O}\left(\sqrt{\frac{(L_0-1)\log(T/\delta)}{\alpha_{i^\star}}} + \frac{\theta_{i^\star} + \log(T/\delta)}{\alpha_{i^\star}}\right) \\ &\qquad\qquad\text{(by the termination condition Eq. (4))} \\ &\geq \frac{R^{\pi^\star}_{L_0-1,i^\star}}{\alpha_{i^\star}} - \frac{\mathcal{R}(N_{L_0-1,i}, \theta_i)}{\alpha_i} - \widetilde{\mathcal{O}}\left(\sqrt{\frac{L_0}{\alpha_{i^\star}}} + \frac{\mathcal{R}(N_{L_0-1,i^\star}, \theta_{i^\star})}{\alpha_{i^\star}}\right) \\ &\qquad\qquad\text{(by Lemma 16 and that } \mathcal{R}(\cdot,\theta) \geq \theta\text{)} \\ &\geq \frac{R^{\pi^\star}_{L_0-1,i}}{\alpha_{i^\star}} - \frac{\mathcal{R}(N_{L_0-1,i}, \theta_i)}{\alpha_i} - \widetilde{\mathcal{O}}\left(\sqrt{\frac{L_0}{\alpha_{i^\star}}} + \frac{\mathcal{R}(N_{L_0-1,i}, \theta_{i^\star})}{\alpha_{i^\star}}\right) \end{aligned} \tag{23}$$

where the last inequality is because by Lemma 17 we have

$$\left|\frac{1}{\alpha_i}R^{\pi^\star}_{L_0-1,i} - \frac{1}{\alpha_{i^\star}}R^{\pi^\star}_{L_0-1,i^\star}\right| \leq \widetilde{\mathcal{O}}\left(\sqrt{\frac{L_0}{\alpha_i}} + \frac{1}{\alpha_i} + \sqrt{\frac{L_0}{\alpha_{i^\star}}} + \frac{1}{\alpha_{i^\star}}\right) = \widetilde{\mathcal{O}}\left(\sqrt{\frac{L_0}{\alpha_{i^\star}}} + \frac{1}{\alpha_{i^\star}}\right).$$

Combining Eq. (23) with Eq. (21) and Eq. (22) and an union bound over i's, we get that with probability at least $1 - \mathcal{O}(k_{\max}\delta)$,

$$\begin{aligned}
&\sum_{t=1}^{L_0}\left(r_t^{\pi^\star} - r_t\right) \\
&\leq 1 + \sum_{i=k}^{k_{\max}} \mathcal{R}(N_{L_0-1,i}, \theta_i) + \sum_{i<i^\star} \alpha_i \times \widetilde{\mathcal{O}}\left(\sqrt{\frac{L_0}{\alpha_{i^\star}}} + \frac{\mathcal{R}(N_{L_0-1,i^\star}, \theta_{i^\star})}{\alpha_{i^\star}}\right) \\
&\leq 1 + \sum_{i=k}^{k_{\max}} \mathcal{R}(N_{L_0-1,i}, \theta_i) + \widetilde{\mathcal{O}}\left(\mathbf{1}[i^\star > k]\left(\sqrt{\frac{L_0}{\alpha_{i^\star}}} + \frac{\mathcal{R}(N_{L_0-1,i^\star}, \theta_{i^\star})}{\alpha_{i^\star}}\right)\right) \qquad (24)
\end{aligned}$$

where in the last inequality we use $\sum_{i<i^\star} \alpha_i \leq 1$.

■

Proof of Theorem 3. Recall that we define

$$Z = \begin{cases} c_{\max} & \text{if ALG is type-a,} \\ c_{\max}\sqrt{T} & \text{if ALG is type-r.} \end{cases}$$

Let $i^\star$ be the smallest $i \in [k, k_{\max}]$ such that $C \leq 2^i$. By Lemma 2 and by the choice of θ_i in Eq. (3), with probability at least $1 - \mathcal{O}(k_{\max}\delta)$, the regret within an epoch is upper bounded by

$$\begin{aligned}
&\widetilde{\mathcal{O}}\left(\sum_{i=k}^{k_{\max}}\left(\sqrt{\beta_1\alpha_i T} + \beta_2\alpha_i 2^i + \beta_2 Z + \beta_3\right) + \mathbf{1}[k < i^\star]\left(\sqrt{\frac{T}{\alpha_{i^\star}}} + \frac{\sqrt{\beta_1\alpha_{i^\star}T} + \beta_2\alpha_{i^\star}2^{i^\star} + \beta_2 Z + \beta_3}{\alpha_{i^\star}}\right)\right) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 T} + \beta_2 2^k + \beta_2 z + \beta_3 + \mathbf{1}[k < i^\star]\left(\sqrt{\frac{\beta_1 T}{\alpha_{i^\star}}} + \beta_2 2^{i^\star} + \frac{\beta_2 z + \beta_3}{\alpha_{i^\star}}\right)\right) \\
&\hspace{20em} (\alpha_i 2^i \leq 2^k \text{ by the choice of } \alpha_i) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 T} + \beta_2(2^k + 2^{i^\star}) + \beta_2 Z + \beta_3 + \sqrt{\frac{\beta_1 T \cdot 2^{i^\star}}{2^k}} + \frac{(\beta_2 Z + \beta_3)2^{i^\star}}{2^k}\right) \\
&\hspace{20em} (\text{by the choice of } \alpha_{i^\star}) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 T} + \beta_2(2^k + 2^{i^\star}) + \beta_2 Z + \beta_3 + \beta_2 2^{i^\star} + \frac{\beta_1 T}{\beta_2 2^k} + \beta_2 2^{i^\star} \times \frac{\beta_2 Z + \beta_3}{\beta_2 2^k}\right) \quad (\text{AM-GM}) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 T} + \beta_2(2^k + 2^{i^\star}) + \beta_2 Z + \beta_3\right). \\
&\hspace{12em} (\text{using } \beta_2 2^k \geq \sqrt{\beta_1 T} + \beta_2 Z + \beta_3 \text{ by the choice of } k_{\text{init}})
\end{aligned}$$

Notice that in COBE we start from $k = k_{\text{init}}$. If $k_{\text{init}} \geq i^\star$, then by Lemma 1, the algorithm will run with $k = k_{\text{init}}$ throughout all T rounds. On the other hand, if $k_{\text{init}} < i^\star$, the k used in COBE might increase from k_{init}. However, if $k = i^\star$ is ever reached, again by Lemma 1, the learner will use this k throughout the rest of the steps. In short, the k's used in COBE are upper bounded by $\max\{k_{\text{init}}, i^\star\}$ with high probability. Since there are at most $\mathcal{O}(k_{\max})$ epochs, the overall regret is upper bounded by

$$\mathcal{O}(k_{\max}) \times \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 T} + \beta_2(2^{\max\{k_{\text{init}}, i^\star\}} + 2^{i^\star}) + \beta_2 Z + \beta_3\right) = \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 T} + \beta_2(C + Z) + \beta_3\right)$$

with probability at least $1-\mathcal{O}(k_{\max}^2\delta)$ Considering the difference definitions of Z for type-a and type-r base algorithms finishes the proof. ■

Appendix D. Omitted Proofs in Section 5.2

Proof of Lemma 5. By Lemma 2,

$$
\begin{aligned}
&\sum_{t=1}^{L_0}(r_t^{\pi^\star}-r_t)\\
&\le \sum_{i=k}^{k_{\max}}\left(\min\left\{\sqrt{\beta_1 N_{L_0,i}},\, \frac{\beta_1}{\Delta}\right\}+\beta_2\alpha_i 2^i+\beta_2 c_{\max}+\beta_3\right)\\
&\qquad\qquad +\widetilde{\mathcal{O}}\left(\sqrt{\frac{L}{\alpha_{i^\star}}}+\frac{\min\left\{\sqrt{\beta_1 N_{L_0,i^\star}},\frac{\beta_1}{\Delta}\right\}+\beta_2\alpha_{i^\star}2^{i^\star}+\beta_2 c_{\max}+\beta_3}{\alpha_{i^\star}}\right)\\
&\le \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 L}+\beta_2(2^k+2^{i^\star})+\frac{\beta_2 c_{\max}+\beta_3}{\alpha_{i^\star}}+\sqrt{\frac{L}{\alpha_{i^\star}}}+\frac{\sqrt{\beta_1 N_{L_0,i^\star}}}{\alpha_{i^\star}}\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(by the definition of } \alpha_i,\ \alpha_i 2^i=\mathcal{O}(2^k+\sqrt{\beta_1 L}/\beta_2))\\
&=\widetilde{\mathcal{O}}\left(\beta_2(2^k+2^{i^\star})+\frac{\sqrt{\beta_1}+\beta_2 c_{\max}+\beta_3}{\alpha_{i^\star}}+\sqrt{\frac{\beta_1 L}{\alpha_{i^\star}}}\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(using } N_{L_0,i^\star}=\widetilde{\mathcal{O}}\left(\alpha_{i^\star}L_0+1\right) \text{ by Lemma 14)}\\
&=\widetilde{\mathcal{O}}\left(\beta_2(2^k+2^{i^\star})+\left(\sqrt{\beta_1}+\beta_2 c_{\max}+\beta_3\right)\left(1+\frac{2^{i^\star}}{\frac{\sqrt{\beta_1 L}}{\beta_2}+2^k}\right)+\sqrt{\beta_1 L\left(1+\frac{2^{i^\star}}{\frac{\sqrt{\beta_1 L}}{\beta_2}+2^k}\right)}\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(by the definition of } \alpha_{i^\star})\\
&=\widetilde{\mathcal{O}}\left(\beta_2(2^k+2^{i^\star})+\left(\sqrt{\beta_1}+\beta_2 c_{\max}+\beta_3\right)\left(1+\frac{\beta_2 2^{i^\star}}{\beta_2 2^k}\right)+\sqrt{\beta_1 L\left(1+\frac{\beta_2 2^{i^\star}}{\sqrt{\beta_1 L}}\right)}\right)\\
&=\widetilde{\mathcal{O}}\left(\sqrt{\beta_1 L}+\beta_2 2^{i^\star}+\beta_2 c_{\max}+\beta_3+\sqrt{\sqrt{\beta_1 L}\times\beta_2 2^{i^\star}}\right)\\
&\qquad\qquad (\beta_2 2^k\ge\beta_2 2^{k_{\text{init}}}\ge\sqrt{\beta_1}+\beta_2 c_{\max}+\beta_3 \text{ and } \beta_2 2^k\le\sqrt{\beta_1 L} \text{ as chosen in G-COBE})\\
&=\widetilde{\mathcal{O}}\left(\sqrt{\beta_1 L}+\beta_2 C_{L_0}+\beta_2 c_{\max}+\beta_3\right).\\
&\qquad\qquad\qquad \text{(AM-GM and } 2^{i^\star}=\mathcal{O}(\max\{C_{L_0},2^k\}) \text{ by the definition of } i^\star \text{ in Lemma 2)}
\end{aligned}
$$

■

Proof of Lemma 6. Since $2^k\ge 32C_{L_0}\ge 32C_t$ for all t during execution, by Lemma 1, with probability at least $1-\mathcal{O}(k_{\max}\delta)$, Eq. (4) will not hold. Therefore, BASIC will finish all L steps (thus $L_0 = L$) and return **true**. By Lemma 16, we have that for ALG_k, with probability $1 - \mathcal{O}(\delta)$,

$$\begin{aligned}
&\sum_{t=1}^{L}(r_t^{\pi^*} - r_t)\mathbf{1}[i_t = k] \\
&\leq \frac{\beta_1}{\Delta} + \beta_2\theta_k + \beta_3 \\
&\leq \frac{\beta_1}{\Delta} + \beta_2\left(\frac{5}{4} \times 2^k + 21c_{\max}\log(T/\delta)\right) + \beta_3 && \text{(by the definition of } \theta_k) \\
&\leq \frac{2\beta_4}{\Delta} + \frac{5}{4}\beta_2 2^k. && \text{(by the definition of } \beta_4 \text{ and that } \Delta \leq 1)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\sum_{t=1}^{L}(r_t^{\pi^*} - r_t)\mathbf{1}[i_t = k] \\
&\geq \sum_{t=1}^{L}(\mu^{\pi^*} - \mu^{\pi_t})\mathbf{1}[i_t = k] + \sum_{t=1}^{L}\left(r_t^{\pi^\star} - \mu_t^{\pi^\star}(x_t) + \mu_t^{\pi_t}(x_t) - r_t^{\pi_t}\right)\mathbf{1}[i_t = k] - 2\sum_{t=1}^{L}\max_{\pi}|\mu^{\pi} - \mu_t^{\pi}| \\
&\geq \Delta\left(\sum_{t=1}^{L}\mathbf{1}[\pi_t \neq \pi^*]\mathbf{1}[i_t = k]\right) - 8\sqrt{\sum_{t=1}^{L}\mathbf{1}[\pi_t \neq \pi^*]\mathbf{1}[i_t = k]\log(T/\delta)} - 18\log(T/\delta) - 2C_L \\
&\qquad \text{(by Lemma 11 with } y_t = \mathbf{1}[i_t = k]\mathbf{1}[\pi_t \neq \pi^\star],\ X_t = r_t^{\pi^\star} - \mu_t^{\pi^\star}(x_t) + \mu_t^{\pi_t}(x_t) - r_t^{\pi_t}) \\
&\geq \frac{3}{4}\Delta\left(\sum_{t=1}^{L}\mathbf{1}[\pi_t \neq \pi^*]\mathbf{1}[i_t = k]\right) - \frac{34\log(T/\delta)}{\Delta} - 2C_L. && \text{(AM-GM and that } \Delta \leq 1)
\end{aligned}$$

Combining the two inequalities above, and using that $\beta_1 \geq 5\log(T/\delta), \beta_2 \geq 1, \beta_3 \geq 5\log(T/\delta)$, we get

$$\begin{aligned}
&\sum_{t=1}^{L}\mathbf{1}[\pi_t \neq \pi^*]\mathbf{1}[i_t = k] \\
&\leq \frac{4}{3\Delta}\left(\frac{2\beta_4}{\Delta} + \frac{5}{4}\beta_2 2^k + \frac{34\log(T/\delta)}{\Delta} + 2C_L\right) \\
&\leq \frac{4}{3\Delta}\left(\frac{2\beta_4}{\Delta} + \frac{7\beta_4}{\Delta} + 2C_L\right) + \frac{5}{3\Delta}\beta_2 2^k && (\beta_4 \geq \beta_1 \geq 5\log(T/\delta)) \\
&\leq \frac{4}{3\Delta}\beta_2 2^k + \frac{5}{3\Delta}\beta_2 2^k && \text{(by the condition specified in the lemma and that } L = L_0) \\
&= \frac{3\beta_2 2^k}{\Delta}. && (25)
\end{aligned}$$

We also have

$$
\begin{aligned}
&\sum_{t=1}^{L} \mathbf{1}[i_t = k] \\
&\geq \frac{3}{4}\alpha_k L - 21\log(T/\delta) && \text{(by Lemma 14)} \\
&\geq \frac{3}{8}\frac{(\beta_2 2^k)^2}{\beta_4} - 5\beta_3 && (\alpha_k \geq \tfrac{1}{2} \text{ and } \sqrt{\beta_4 L} \geq \beta_2 2^k \text{ and } \beta_3 \geq 5\log(T/\delta)) \\
&\geq \frac{12\beta_2 2^k}{\Delta} - \beta_2 2^k && (\ \beta_2 2^k \geq \tfrac{32\beta_4}{\Delta} \geq 32\beta_3 \text{ by the condition specified in the lemma}) \\
&\geq \frac{11\beta_2 2^k}{\Delta}. && (\Delta \leq 1)
\end{aligned} \tag{26}
$$

Combining Eq. (25) and Eq. (26) proves the lemma. ■

Appendix E. Omitted Proofs in Section 5.3

For TwoModelSelect, we define the following notations:

Definition 18 *Let $\mathcal{E}_j$ be the set of rounds in epoch j, i.e., $\mathcal{E}_j \triangleq [t_j, t_{j+1} - 1]$. Let $\mathcal{E}_j'$ be the set of rounds in epoch j except for the last round, i.e., $\mathcal{E}_j' \triangleq [t_j, t_{j+1} - 2]$ (might be empty if $t_{j+1} = t_j + 1$).*

Definition 19 *Let $\mathcal{I} \subseteq \mathcal{E}_j$ be any interval in epoch j. Define*

$$
\begin{aligned}
\widehat{N}_{\mathcal{I},0} &\triangleq \sum_{t\in\mathcal{I}} \mathbf{1}[Y_t = 0], & \widehat{N}_{\mathcal{I},1} &\triangleq \sum_{t\in\mathcal{I}} \mathbf{1}[Y_t = 1], \\
\widehat{R}_{\mathcal{I},0} &\triangleq \frac{1}{1-p_j}\sum_{t\in\mathcal{I}} \mathbf{1}[Y_t = 0]r_t, & \widehat{R}_{\mathcal{I},1} &\triangleq \frac{1}{p_j}\sum_{t\in\mathcal{I}} \mathbf{1}[Y_t = 1]r_t, \\
\widehat{R}^{\pi}_{\mathcal{I},0} &\triangleq \frac{1}{1-p_j}\sum_{t\in\mathcal{I}} \mathbf{1}[Y_t = 0]r^{\pi}_t, & \widehat{R}^{\pi}_{\mathcal{I},1} &\triangleq \frac{1}{p_j}\sum_{t\in\mathcal{I}} \mathbf{1}[Y_t = 1]r^{\pi}_t.
\end{aligned}
$$

Definition 20 *With abuse of notations, define $C_{\mathcal{I}} \triangleq \sum_{\tau\in\mathcal{I}} c_\tau$ (recall that we also define $C_t = \sum_{\tau=1}^{t} c_\tau$).*

Definition 21 $\Theta_j \triangleq \frac{5}{4}p_j C_{\mathcal{E}_j} + 21c_{\max}\log(T/\delta)$.

Below, we first establish some basic lemmas:

Lemma 22 *Let $\mathcal{R}(t,\theta) = \sqrt{\beta_1 t} + \beta_2\theta + \beta_3$ with $\beta_3 \geq 10\sqrt{\beta_1\log(T/\delta)}$. Then with probability at least $1 - \mathcal{O}(\delta)$, the following holds for all $\mathcal{I} = [t_j, t] \subseteq \mathcal{E}_j$ and any $\theta \geq 0$:*

$$
\frac{1}{2}\mathcal{R}(p_j|\mathcal{I}|,\theta) \leq \mathcal{R}(\widehat{N}_{\mathcal{I},1},\theta) \leq \frac{3}{2}\mathcal{R}(p_j|\mathcal{I}|,\theta).
$$

Proof With probability at least $1-\mathcal{O}(\delta)$,

$$\begin{aligned}
\mathcal{R}(\widehat{N}_{\mathcal{I},1},\theta) &= \sqrt{\beta_1 \widehat{N}_{\mathcal{I},1}} + \beta_2\theta + \beta_3 \\
&\le \sqrt{\frac{5}{4}\beta_1 p_j|\mathcal{I}| + 21\beta_1\log(T/\delta)} + \beta_2\theta + \beta_3 \quad \text{(by the same argument as Lemma 14)} \\
&\le \frac{3}{2}\sqrt{\beta_1 p_j|\mathcal{I}|} + \beta_2\theta + 5\sqrt{\beta_1\log(T/\delta)} + \beta_3 \\
&\le \frac{3}{2}\mathcal{R}(p_j|\mathcal{I}|,\theta). \qquad (\beta_3 \ge 10\sqrt{\beta_1\log(T/\delta)})
\end{aligned}$$

If $p_j|\mathcal{I}| \ge 42\log(T/\delta)$, then

$$\begin{aligned}
\mathcal{R}(\widehat{N}_{\mathcal{I},1},\theta) &= \sqrt{\beta_1 \widehat{N}_{\mathcal{I},1}} + \beta_2\theta + \beta_3 \\
&\ge \sqrt{\frac{3}{4}\beta_1 p_j|\mathcal{I}| - 21\beta_1\log(T/\delta)} + \beta_2\theta + \beta_3 \quad \text{(by the same argument as Lemma 14)} \\
&\ge \frac{1}{2}\sqrt{\beta_1 p_j|\mathcal{I}|} + \beta_2\theta + \beta_3 \\
&\ge \frac{1}{2}\mathcal{R}(p_j|\mathcal{I}|,\theta);
\end{aligned}$$

otherwise, we have $p_j|\mathcal{I}| < 42\log(T/\delta)$ and

$$\begin{aligned}
\mathcal{R}(\widehat{N}_{\mathcal{I},1},\theta) &\ge \beta_2\theta + \beta_3 \\
&\ge 5\sqrt{\beta_1\log(T/\delta)} + \beta_2\theta + \frac{1}{2}\beta_3 \qquad (\beta_3 \ge 10\sqrt{\beta_1\log(T/\delta)}) \\
&\ge 5\sqrt{\beta_1 \times \frac{1}{42}p_j|\mathcal{I}|} + \beta_2\theta + \frac{1}{2}\beta_3 \\
&\ge \frac{1}{2}\mathcal{R}(p_j|\mathcal{I}|,\theta).
\end{aligned}$$

■

Lemma 23 *With probability at least $1-\mathcal{O}(\delta)$, for all interval $\mathcal{I}=[t_j,t]\subseteq\mathcal{E}_j$,*

$$\left|\widehat{R}_{\mathcal{I},1} - |\mathcal{I}|\mu^{\pi'}\right| \le \frac{3}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{I}|,\Theta_j\right), \tag{27}$$

$$\left|\widehat{R}_{\mathcal{I},0} - |\mathcal{I}|\mu^{\widehat{\pi}}\right| \le 4\sqrt{|\mathcal{I}|\log(1/\delta)} + 4\log(T/\delta) + C_{\mathcal{I}} \le \frac{1}{p_j}\mathcal{R}_{\mathcal{B}}(p_j|\mathcal{I}|,\Theta_j). \tag{28}$$

where $\pi' = \operatorname{argmax}_{\pi'\in\Pi\setminus\{\widehat{\pi}\}}\mu^{\pi'}$.

Proof By the same argument as Lemma 15, the corruption experienced by $\mathcal{B}_{\widehat{\pi}}$ in $\mathcal{I}$ is upper bounded by Θ_j. By the regret guarantee of $\mathcal{B}_{\widehat{\pi}}$, we have with probability at least $1-\mathcal{O}(\delta)$,

$$\begin{aligned}
\widehat{R}_{\mathcal{I},1} &\geq \widehat{R}^{\pi'}_{\mathcal{I},1} - \frac{1}{p_j}\mathcal{R}_{\mathcal{B}}\left(\widehat{N}_{\mathcal{I},1},\Theta_j\right) \\
&\geq \sum_{t\in\mathcal{I}}\mu^{\pi'}_t - 2\sqrt{\frac{2|\mathcal{I}|\log(T/\delta)}{p_j}} - \frac{2\log(T/\delta)}{p_j} - \frac{2}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{I}|,\Theta_j\right) \\
&\hspace{12em}\text{(by Lemma 9 with a union bound over } |\mathcal{I}| \text{ and Lemma 22)} \\
&\geq |\mathcal{I}|\mu^{\pi'} - C_{\mathcal{I}} - 2\sqrt{\frac{2|\mathcal{I}|\log(T/\delta)}{p_j}} - \frac{2\log(T/\delta)}{p_j} - \frac{2}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{I}|,\Theta_j\right) \\
&\geq |\mathcal{I}|\mu^{\pi'} - \frac{3}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{I}|,\Theta_j\right). \\
&\hspace{10em}\text{(by the assumption } \beta_1\geq 8\log(T/\delta), \beta_2\geq 1, \beta_3\geq 5\log(T/\delta))
\end{aligned}$$

Again by Lemma 9 with a union bound over $|\mathcal{I}|$, we also have with probability $1-\mathcal{O}(\delta)$,

$$\begin{aligned}
\widehat{R}_{\mathcal{I},1} &\leq \sum_{t\in\mathcal{I}}\mu^{\pi'}_t + 2\sqrt{\frac{2|\mathcal{I}|\log(T/\delta)}{p_j}} + \frac{2\log(T/\delta)}{p_j} \\
&\leq |\mathcal{I}|\mu^{\pi'} + 2\sqrt{\frac{2|\mathcal{I}|\log(T/\delta)}{p_j}} + \frac{2\log(T/\delta)}{p_j} + C_{\mathcal{I}} \\
&\leq |\mathcal{I}|\mu^{\pi'} + \frac{1}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{I}|,\Theta_j\right).
\end{aligned}$$

Combining them, we get Eq. (27). The first inequality in Eq. (28) can be obtained by Lemma 9 with the fact that $1-p_j\geq\frac{1}{2}$ and $\left|\sum_{t\in\mathcal{I}}\left(\mu^{\pi'}_t-\mu^{\pi'}\right)\right|\leq C_{\mathcal{I}}$; the second inequality in Eq. (28) can be obtained using the assumptions on β_1,β_2,β_3. ■

Lemma 24 *For all j, $\widehat{\Delta}_j\leq 1$.*

Proof When $j=1$, $\widehat{\Delta}_1\leq 1$ by definition. Assume that $j+1\geq 2$ is the first j such that $\widehat{\Delta}_{j+1}>1$. By the way we update $\widehat{\Delta}_j$, it must be that $\widehat{\Delta}_j\geq\frac{1}{1.25}$ and that at the end of epoch j, Eq. (12) is triggered.

However, notice that in Eq. (12), the left-hand side $\widehat{R}_0=\frac{1}{1-p_j}\sum_{\tau=t_j}^{t}r_\tau\mathbf{1}[Y_\tau=0]\leq\frac{1}{1-p_j}M_j\leq 2M_j$ since $p_j=\frac{\beta_4}{2M_j\widehat{\Delta}_j^2}\leq\frac{1}{2}$ by Eq. (13), but the right-hand side of Eq. (12) involves a term $3M_j\widehat{\Delta}_j\geq 3M_j\times\frac{1}{1.25}>2M_j$. Therefore, Eq. (12) is impossible to be triggered at this j, contradicting our assumption. ■

Lemma 25 *With probability at least $1-\mathcal{O}(\delta)$, for all interval $\mathcal{I}=[t_j,t]\subseteq\mathcal{E}_j$,*

$$\frac{1}{p_j}\mathcal{R}_{\mathcal{B}}\left(\widehat{N}_{\mathcal{I},1},\Theta_j\right)\leq\frac{2}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{I}|,\Theta_j+\frac{p_j\sqrt{\beta_1 L}}{\beta_2}\right)\leq 0.02M_j\widehat{\Delta}_j+2.5\beta_2 C_{\mathcal{E}_j}\log(T/\delta)+2\sqrt{\beta_1 L}.$$

Proof

$$
\begin{aligned}
&\frac{1}{p_j}\mathcal{R}_{\mathcal{B}}\left(\widehat{N}_{\mathcal{I},1},\Theta_j\right) \\
&\le \frac{2}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{I}|,\Theta_j+\frac{p_j\sqrt{\beta_1 L}}{\beta_2}\right) && \text{(by Lemma 22)} \\
&= 2\left(\sqrt{\frac{\beta_1|\mathcal{I}|}{p_j}}+\frac{\beta_2\Theta_j}{p_j}+\frac{\beta_3}{p_j}\right)+2\sqrt{\beta_1 L} \\
&\le 2\left(\sqrt{\frac{2\beta_1}{\beta_4}|\mathcal{I}|M_j\widehat{\Delta}_j}+\frac{5}{4}\beta_2 C_{\mathcal{E}_j}+\frac{21\beta_2 c_{\max}\log(T/\delta)+\beta_3}{\beta_4}\times 2M_j\widehat{\Delta}_j^2\right)+2\sqrt{\beta_1 L} \\
&\le 0.02M_j\widehat{\Delta}_j+2.5\beta_2 C_{\mathcal{E}_j}+2\sqrt{\beta_1 L}. \quad \text{(by the definition of } \beta_4 \text{ and that } \widehat{\Delta}_j\le 1 \text{ by Lemma 24)}
\end{aligned}
$$

■

Proof of Lemma 7. Let T_0 be the round at which TwoModelSelect terminates. In the following proof, we assume that the high-probability events defined in previous lemmas hold.

Case 1. $\widehat{\pi}\neq\pi^\star$**.**

$$
\begin{aligned}
&\sum_{t\in\mathcal{E}_j'}(r_t^{\pi^\star}-r_t) \\
&= (1-p_j)\left(\widehat{R}^{\pi^\star}_{\mathcal{E}_j',0}-\widehat{R}^{\widehat{\pi}}_{\mathcal{E}_j',0}\right)+p_j\left(\widehat{R}^{\pi^\star}_{\mathcal{E}_j',1}-\widehat{R}_{\mathcal{E}_j',1}\right) \\
&\le (1-p_j)\left(\widehat{R}^{\pi^\star}_{\mathcal{E}_j',1}-\widehat{R}_{\mathcal{E}_j',0}\right)+(1-p_j)\left(2\sqrt{\frac{2|\mathcal{E}_j'|\log(T/\delta)}{p_j}}+\frac{2\log(T/\delta)}{p_j}\right)+p_j\left(\widehat{R}^{\pi^\star}_{\mathcal{E}_j',1}-\widehat{R}_{\mathcal{E}_j',1}\right) \\
&\qquad\qquad \text{(when } Y_t=0 \text{ we execute } \widehat{\pi}\text{, and by Lemma 9)} \\
&\le (1-p_j)\left(\widehat{R}_{\mathcal{E}_j',1}-\widehat{R}_{\mathcal{E}_j',0}\right)+\frac{1}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{E}_j'|,0\right)+\left(\widehat{R}^{\pi^\star}_{\mathcal{E}_j',1}-\widehat{R}_{\mathcal{E}_j',1}\right) \\
&\qquad\qquad \text{(by the assumptions on } \beta_1,\beta_3) \\
&\le \frac{5}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{E}_j'|,\frac{p_j\sqrt{\beta_1 L}}{\beta_2}\right)-\frac{1}{2}|\mathcal{E}_j'|\widehat{\Delta}_j+\frac{1}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{E}_j'|,0\right)+\frac{1}{p_j}\mathcal{R}_{\mathcal{B}}\left(\widehat{N}_{\mathcal{E}_j',1},\Theta_j\right) \\
&\qquad\qquad \text{(by Eq. (11), and Assumption 5 with the assumption that } \pi^\star\in\Pi\backslash\{\widehat{\pi}\}) \\
&\le \frac{8}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j|\mathcal{E}_j'|,\Theta_j+\frac{p_j\sqrt{\beta_1 L}}{\beta_2}\right)-\frac{1}{2}|\mathcal{E}_j|\widehat{\Delta}_j && \text{(by Lemma 22)} \\
&\le 0.16M_j\widehat{\Delta}_j+20\beta_2 C_{\mathcal{E}_j}+16\sqrt{\beta_1 L}-0.5|\mathcal{E}_j|\widehat{\Delta}_j. && \text{(by Lemma 25)}
\end{aligned}
$$

For $j\ge 2$, the last expression is further upper bounded by

$$
\begin{aligned}
&\left(0.32|\mathcal{E}_{j-1}|+\frac{\beta_4}{\widehat{\Delta}_j^2}\right)\widehat{\Delta}_j+20\beta_2 C_{\mathcal{E}_j}+16\sqrt{\beta_1 L}-0.5|\mathcal{E}_j|\widehat{\Delta}_j && \text{(by the definition of } M_j) \\
&\le \frac{\beta_4}{\widehat{\Delta}_j}+20\beta_2 C_{\mathcal{E}_j}+16\sqrt{\beta_1 L}+0.4|\mathcal{E}_{j-1}|\widehat{\Delta}_{j-1}-0.5|\mathcal{E}_j|\widehat{\Delta}_j && (\widehat{\Delta}_j\le 1.25\widehat{\Delta}_{j-1}) \\
&\le \widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L}+\beta_2 C_{\mathcal{E}_j}+\beta_4\right)+0.4|\mathcal{E}_{j-1}|\widehat{\Delta}_{j-1}-0.5|\mathcal{E}_j|\widehat{\Delta}_j; && \left(\widehat{\Delta}_j\ge\widehat{\Delta}_1=\min\left\{\sqrt{\tfrac{\beta_4}{L}},1\right\}\right)
\end{aligned}
$$

for $j=1$, it is upper bounded by

$$\frac{0.16\beta_4}{\widehat{\Delta}_1}+20\beta_2 C_{\mathcal{E}_1}+16\sqrt{\beta_1 L}-0.5|\mathcal{E}_1|\widehat{\Delta}_1=\widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L}+\beta_2 C_{\mathcal{E}_1}+\beta_4\right)-0.5|\mathcal{E}_1|\widehat{\Delta}_1.$$

Summing up the above bound over j (and noticing that the number of epochs is upper bounded by $3\log^2 T$), we see that

$$\begin{aligned}
&\sum_{t=1}^{T_0}(r_t^{\pi^\star}-r_t)\\
&\leq \widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L}+\beta_2 C_{\mathcal{E}_1}+\beta_4\right)-0.5|\mathcal{E}_1|\widehat{\Delta}_1+\sum_{j=2}^{3\log^2 T}\left(\widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L}+\beta_2 C_{\mathcal{E}_j}+\beta_4\right)+0.4|\mathcal{E}_{j-1}|\widehat{\Delta}_{j-1}-0.5|\mathcal{E}_j|\widehat{\Delta}_j\right)\\
&=\widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L}+\beta_2 C_{T_0}+\beta_4\right).
\end{aligned}$$

Case 2. $\widehat{\pi}=\pi^\star$**.**

$$\begin{aligned}
\sum_{t\in\mathcal{E}_j'}\left(r_t^{\pi^\star}-r_t\right)&=p_j\left(\widehat{R}_{t,1}^{\pi^\star}-\widehat{R}_{t,1}\right)\\
&\leq p_j\left(\widehat{R}_{t,0}^{\pi^\star}-\widehat{R}_{t,1}\right)+\widetilde{\mathcal{O}}\left(\sqrt{p_j|\mathcal{E}_j'|}+1\right) && \text{(Lemma 9)}\\
&\leq p_j\left(\widehat{R}_{t,0}-\widehat{R}_{t,1}\right)+\widetilde{\mathcal{O}}\left(\sqrt{p_j M_j}+1\right) && (\text{since } \widehat{\pi}=\pi^\star)\\
&\leq \widetilde{\mathcal{O}}\left(p_j M_j\widehat{\Delta}_j+\sqrt{\beta_1 L}+\sqrt{p_j M_j}\right) && \text{(by Eq. (12))}\\
&\leq \widetilde{\mathcal{O}}\left(\frac{\beta_4}{\widehat{\Delta}_j}+\sqrt{\beta_1 L}\right) && (\text{by the definition of } p_j)\\
&=\widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L}+\beta_4\right). && \left(\widehat{\Delta}_j\geq\widehat{\Delta}_1=\min\left\{\sqrt{\tfrac{\beta_4}{L}},1\right\}\right)
\end{aligned}$$

Similarly, summing over epochs and using the fact that the number of epochs is upper bounded by $\mathcal{O}(\log^2 T)$ we get the desired bound.

■

Proof of Lemma 8. The condition in the lemma implies $\Delta\geq 16\sqrt{\frac{\beta_4}{L}}=16\widehat{\Delta}_1$. Below we prove by induction that $\Delta\geq\widehat{\Delta}_j$ for all j. This holds for $j=1$. Notice that $\widehat{\Delta}_j$ only increases when the second break condition Eq. (12) holds. If Eq. (12) holds, we have

$$\begin{aligned}
|\mathcal{E}_j|\Delta&=|\mathcal{E}_j|(\mu^{\pi^*}-\mu^{\pi'})\\
&\geq \widehat{R}_{\mathcal{E}_j,0}-\widehat{R}_{\mathcal{E}_j,1}-\frac{4}{p_j}\mathcal{R}_{\mathcal{B}}\left(\widehat{N}_{\mathcal{E}_j,1},\Theta_j\right) && \text{(Lemma 23)}\\
&\geq 3M_j\widehat{\Delta}_j+9\sqrt{\beta_1 L}-4\left(0.02M_j\widehat{\Delta}_j+2.5\beta_2 C_{\mathcal{E}_j}+2\sqrt{\beta_1 L}\right).\\
& && \text{(by Eq. (12) and Lemma 25)}\\
&\geq 2.5M_j\widehat{\Delta}_j. && \left(\text{by the condition specified in the lemma, we have } \sqrt{\beta_1 L}\geq 10\beta_2 C_{\mathcal{E}_j}\right)
\end{aligned}$$

Because $|\mathcal{E}_j| \le M_j$, we have $\widehat{\Delta}_j \le \frac{1}{2.5}\Delta$. Therefore, after the update, $\widehat{\Delta}_{j+1} = 1.25\widehat{\Delta}_j \le \Delta$ still holds.

Next, we show that the first break condition Eq. (11) will not hold with high probability: at any time t within epoch j,

$$
\begin{aligned}
&\widehat{R}_{[t_j,t-1],0} - \widehat{R}_{[t_j,t-1],1} - \frac{1}{2}(t-t_j)\widehat{\Delta}_j \\
&= \left(\widehat{R}_{[t_j,t-1],0} - (t-t_j)\mu^{\pi^*}\right) + (t-t_j)\left(\mu^{\pi^*} - \mu^{\pi'} - \frac{1}{2}\widehat{\Delta}_j\right) + \left((t-t_j)\mu^{\pi'} - \widehat{R}_{[t_j,t],1}\right) \\
&\ge -\frac{4}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j(t-t_j), p_j C_{\mathcal{E}_j}\right) + (t-t_j)\left(\Delta - \frac{1}{2}\widehat{\Delta}_j\right) \qquad \text{(Lemma 23)} \\
&\ge -\frac{4}{p_j}\mathcal{R}_{\mathcal{B}}\left(p_j(t-t_j), \frac{p_j\sqrt{\beta_1 L}}{\beta_2}\right).
\end{aligned}
$$

($\beta_2 C_{\mathcal{E}_j} \le \sqrt{\beta_1 L}$ as assumed in the lemma; $\Delta \ge \widehat{\Delta}_j$ for all j as we just showed above)

Therefore, the first break condition will not be triggered. Overall, with high probability, $\widehat{\Delta}_j$ is non-decreasing with j.

Under this high-probability event, since $\widehat{\Delta}_j$ never decreases, the number of times $\widehat{\Delta}_j$ increases is upper bounded by $\log_{1.25}\frac{\Delta}{\widehat{\Delta}_1} \le \log_{1.25}\sqrt{\frac{L}{\beta_4}} \le \frac{1}{2}\log_{1.25} T \le 2\log_2 T$. Furthermore, between two times $\widehat{\Delta}_j$ increases, since Eq. (11) and Eq. (12) are not triggered, the epoch length is at least two times the previous one (by Eq. (13)). Therefore, between two times $\widehat{\Delta}_j$ increases, the number of epochs is upper bounded by $\log_2 T$. Overall, the total number of epochs is upper bounded by $2\log_2 T \times \log_2 T = 2\log^2 T$. Since we allow the maximum number of epochs to be $3\log^2 T$ in Algorithm 4, it will not end before the number of rounds reaches T. ∎

Proof of Theorem 4. Let $L^\star$ be the smallest L such that

$$
32\left(\frac{\beta_4}{\Delta} + \beta_2 C\right) \le \sqrt{\beta_4 L}.
$$

In the for-loops where the learner uses $L \le L^\star$, by Lemma 5 and Lemma 7, the sum of regret in Phase 1 and Phase 2 is upper bounded by

$$
\widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L} + \beta_2 C + \beta_4\right) = \widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L^\star} + \beta_2 C + \beta_4\right) = \widetilde{\mathcal{O}}\left(\frac{\beta_4}{\Delta} + \beta_2 C\right).
$$

In the for-loop where the learner first time uses $L > L^\star$, we have

$$
\beta_2 2^k \ge \sqrt{\beta_4(L-1)} \ge \sqrt{\beta_4 L^\star} \ge 32\left(\frac{\beta_4}{\Delta} + \beta_2 C\right)
$$

where the first inequality is by the choice of L in COBE. By Lemma 6, with probability at least $1 - \mathcal{O}(\delta)$, $\widehat{\pi} = \pi^\star$. Further by Lemma 8, with high probability, Phase 2 will continue until the total number of rounds reaches T. In this case, using Lemma 5 and Lemma 7, we can still bound the regret in the remaining steps by

$$
\widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L} + \beta_2 C + \beta_4\right) = \widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L^\star} + \beta_2 C + \beta_4\right) = \widetilde{\mathcal{O}}\left(\frac{\beta_4}{\Delta} + \beta_2 C\right).
$$

By the discussions above, we also see that with high probability, in all for-loops, the learner uses $L < 2L^\star$ (because the algorithm will be locked in Phase 2 when the first time $L > L^\star$ happens). Therefore, by the condition of starting Phase 3, Phase 3 can only be reached when $L^\star = \Omega(T)$. In this case, the regret incurred in Phase 3, by Theorem 3, is upper bounded by

$$\widetilde{\mathcal{O}}\left(\sqrt{\beta_1 T} + \beta_2 C + \beta_3\right) = \widetilde{\mathcal{O}}\left(\sqrt{\beta_1 L^\star} + \beta_2 C + \beta_3\right) = \widetilde{\mathcal{O}}\left(\frac{\beta_4}{\Delta} + \beta_2 C + \beta_4\right) = \widetilde{\mathcal{O}}\left(\frac{\beta_4}{\Delta} + \beta_2 C\right).$$

Overall, after summing the regret in all phases and using the fact that the for-loop only repeat $\widetilde{\mathcal{O}}(1)$ times, we see that the total regret can be upper bounded by $\widetilde{\mathcal{O}}\left(\frac{\beta_4}{\Delta} + \beta_2 C\right)$. To show that the algorithm also simultaneously guarantees a bound of $\widetilde{\mathcal{O}}\left(\sqrt{\beta_4 T} + \beta_2 C + \beta_4\right)$, simply bound the regret in all phases by $\widetilde{\mathcal{O}}\left(\sqrt{\beta_4 L} + \beta_2 C + \beta_4\right) = \widetilde{\mathcal{O}}\left(\sqrt{\beta_4 T} + \beta_2 C + \beta_4\right)$.

■

Appendix F. The Implementation of the Leave-one-policy-out MDP

We consider a tabular MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, p, H)$ with a fixed initial state $s_1 \in \mathcal{S}$. Let $\Pi_{\mathcal{M}}$ denote the set of all deterministic policies in $\mathcal{M}$. Now, given a deterministic policy $\widehat{\pi} \in \Pi_{\mathcal{M}}$, our goal is to construct another MDP $\mathcal{M}'$, such that the policy set of $\mathcal{M}'$ includes all policies in $\mathcal{M}$ except for $\widehat{\pi}$, and that for any $\pi \in \Pi_{\mathcal{M}} \backslash \{\widehat{\pi}\}$, the expected reward in $\mathcal{M}$ and $\mathcal{M}'$ is the same.

MDP $\mathcal{M}'$ has state space $\{s_0\} \cup \mathcal{S} \times \mathcal{S}$ and horizon $H+1$. In $\mathcal{M}'$, the agent starts in the initial state s_0 and takes one of S actions which makes it transition to one of S copies of the original MDP $\mathcal{M}$. The s-th copy of $\mathcal{M}$ is denoted by $\mathcal{M}_s$ and is identical to $\mathcal{M}$ except that the agent is not allowed to take the actions prescribed by $\widehat{\pi}$ in state s.

Note that we can obtain samples for $\mathcal{M}'$ by playing in $\mathcal{M}$, and that

$$\max_{\pi \in \Pi_{\mathcal{M}'}} \mu^\pi_{\mathcal{M}'} = \max_{\pi \in \Pi_{\mathcal{M}} \backslash \{\widehat{\pi}\}} \mu^\pi_{\mathcal{M}} \tag{29}$$

where $\mu^\pi_{\mathcal{M}}$ denotes the expected reward of policy π under MDP $\mathcal{M}$. To see this, simply notice that for any $\pi \in \Pi_{\mathcal{M}} \setminus \{\widehat{\pi}\}$ which differs from $\pi^\star$ on state s, one can find a policy $\pi' \in \Pi_{\mathcal{M}'}$ that first goes to $\mathcal{M}_s$ in $\mathcal{M}'$ in the first step, and then follow π in the rest of the steps. This policy π' gives the same expected reward as π. Conversely, for any $\pi' \in \Pi_{\mathcal{M}'}$, there is a policy $\pi \in \Pi_{\mathcal{M}} \setminus \{\widehat{\pi}\}$ which simply equals to π' on its 2 to $H+1$ steps. This π gives the same expected reward as π'.

Although $\mathcal{M}'$ has S^2+1 states, and the total number of actions is $(SA-1) \times S + S$ (where $SA-1$ is the total number of actions in each copy of $\mathcal{M}$, and the additional S is the number of actions on s_0), running UCBVI on $\mathcal{M}'$ can in fact yield the same gap-independent bound as running it in $\mathcal{M}$ if we share the samples among different copies of $\mathcal{M}$. To see this in the uncorrupted case, notice that in the analysis of the UCBVI algorithm (see, e.g., (Azar et al., 2017), or Chapter 7 of (Agarwal et al., 2020b)), the regret bound is a sum of terms of the form $\sum_t \sum_h \frac{\mathrm{poly}(S,A,H)}{\sqrt{n_t(s_h^t, a_h^t)}}$ or $\sum_t \sum_h \frac{\mathrm{poly}(S,A,H)}{n_t(s_h^t, a_h^t)}$. When the samples of the S copies of $\mathcal{M}$ are shared, these sum will only scale with the original number of states and actions. This can also be proved formally through the use of feedback graphs (Dann et al., 2020). In the corrupted case, the amount of corruption (i.e., $c_t = H \cdot \sup_{s,a,V} |(\mathcal{T}V - \mathcal{T}_t V)(s,a)|$) remains the same in $\mathcal{M}$ and in $\mathcal{M}'$. Therefore, the overall regret bound in $\mathcal{M}'$ under corruption remains the same order as that in $\mathcal{M}$.

Table 2: A refined version of Table 1 with all dependencies explicitly written out. See Table 1 for the restrictions of every algorithm. S and A are the number of states and actions respectively; d is the feature dimension in linear settings; $\dim_{\mathrm{DE}}$ is the Bellman-eluder dimension defined in (Jin et al., 2021a); G is the GapComplexity defined in (Simchowitz and Jamieson, 2019); Δ is the gap between the expected reward of the best and second-best policy. It holds that $\mathrm{G} \le \frac{SA}{\Delta}$. $C^{\mathsf{a}} = \sum_t c_t$ and $C^{\mathsf{r}} = \sqrt{T \sum_t c_t^2}$, where c_t is the amount of corruption in round t.

Setting	Algorithm	$\mathrm{Reg}(T)$ in $\widetilde{\mathcal{O}}(\cdot)$
Tabular MDP	(Lykouris et al., 2021)	$\mathrm{poly}(H)\big((1 + C^{\mathsf{a}}) \min\{\mathrm{G}, \sqrt{SAT}\} + S^2AC^{\mathsf{a}} + SA(C^{\mathsf{a}})^2\big)$
	(Chen et al., 2021b)	$\mathrm{poly}(H)\big(\min\{\frac{S^2A^{3/2}}{\Delta}, \sqrt{S^4A^3T}\} + S^2A^2C^{\mathsf{a}} + (C^{\mathsf{a}})^2\big)$
	(Jin et al., 2021b)	$\mathrm{poly}(H)\big(\min\{\mathrm{G}, \sqrt{S^2A^2T}\} + C^{\mathsf{a}}\big)$
	COBE + UCBVI	$\mathrm{poly}(H)\big(\sqrt{SAT} + S^2A + SAC^{\mathsf{a}}\big)$
	G-COBE + UCBVI	$\mathrm{poly}(H)\big(\min\{\frac{S^2A}{\Delta}, \sqrt{S^2AT}\} + S^2A + SAC^{\mathsf{a}}\big)$
Linear bandit	(Li et al., 2019)	$\frac{d^6}{\Delta^2} + \frac{d^{5/2}C^{\mathsf{a}}}{\Delta}$
	(Bogunovic et al., 2020)	$d\sqrt{T} + C^{\mathsf{a}}\sqrt{dT}$
	(Bogunovic et al., 2021)	$d\sqrt{T} + d^2(C^{\mathsf{a}})^2$
	(Lee et al., 2021)	$\min\{\frac{d^2}{\Delta}, d\sqrt{T}\} + C^{\mathsf{a}}$
	COBE + PE	$d\sqrt{T} + d^{3/2}C^{\mathsf{a}} + d^{3/2}$
	G-COBE + PE	$\min\{\frac{d^2}{\Delta}, d\sqrt{T}\} + d^{3/2}C^{\mathsf{a}} + d^2$
Linear contextual bandit	(Foster et al., 2020)	$d\sqrt{T} + \sqrt{d}C^{\mathsf{r}}$
	COBE + OFUL	$d\sqrt{T} + dC^{\mathsf{r}}$
	COBE + VOFUL	$d^{4.5}\sqrt{T} + d^4C^{\mathsf{a}}$
Linear MDP	(Lykouris et al., 2021)	$\mathrm{poly}(H)\big(C^{\mathsf{a}}\sqrt{(d^3 + dA)T} + (C^{\mathsf{a}})^2\sqrt{dT}\big)$
	(Wei and Luo, 2021)	$\mathrm{poly}(H)\big(\sqrt{d^3T} + (C^{\mathsf{a}})^{1/3}d^{4/3}T^{2/3}\big)$
	COBE + LSVI-UCB	$\mathrm{poly}(H)\big(\sqrt{d^3T} + dC^{\mathsf{r}}\big)$
	COBE + VARLin	$\mathrm{poly}(H)\big(d^{4.5}\sqrt{T} + d^4C^{\mathsf{a}}\big)$
Low BE-dimension	COBE + GOLF	$\mathrm{poly}(H)\big(\sqrt{\dim_{\mathrm{DE}} \cdot T} + \sqrt{\dim_{\mathrm{DE}}}C^{\mathsf{r}}\big)$

Appendix G. Base Algorithms

In this section, we describe and analyze the base algorithms for all settings considered in Table 1. The precise regret bounds achieved by our approaches and comparison with previous works are summarized in Table 2, which complements Table 1. The proofs are sometimes brief since they mostly follow standard analysis appeared in previous works. More details can be found in the references.

G.1. Robust UCBVI for tabular MDPs

A Robust UCBVI algorithm is presented by Lykouris et al. (2021) in their Appendix B. We translate it to our setting (i.e., our trajectory reward is bounded in $[0, 1]$, and our definition of C^{a} already includes an H factor). The resulting algorithm essentially runs the standard UCBVI algorithm

(Azar et al., 2017) with enlarged bonuses

$$b_t(s,a) = \min\left\{2\sqrt{\frac{2\ln(64SAHT^2/\delta)}{n_t(s,a)}} + \frac{C^{\mathsf{a}}}{n_t(s,a)},\ 1\right\}$$

where $n_t(s,a)$ is the number of visits to (s,a) before episode t, and C^{a} is a given upper bound of the total corruption. This algorithm achieves the following bound (c.f. Eq. (B.1) in Lykouris et al. (2021)):

$$\sum_{\tau=1}^{t}\left(V_1^\star(s_1^t) - V_1^{\pi_t}(s_1^t)\right) \le \mathrm{poly}(H) \times \widetilde{\mathcal{O}}\left(\min\left\{\mathsf{GapComplexity}, \sqrt{SAt}\right\} + S^2A + C^{\mathsf{a}}SA\right). \tag{30}$$

Furthermore, by their definition of $\mathsf{GapComplexity}$ (with proper scaling for our setting), it holds that $\mathsf{GapComplexity} \le \frac{SA}{\mathsf{gap}_{\min}} \le \frac{SA}{\Delta}$, where $\mathsf{gap}_{\min} \triangleq \min_{s,a\neq\pi_h^\star(s),h}\left(V_h^\star(s) - Q_h^\star(s,a)\right)$ and the second inequality is by the performance difference lemma,

$$\begin{aligned}
\Delta = \min_{\pi\neq\pi^\star}\Delta_\pi &= \min_{\pi\neq\pi^\star}\mathbb{E}\left[\sum_{h,s,a}\Pr[s_h = s](V_h^\star(s) - Q_h^\star(s,a)) \,\middle|\, \pi\right] \\
&\le \min_{h,s,a\neq\pi_h^\star(s)}\mathbb{E}\left[\Pr[s_h = s](V_h^\star(s) - Q_h^\star(s,a)) \,\middle|\, \pi : \pi_h(s') = \pi_h^\star(s')\forall s' \neq s, \pi_h(s) = a\right] \\
&\qquad\qquad\text{(Let } \pi \text{ be the policy that only differs from } \pi^\star \text{ on state } s \text{ at level } h\text{)} \\
&\le \mathsf{gap}_{\min}.
\end{aligned}$$

Below we use these facts to derive our bound.

Theorem 26 *For finite-horizon tabular MDPs,* COBE *with Robust UCBVI as the base algorithm guarantees* $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\mathrm{poly}(H) \times \left(\sqrt{SAT} + S^2A + SAC^{\mathsf{a}}\right)\right)$*;* G-COBE *with Robust UCBVI guarantees* $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\mathrm{poly}(H) \times \left(\min\left\{\frac{S^2A}{\Delta}, \sqrt{S^2AT}\right\} + S^2A + SAC^{\mathsf{a}}\right)\right)$.

Proof By Eq. (30) and Azuma's inequality, we have

$$\sum_{\tau=1}^{t}\left(r_\tau^{\pi^\star} - r_\tau\right) \le \mathrm{poly}(H) \times \widetilde{\mathcal{O}}\left(\sqrt{SAt} + S^2A + SAC^{\mathsf{a}}\right),$$

which satisfies Eq. (5) with $\beta_1 = \widetilde{\Theta}(\mathrm{poly}(H)SA)$, $\beta_2 = \widetilde{\Theta}(\mathrm{poly}(H)SA)$, $\beta_3 = \widetilde{\Theta}(\mathrm{poly}(H)S^2A)$. Applying Theorem 3 with these parameters we get the bound for COBE.

Let $N_t^{\neq\pi^\star} = \sum_{\tau=1}^{t}\mathbf{1}[\pi_\tau \neq \pi^\star]$ be the number of times the learner chooses sub-optimal policies. Using Eq. (30) and noticing that the left-hand side of it is lower bounded by $N_t^{\neq\pi^\star}\Delta$, we get

$$N_t^{\neq\pi^\star}\Delta \le \mathrm{poly}(H) \times \widetilde{\mathcal{O}}\left(\frac{SA}{\Delta} + S^2A + SAC^{\mathsf{a}}\right).$$

Algorithm 5 Robust Phased Elimination

input: C^{a}
define: $m_0 = 4d([\log\log d]_+ + 18)$
initialize: $\mathcal{A}_0 \leftarrow \mathcal{A}$
for $k = 0, 1, 2, \dots$ **do**

Let $m_k = 2^{k-1} m_0$.
Compute $\zeta_k : \mathcal{A}_k \to [0, 1]$ such that

$$\max_{a\in\mathcal{A}_k} \|a\|_{\Gamma(\zeta_k)^{-1}} \le 2d \quad \text{and} \quad |\mathrm{supp}(\zeta_k)| \le m_0$$

where $\Gamma(\zeta_k) = \sum_{a\in\mathcal{A}_k} \zeta_k(a) a a^\top$.
Set

$$u_k(a) = \begin{cases} 0 & \text{if } \zeta_k(a) = 0 \\ \left\lceil m_k \max\{\zeta_k(a), \frac{1}{m_0}\}\right\rceil & \text{otherwise} \end{cases}$$

Draw each action $a \in \mathcal{A}_k$ exactly $u_k(a)$ times, and get action-reward pairs $(a_\tau, r_\tau)_{\tau=1}^{u_k}$, where $u_k = \sum_{a\in\mathcal{A}_k} u_k(a)$.
Estimate parameter:

$$w_k = \Gamma_k^{-1} \sum_{\tau=1}^{u_k} a_\tau r_\tau, \qquad \text{where } \Gamma_k = \sum_{a\in\mathcal{A}_k} u_k(a) a a^\top.$$

Update the active set

$$\mathcal{A}_{k+1} \leftarrow \left\{ a \in \mathcal{A}_k \;:\; \max_{a'\in\mathcal{A}_k} w_k^\top (a' - a) \le 4d\sqrt{\frac{1}{m_k}\log(T/\delta)} + \frac{4\sqrt{2d}m_0}{m_k} C^{\mathsf{a}} \right\}.$$

end

Therefore, by Azuma's inequality, we have with probability $1 - \mathcal{O}(\delta)$,

$$\begin{aligned}\sum_{\tau=1}^{t} (r_\tau^{\pi^\star} - r_\tau) &\le \mathrm{poly}(H) \times \widetilde{\mathcal{O}}\left(\min\left\{ \frac{SA}{\Delta}, \sqrt{SAt} \right\} + S^2 A + C^{\mathsf{a}} SA + \sqrt{N_t^{\neq\pi^\star}} \right) \\ &\le \mathrm{poly}(H) \times \widetilde{\mathcal{O}}\left(\min\left\{ \frac{SA}{\Delta}, \sqrt{SAt} \right\} + S^2 A + C^{\mathsf{a}} SA \right).\end{aligned}$$

This satisfies Eq. (7) with $\beta_1 = \widetilde{\Theta}(\mathrm{poly}(H)SA)$, $\beta_2 = \widetilde{\Theta}(\mathrm{poly}(H)SA)$, $\beta_3 = \widetilde{\Theta}(\mathrm{poly}(H)S^2 A)$ (therefore, $\beta_4 = \widetilde{\Theta}(\mathrm{poly}(H)S^2 A)$). Applying Theorem 4 with these parameters we get the bound for G-COBE. ∎

G.2. Robust Phased Elimination for linear bandits

The Robust Phased Elimination (Algorithm 5) is exactly the Algorithm 1 of Bogunovic et al. (2021) with the choice of parameters specified in their Theorem 1 (i.e., $\nu = \frac{1}{m_0}$ in their notations). Its gap-independent bound is shown below:

Lemma 27 *With probability at least $1 - \mathcal{O}(\delta)$, Robust Phased Elimination ensures*

$$\sum_{\tau=1}^{t}(r_\tau^{a^\star} - r_\tau) = \widetilde{\mathcal{O}}\left(d\sqrt{t\log(T/\delta)} + C^{\mathsf{a}} d^{3/2}\log T\right).$$

Proof By Theorem 1 of Bogunovic et al. (2021), we have

$$\sum_{\tau=1}^{t}(\mu^{a^\star} - \mu^{a_t}) = \widetilde{\mathcal{O}}\left(\sqrt{dt\log(T|\mathcal{A}|/\delta)} + C^{\mathsf{a}} d^{3/2}\log T\right) = \widetilde{\mathcal{O}}\left(d\sqrt{t\log(T/\delta)} + C^{\mathsf{a}} d^{3/2}\log T\right)$$

where for simplicity we assume $|\mathcal{A}| = \mathcal{O}(T^d)$ without loss of generality. The conclusion follows by noticing that $\sum_{\tau=1}^{t}|\mu^{a^\star} - \mu_\tau^{a^\star}| \le C^{\mathsf{a}}$, $\sum_{\tau=1}^{t}|\mu^{a_\tau} - \mu_\tau^{a_\tau}| \le C^{\mathsf{a}}$ by the definition of corruption, and that with probability $1-\mathcal{O}(\delta)$, $\left|\sum_{\tau=1}^{t}(r_\tau^{a^\star} - \mu_\tau^{a^\star})\right| = \mathcal{O}(\sqrt{t\log(T/\delta)})$, and $\left|\sum_{\tau=1}^{t}(r_\tau - \mu_\tau^{a_\tau})\right| = \mathcal{O}(\sqrt{t\log(T/\delta)})$ by Azuma's inequality. ■

Next, we further show that the same algorithm achieves a gap-dependent bound. We first restate an intermediate result of Bogunovic et al. (2021).

Lemma 28 (Appendix A.2 of Bogunovic et al. (2021)) *Robust Phased Elimination ensures that with probability at least $1-\mathcal{O}(\delta)$, $a^\star \in \mathcal{A}_k$ for all k.*

The gap-dependent bound of Robust Phased Elimination is then given by the following lemma.

Lemma 29 *With probability $1-\mathcal{O}(\delta)$, Robust Phased Elimination ensures*

$$\sum_{\tau=1}^{t}\left(r_\tau^{a^\star} - r_\tau\right) = \widetilde{\mathcal{O}}\left(\frac{d^2\log(T/\delta)}{\Delta} + C^{\mathsf{a}} d^{3/2}\log T\right).$$

Proof By Eq. (58) of (Bogunovic et al., 2021), for all $a \in \mathcal{A}_k$, $a \neq a^\star$, we have

$$\Delta \le w^{\star\top}(a^\star - a) \le 8d\sqrt{\frac{\log(T/\delta)}{m_k}} + \frac{8\sqrt{2}dm_0}{m_k}C^{\mathsf{a}}$$

where the first inequality is by our assumption. Solving the inequality we get

$$m_k \le \frac{256d^2\log(T/\delta)}{\Delta^2} + \frac{16\sqrt{2}dm_0C^{\mathsf{a}}}{\Delta}. \tag{31}$$

This means that as long as m_k grows larger than the right-hand side of Eq. (31), no sub-optimal arm can remain in $\mathcal{A}_k$. Let $k^\star$ be the smallest k such that m_k is larger than the right-hand side of Eq. (31). Then we only need to calculate the regret incurred in epochs $1, \dots, k^\star - 1$. By the same calculation as Eq. (48)-(55) in (Bogunovic et al., 2021), we get that with probability at least $1-\mathcal{O}(\delta)$,

$$
\begin{aligned}
\sum_{\tau=1}^{t}(r_\tau^{a^\star}-r_\tau) &\le \sum_{k=1}^{k^\star-1}\sum_{\tau\in\text{epoch}(k)}(r_\tau^{a^\star}-r_\tau) \\
&\le \sum_{k=1}^{k^\star-1}\sum_{\tau\in\text{epoch}(k)}(\mu^{a^\star}-\mu^{a_\tau})+\mathcal{O}\left(\sum_{k=1}^{k^\star-1}\sqrt{m_k\log(T/\delta)}+C^{\mathsf{a}}\right) \\
&\hspace{20em}\text{(Azuma's inequality)} \\
&=\mathcal{O}\left(u_0+\sum_{k=1}^{k^\star-1}\left(d\sqrt{m_k\log(T/\delta)}+C^{\mathsf{a}}m_0\sqrt{d}\right)\right) \\
&\hspace{12em}\text{(By Eq. (48)-(55) in (Bogunovic et al., 2021))} \\
&=\mathcal{O}\left(m_0+d\sqrt{m_{k^\star}\log(T/\delta)}+C^{\mathsf{a}}m_0\sqrt{d}\log T\right) \\
&=\mathcal{O}\left(\frac{d^2\log(T/\delta)}{\Delta}+d^{3/2}C^{\mathsf{a}}\log T\right).
\end{aligned}
$$

■

Theorem 30 *For linear bandits,* COBE *with Robust Phased Elimination as the base algorithm guarantees* $\text{Reg}(T)=\widetilde{\mathcal{O}}\left(d\sqrt{T}+d^{3/2}C^{\mathsf{a}}+d^{3/2}\right)$*;* G-COBE *with Robust Phased Elimination guarantees* $\text{Reg}(T)=\widetilde{\mathcal{O}}\left(\min\left\{\frac{d^2}{\Delta},d\sqrt{T}\right\}+d^{3/2}C^{\mathsf{a}}+d^2\right)$.

Proof By Lemma 27 and Lemma 29, we see that Robust Phased Elimination satisfies Eq. (5) and Eq. (7) with $\beta_1=\widetilde{\Theta}(d^2),\beta_2=\widetilde{\Theta}(d^{3/2}),\beta_3=\Theta(d)$ (thus, $\beta_4=\widetilde{\Theta}(d^2)$). Using them in Theorem 3 and Theorem 4 gives the desired bounds. ■

G.3. Robust OFUL for linear contextual bandits / Robust LSVI-UCB for linear MDPs

From this section, we denote the state, the action, and the reward at the h-th step of the t-th episode as s_h^t, a_h^t, and σ_h^t respectively (same as the $s_{t,h}, a_{t,h}, \sigma_{t,h}$ defined in Section 3).

Below we restate the linear MDP assumption in (Jin et al., 2020b) (adapted to our case where the per-step reward lies in $[0,\frac{1}{H}]$):

Assumption 6 (Finite-horizon Linear MDP) *Let $\phi(s,a)\in\mathbb{R}^d$ be known feature vector for the state-action pair (s,a). Assume that for all (s,a), the reward function can be represented as $\sigma(s,a)=\phi(s,a)^\top\rho$, and the transition kernel can be represented as $p(s'|s,a)=\phi(s,a)^\top\nu(s')$ for some $\nu(s')\in\mathbb{R}^d$. Without loss of generality, we assume that $\|\phi(s,a)\|\le 1$, $\|\rho\|\le\frac{1}{H}\sqrt{d}$, and $\left\|\int\nu(s')\mathrm{d}s'\right\|\le\sqrt{d}$.*

In Algorithm 6, we present a corruption robust version of LSVI-UCB (Jin et al., 2020b) that takes C^{r} as input. Since linear contextual bandit is a special case of linear MDP with $H=1$, we can use the same algorithm to deal with it.

The following lemma is adapted from (Jin et al., 2020b, Lemma B.4).

Algorithm 6 Robust OFUL / Robust LSVI-UCB

input: C^{r}

define: $\zeta = \zeta_0 \cdot \sqrt{d \log(dT/\delta)}$ for $H = 1$ (linear contextual bandit case), or $\zeta = \zeta_0 \cdot d \sqrt{\log(dHT/\delta)}$ for $H > 1$ (linear MDP case), where ζ_0 is a universal constant.

for $t = 1, \dots, T$ **do**

$$\Lambda^t = \sum_{\tau=1}^{t-1} \sum_{h=1}^{H} \phi_h^\tau \phi_h^{\tau\top} + I, \qquad \text{where } \phi_h^\tau \triangleq \phi(s_h^\tau, a_h^\tau).$$

  for $h = H, \dots, 1$ **do**

$$w_h^t = (\Lambda^t)^{-1} \sum_{\tau=1}^{t-1} \sum_{k=1}^{H} \phi_k^{\tau\top} \left(\sigma_k^\tau + V_{h+1}^t(s_{k+1}^\tau)\right)$$

$$Q_h^t(s,a) \leftarrow \min\left\{ w_h^{t\top} \phi(s,a) + \left(4\zeta + C^{\mathsf{r}} \sqrt{\frac{d}{Ht}}\right) \|\phi(s,a)\|_{(\Lambda^t)^{-1}}, \ 1 \right\}$$

$$V_h^t(s) \leftarrow \max_a Q_h^t(s,a)$$

  end

  for $h = 1, \dots, H$ **do**

    Observe s_h^t, choose $a_h^t = \mathrm{argmax}_a\, Q_h^t(s_h^t, a)$, and receive σ_h^t.

  end

end

Lemma 31 *For any π, with probability at least $1 - \mathcal{O}(\delta)$, the following holds for all s, a:*

$$\phi(s,a)^\top (w_h^t - w_h^\pi) = \mathbb{E}_{s' \sim p_t(\cdot|s,a)} \left[V_{h+1}^t(s') - V_{h+1}^\pi(s')\right] + \varepsilon_h^t(s,a)$$

for some $\varepsilon_h^t(s,a)$ that satisfies

$$|\varepsilon_h^t(s,a)| \le \left(4\zeta + C^{\mathsf{r}} \sqrt{\frac{d}{Ht}}\right) \|\phi(s,a)\|_{(\Lambda^t)^{-1}} + \frac{2}{H} c_t \tag{32}$$

Proof For any (s,a) and any π,

$$\begin{aligned}
&\phi(s,a)^\top (w_h^t - w_h^\pi) \\
&= \phi(s,a)^\top \left((\Lambda^t)^{-1} \sum_{\tau=1}^{t-1} \sum_{k=1}^{H} \phi_k^\tau \Big(\sigma_k^\tau + V_{h+1}^t(s_{k+1}^\tau)\Big) - w_h^\pi \right) \\
&= \phi(s,a)^\top (\Lambda^t)^{-1} \left(\sum_{\tau=1}^{t-1} \sum_{k=1}^{H} \phi_k^\tau \left(\sigma_k^\tau + V_{h+1}^t(s_{k+1}^\tau) - \phi_k^{\tau\top} \rho - \phi_k^{\tau\top} \int \nu(s') V_{h+1}^\pi(s') \mathrm{d}s' \right) - w_h^\pi \right) \\
&= \phi(s,a)^\top (\Lambda^t)^{-1} \left(\sum_{\tau=1}^{t-1} \sum_{k=1}^{H} \phi_k^\tau \left(\sigma_k^\tau + V_{h+1}^t(s_{k+1}^\tau) - \sigma(s_k^\tau, a_k^\tau) - \mathbb{E}_{s' \sim p(\cdot|s_k^\tau, a_k^\tau)} V_{h+1}^\pi(s') \right) - w_h^\pi \right)
\end{aligned}$$

$$
\begin{aligned}
&= \underbrace{\phi(s,a)^\top(\Lambda^t)^{-1}\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}\phi_k^\tau\left(\sigma_k^\tau - \sigma_\tau(s_k^\tau, a_k^\tau)\right)}_{\textbf{term}_1} \\
&\quad + \underbrace{\phi(s,a)^\top(\Lambda^t)^{-1}\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}\phi_k^\tau\left(V_{h+1}^t(s_{k+1}^\tau) - \mathbb{E}_{s'\sim p_\tau(\cdot|s_k^\tau,a_k^\tau)}V_{h+1}^t(s')\right)}_{\textbf{term}_3} \\
&\quad + \underbrace{\phi(s,a)^\top(\Lambda^t)^{-1}\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}\phi_k^\tau\left(\sigma_\tau(s_k^\tau, a_k^\tau) + \mathbb{E}_{s'\sim p_\tau(\cdot|s_k^\tau,a_k^\tau)}V_{h+1}^t(s') - \sigma(s_k^\tau, a_k^\tau) - \mathbb{E}_{s'\sim p(\cdot|s_k^\tau,a_k^\tau)}V_{h+1}^t(s')\right)}_{\textbf{term}_3} \\
&\quad + \underbrace{\phi(s,a)^\top(\Lambda^t)^{-1}\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}\phi_k^\tau\left(\mathbb{E}_{s'\sim p(\cdot|s_k^\tau,a_k^\tau)}V_{h+1}^t(s') - \mathbb{E}_{s'\sim p(\cdot|s_k^\tau,a_k^\tau)}V_{h+1}^\pi(s')\right)}_{\textbf{term}_4} \quad \underbrace{-\phi(s,a)^\top(\Lambda^t)^{-1}w_h^\pi}_{\textbf{term}_5}
\end{aligned}
$$

$\textbf{term}_1$ and $\textbf{term}_2$ are of the same form, Their absolute values $|\textbf{term}_1|$ and $|\textbf{term}_2|$ can both be upper bounded by $\zeta\|\phi(s,a)\|_{(\Lambda^t)^{-1}}$ with a similar proof as Lemma B.3 and Lemma D.4 of (Jin et al., 2020b) (notice that our range of reward is smaller than theirs by a $\frac{1}{H}$ factor).

Then notice that

$$
\textbf{term}_3 = \phi(s,a)^\top(\Lambda^t)^{-1}\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}\phi_k^\tau\left(\mathcal{T}_\tau V_{h+1}^t(s_k^\tau, a_k^\tau) - \mathcal{T}V_{h+1}^t(s_k^\tau, a_k^\tau)\right)
$$

($\mathcal{T}_t$ and $\mathcal{T}$ are Bellman operators defined in Section 3)

and $|\textbf{term}_3|$ is upper bounded by

$$
\begin{aligned}
&\|\phi(s,a)\|_{(\Lambda^t)^{-1}}\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}\|\phi_k^\tau\|_{(\Lambda^t)^{-1}} \times \frac{1}{H}c_\tau \\
&\qquad\qquad \text{(recall that } c_t \triangleq H\cdot\sup_{s,a}\sup_{V\in[0,1]^{\mathcal{S}}}|(\mathcal{T}V - \mathcal{T}_tV)(s,a)|) \\
&\le \frac{1}{H}\|\phi(s,a)\|_{(\Lambda^t)^{-1}}\sqrt{\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}c_\tau^2}\sqrt{\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}\|\phi_k^\tau\|^2_{(\Lambda^t)^{-1}}} \\
&\le \frac{C^{\mathrm{r}}}{\sqrt{Ht}} \times \sqrt{d}\|\phi(s,a)\|_{(\Lambda^t)^{-1}}. \qquad \text{(by Lemma D.1 of (Jin et al., 2020b))}
\end{aligned}
$$

$$
\begin{aligned}
\textbf{term}_4 &= \phi(s,a)^\top(\Lambda^t)^{-1}\sum_{\tau=1}^{t-1}\sum_{k=1}^{H}\phi_k^\tau\phi_k^{\tau\top}\left(\rho + \int\nu(s')\left(V_{h+1}^t(x') - V_{h+1}^\pi(s')\right)\mathrm{d}s'\right) \\
&= \phi(s,a)^\top\left(\rho + \int\nu(s')\left(V_{h+1}^t(s') - V_{h+1}^\pi(s')\right)\mathrm{d}s'\right)
\end{aligned}
$$

$$\underbrace{-\phi(s,a)^\top(\Lambda^t)^{-1}\left(\rho+\int\nu(s')\left(V^t_{h+1}(s')-V^\pi_{h+1}(s')\right)\mathrm{d}s'\right)}_{\textbf{term}_5}$$

$$\begin{aligned}
&=\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[V^t_{h+1}(s')-V^\pi_{h+1}(s')\right]+\textbf{term}_5\\
&=\mathbb{E}_{s'\sim p_t(\cdot|s,a)}\left[V^t_{h+1}(s')-V^\pi_{h+1}(s')\right]+\textbf{term}_5\\
&\quad+\underbrace{\left(\sigma(s,a)+\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[V^t_{h+1}(s')\right]-\sigma_t(s,a)-\mathbb{E}_{s'\sim p_t(\cdot|s,a)}\left[V^t_{h+1}(s')\right]\right)}_{\textbf{term}_6}\\
&\quad-\underbrace{\left(\sigma(s,a)+\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[V^\pi_{h+1}(s')\right]-\sigma_t(s,a)-\mathbb{E}_{s'\sim p_t(\cdot|s,a)}\left[V^\pi_{h+1}(s')\right]\right)}_{\textbf{term}_7}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
|\textbf{term}_5|&\le\|\phi(s,a)\|_{(\Lambda^t)^{-1}}\left\|\rho+\int\nu(s')\left(V^t_{h+1}(s')-V^\pi_{h+1}(s')\right)\mathrm{d}s'\right\|_{(\Lambda^t)^{-1}}\\
&\le\|\phi(s,a)\|_{(\Lambda^t)^{-1}}\left\|\rho+\int\nu(s')\left(V^t_{h+1}(s')-V^\pi_{h+1}(s')\right)\mathrm{d}s'\right\|\\
&\le 2\sqrt{d}\|\phi(s,a)\|_{(\Lambda^t)^{-1}} \qquad\text{(by Assumption 6)}\\
&\le\zeta\|\phi(s,a)\|_{(\Lambda^t)^{-1}},\\
|\textbf{term}_6|&\le\sup_{s',a',V\in[0,1]^{\mathcal{S}}}\left|\left(\mathcal{T}_tV-\mathcal{T}V\right)(s',a')\right|=\frac{1}{H}c_t\\
|\textbf{term}_7|&\le\sup_{s',a',V\in[0,1]^{\mathcal{S}}}\left|\left(\mathcal{T}_tV-\mathcal{T}V\right)(s',a')\right|=\frac{1}{H}c_t
\end{aligned}$$

Finally,

$$\textbf{term}_5\le\|\phi(s,a)\|_{(\Lambda^t)^{-1}}\|w^\pi_h\|_{(\Lambda^t)^{-1}}\le\sqrt{d}\|\phi(s,a)\|_{(\Lambda^t)^{-1}}\le\zeta\|\phi(s,a)\|_{(\Lambda^t)^{-1}}.$$

Combining all terms above finishes the proof. ∎

Lemma 32 *With probability at least $1-\mathcal{O}(\delta)$, $V^t_h(s)\ge V^\star_h(s)-2c_t$ for all t,h,s.*

Proof We use induction to show that with probability at least $1-\mathcal{O}(\delta)$, $Q^t_h(s,a)\ge Q^\pi_h(s,a)-\frac{2(H+1-h)}{H}c_t$ for all h,s,a and any π. Consider the case $h=H$,

$$\begin{aligned}
&Q^t_H(s,a)\\
&=\min\left\{w^{t\top}_H\phi(s,a)+\left(4\zeta+C^{\mathsf{r}}\sqrt{\frac{d}{Ht}}\right)\|\phi(s,a)\|_{(\Lambda^t)^{-1}},\ 1\right\}\\
&=\min\left\{w^{\pi\top}_H\phi(s,a)+\left(4\zeta+C^{\mathsf{r}}\sqrt{\frac{d}{Ht}}\right)\|\phi(s,a)\|_{(\Lambda^t)^{-1}}+\varepsilon^t_H(s,a),\ Q^\pi_H(s,a)\right\}\\
&\qquad\text{(by Lemma 31)}
\end{aligned}$$

$$
\begin{aligned}
&\geq \min\left\{Q_H^\pi(s,a) - \frac{2}{H}c_t,\ \ Q_H^\pi(s,a)\right\} && \text{(by Lemma 31)}\\
&= Q_H^\pi(s,a) - \frac{2}{H}c_t.
\end{aligned}
$$

Suppose that the induction hypothesis holds for $h+1$, then

$$
\begin{aligned}
&Q_h^t(s,a)\\
&= \min\left\{w_h^{t\top}\phi(s,a) + \left(4\zeta + C^{\mathsf{r}}\sqrt{\frac{d}{Ht}}\right)\|\phi(s,a)\|_{(\Lambda^t)^{-1}},\ \ 1\right\}\\
&= \min\left\{w_h^{\pi\top}\phi(s,a) + \mathbb{E}_{s'\sim p_t(\cdot|s,a)}[V_{h+1}^t(s') - V_{h+1}^\pi(s')] + \left(4\zeta + C^{\mathsf{r}}\sqrt{\frac{d}{Ht}}\right)\|\phi(s,a)\|_{(\Lambda^t)^{-1}} + \varepsilon_h^t(s,a),\ \ 1\right\}\\
&\geq \min\left\{Q_h^\pi(s,a) + \mathbb{E}_{s'\sim p_t(\cdot|s,a)}[V_{h+1}^t(s') - V_{h+1}^\pi(s')] - \frac{2}{H}c_t,\ \ Q_h^\pi(s,a)\right\} && \text{(by Lemma 31)}
\end{aligned}
$$

Notice that by the induction hypothesis, we have for any s, $V_{h+1}^\pi(s) \leq \max_a Q_{h+1}^\pi(s,a) \leq \max_a Q_{h+1}^t(s,a) + \frac{2(H-h)}{H}c_t = V_{h+1}^t(s) + \frac{2(H-h)}{H}c_t$. Therefore, the last expression can further be lower bounded by

$$
\begin{aligned}
&\min\left\{Q_h^\pi(s,a) - \frac{2(H-h)}{H}c_t - \frac{2}{H}c_t,\ \ Q_h^\pi(s,a)\right\}\\
&= Q_h^\pi(s,a) - \frac{2(H-h+1)}{H}c_t,
\end{aligned}
$$

which finishes the induction. Note that $Q_h^t(s,a) \geq Q_h^\pi(s,a) - \frac{2(H-h+1)}{H}c_t$ implies the lemma since

$$
\begin{aligned}
V_h^t(s) &= \max_a Q_h^t(s,a)\\
&\geq \max_a Q_h^\star(s,a) - \frac{2(H-h+1)}{H}c_t && \text{(let } \pi = \pi^\star\text{)}\\
&\geq V_h^\star(s) - 2c_t.
\end{aligned}
$$

■

Lemma 33 *Robust OFUL / Robust LSVI-UCB ensures with probability at least* $1-\mathcal{O}(\delta)$

$$
\sum_{\tau=1}^t \left(r_\tau^{\pi^\star} - r_\tau\right) = \widetilde{\mathcal{O}}\left(\zeta\sqrt{dHt} + dC^{\mathsf{r}}\right).
$$

Proof Note that for all t, h,

$$
\begin{aligned}
&\sum_{\tau=1}^t \left(V_h^\tau(s_h^\tau) - V_h^{\pi_\tau}(s_h^\tau)\right) = \sum_{\tau=1}^t \left(Q_h^\tau(s_h^\tau, a_h^\tau) - Q_h^{\pi_\tau}(s_h^\tau, a_h^\tau)\right)\\
&\leq \sum_{\tau=1}^t \phi_h^\tau(w_h^\tau - w_h^{\pi_\tau}) + \widetilde{\mathcal{O}}\left(\sum_{\tau=1}^t \left(\zeta + C^{\mathsf{r}}\sqrt{\frac{d}{H\tau}}\right)\|\phi_h^\tau\|_{(\Lambda^\tau)^{-1}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{\tau=1}^{t} \mathbb{E}_{s'\sim p_\tau(\cdot|s_h^\tau,a_h^\tau)}\left[V_{h+1}^\tau(s') - V_{h+1}^{\pi_\tau}(s')\right] \\
&\qquad + \widetilde{\mathcal{O}}\left(\sum_{\tau=1}^{t}\left(\zeta + C^{\mathsf{r}}\sqrt{\frac{d}{H\tau}}\right)\|\phi_h^\tau\|_{(\Lambda^\tau)^{-1}} + \frac{1}{H}\sum_{\tau=1}^{t} c_\tau\right) && \text{(by Lemma 31)} \\
&= \sum_{\tau=1}^{t}\left(V_{h+1}^\tau(s_{h+1}^\tau) - V_{h+1}^{\pi_\tau}(s_{h+1}^\tau)\right) \\
&\qquad + \widetilde{\mathcal{O}}\left(\sum_{\tau=1}^{t}\left(\zeta + C^{\mathsf{r}}\sqrt{\frac{d}{H\tau}}\right)\|\phi_h^\tau\|_{(\Lambda^\tau)^{-1}} + \frac{1}{H}\sum_{\tau=1}^{t} c_\tau + \sum_{\tau=1}^{t}\epsilon_h^\tau\right), && (33)
\end{aligned}
$$

where we define $\epsilon_h^\tau \triangleq \mathbb{E}_{s'\sim p_\tau(\cdot|s_h^\tau,a_h^\tau)}\left[V_{h+1}^\tau(s') - V_{h+1}^{\pi_\tau}(s')\right] - \left(V_{h+1}^\tau(s_{h+1}^\tau) - V_{h+1}^{\pi_\tau}(s_{h+1}^\tau)\right)$. Thus,

$$
\begin{aligned}
\sum_{\tau=1}^{t}\left(V_1^\star(s_1^\tau) - V_1^{\pi_\tau}(s_1^\tau)\right) &\leq \sum_{\tau=1}^{t}\left(V_1^\tau(s_1^\tau) - V_1^{\pi_\tau}(s_1^\tau)\right) + \mathcal{O}\left(\sum_{\tau=1}^{t} c_\tau\right) && \text{(by Lemma 32)} \\
&\leq \widetilde{\mathcal{O}}\left(\sum_{\tau=1}^{t}\sum_{h=1}^{H}\left(\zeta + C^{\mathsf{r}}\sqrt{\frac{d}{Ht}}\right)\|\phi_h^\tau\|_{(\Lambda^\tau)^{-1}} + \sum_{\tau=1}^{t} c_\tau + \sum_{\tau=1}^{t}\sum_{h=1}^{H}\epsilon_h^\tau\right) \\
& && \text{(by Eq. (33))} \\
&\leq \widetilde{\mathcal{O}}\left(\sqrt{\sum_{\tau=1}^{t}\sum_{h=1}^{H}\left(\zeta^2 + \frac{(C^{\mathsf{r}})^2 d}{Ht}\right)}\sqrt{\underbrace{\sum_{\tau=1}^{t}\sum_{h=1}^{H}\|\phi_h^\tau\|^2_{(\Lambda^\tau)^{-1}}}_{=\widetilde{\mathcal{O}}(d)}} + C^{\mathsf{a}} + \sqrt{Ht}\right) \\
& && \text{(Cauchy-Schwarz and Azuma's inequality)} \\
&\leq \widetilde{\mathcal{O}}\left(\zeta\sqrt{dHt} + dC^{\mathsf{r}}\right).
\end{aligned}
$$

Finally, by Azuma's inequality, we get

$$
\sum_{\tau=1}^{t}\left(r_\tau^{\pi^\star} - r_\tau\right) = \widetilde{\mathcal{O}}\left(\zeta\sqrt{dHt} + dC^{\mathsf{r}}\right).
$$

■

Theorem 34 *For linear contextual bandits,* COBE *with Robust OFUL as the base algorithm guarantees* $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(d\sqrt{T} + dC^{\mathsf{r}}\right)$. *For linear MDPs,* COBE *with Robust LSVI-UCB as the base algorithm guarantees* $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\sqrt{d^3HT} + H\sqrt{T} + dC^{\mathsf{r}}\right)$.

Proof By Lemma 33, we see that Robust OFUL (with $\zeta = \widetilde{\Theta}(\sqrt{d})$) satisfies Eq. (5) with $\beta_1 = \widetilde{\Theta}(d^2), \beta_2 = \widetilde{\Theta}(d), \beta_3 = \Theta(1)$. Using them in Theorem 3 gives the desired bound for linear contextual bandits. For Robust LSVI-UCB, we pick $\zeta = \widetilde{\Theta}(d)$ and thus $\beta_1 = \widetilde{\Theta}(d^3H), \beta_2 = \widetilde{\Theta}(d), \beta_3 = \Theta(1)$. Using them in Theorem 3 together with the fact that $c_{\max} = \mathcal{O}(H)$, we get the desired bound for linear MDPs. ■

Algorithm 7 Robust VOFUL / Robust VARLin

input: C^{a}

define: $\ell_j \triangleq 2^{-j}$ and $\text{clip}_j(v) \triangleq \max(\min(v, \ell_j), -\ell_j)$. Let $\mathcal{B}(r)$ be Euclidean ball with radius r.

for $t = 1, 2, \ldots, T$ **do**

$$\mathcal{W}^t = \Bigg\{ w = (w_1, w_2, \ldots, w_H) \in \mathcal{B}(\sqrt{d})^H : \\ \left| \sum_{\tau=1}^{t-1} \text{clip}_j\left((\phi_h^\tau)^\top \xi\right) \left((\phi_h^\tau)^\top w_h - \sigma_h^\tau - V_{h+1}(w_{h+1})(s_{h+1}^\tau) \right) \right| \\ \le 200\ell_j \left(\sqrt{dHt \log(dTH/\delta)} + C^{\mathsf{a}} \right) \\ \forall \xi \in \mathcal{B}\left(2\sqrt{d}\right), \ \forall j \in [\lceil \log_2 T \rceil], \ \forall h \in [H] \Bigg\}$$

where $Q_h(w_h)(s,a) \triangleq w_h^\top \phi(s,a)$ and $V_h(w_h)(s) \triangleq \max_a Q_h(w_h)(s,a)$.
Let

$$w^t = \operatorname*{argmax}_{w \in \mathcal{W}^t} V_1(w_1)(s_1^t),$$

and define $Q_h^t(s,a) \triangleq Q_h(w_h^t)(s,a)$ and $V_h^t(s) \triangleq V_h(w_h^t)(s)$.

for $h = 1, \ldots, H$ **do**

Observe s_h^t, choose $a_h^t = \text{argmax}_a Q_h^t(s_h^t, a)$, and observe σ_h^t.

end

end

G.4. Robust-VOFUL for linear contextual bandits / Robust-VARLin for linear MDPs

In this section, we develop a variant of the algorithm of (Zhang et al., 2021c) that is robust to corruption (Algorithm 7). Notice that their original algorithm is for a different linear model called linear mixture MDP, but we carry the similar idea to the linear MDP setting. Again, the same algorithm works for linear contextual bandits.

Lemma 35 *With probability at least $1-\delta$, the following holds for all $t \in [T]$, $h \in [H]$, $j \in [\lceil \log_2 T \rceil]$, $\xi \in \mathcal{B}(2\sqrt{d})$, $w_{h+1} \in \mathcal{B}(\sqrt{d})$:*

$$\left| \sum_{\tau=1}^{t-1} \mathit{clip}_j\left(\phi_h^{\tau\top}\xi\right) \left(\phi_h^{\tau\top} \left(\rho + \int \nu(s')V(w_{h+1})(s')\mathrm{d}s' \right) - \left(\sigma_h^\tau + V(w_{h+1})(s_{h+1}^\tau)\right) \right) \right| \\ \le 200\ell_j \left(\sum_{\tau=1}^{t-1} c_\tau + \sqrt{dt \log(dHT/\delta)} \right) \tag{34}$$

Proof For a fixed tuple of t, h, j, ξ, w_{h+1}, recall that $\mathbb{E}[\sigma_h^\tau | s_h^\tau, a_h^\tau] = \sigma_\tau(s_h^\tau, a_h^\tau)$ and $s_{h+1}^\tau \sim p_\tau(\cdot | s_h^\tau, a_h^\tau)$. Therefore,

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathrm{clip}_j\left(\phi_h^{\tau\top}\xi\right)\left(\phi_h^{\tau\top}\left(\rho + \int \nu(s')V(w_{h+1})(s')\mathrm{d}s'\right) - \left(\sigma_h^\tau + V(w_{h+1})(s_{h+1}^\tau)\right)\right) \,\middle|\, s_h^\tau, a_h^\tau\right]\right| \\
&= \left|\mathrm{clip}_j\left(\phi_h^{\tau\top}\xi\right)\left(\sigma_\tau\left(s_h^\tau, a_h^\tau\right) + \mathbb{E}_{s'\sim p_\tau(\cdot|s_h^\tau, a_h^\tau)}V(w_{h+1})(s') - \sigma\left(s_h^\tau, a_h^\tau\right) + \mathbb{E}_{s'\sim p(\cdot|s_h^\tau, a_h^\tau)}V(w_{h+1})(s')\right)\right| \\
&\leq \ell_j c_\tau.
\end{aligned}$$

By Azuma's inequality, for a fixed tuple (t, h, j, ξ, w_{h+1}), with probability at least $1-\delta'$,

$$\begin{aligned}
&\left|\sum_{\tau=1}^{t-1}\mathrm{clip}_j\left(\phi_h^{\tau\top}\xi\right)\left(\phi_h^{\tau\top}\left(\rho + \int \nu(s')V(w_{h+1})(s')\mathrm{d}s'\right) - \left(\sigma_h^\tau + V(w_{h+1})(s_{h+1}^\tau)\right)\right)\right| \\
&\qquad \leq \ell_j\left(\sum_{\tau=1}^{t-1}c_\tau + 2\sqrt{t\log(T/\delta')}\right). \qquad (35)
\end{aligned}$$

Next, we take a union bound for Eq. (35) over $t \in [T]$, $h \in [H]$, $j \in [\lceil\log_2 T\rceil]$, and ξ, w_{h+1} in an $\frac{\ell_j}{2T}$-cover of $\mathcal{B}(2\sqrt{d})$ and $\mathcal{B}(\sqrt{d})$ respectively. By (Wu, 2016), the ϵ-covering number of a d-dimensional unit ball is upper bounded by $(3/\epsilon)^d$. Therefore, we get that with probability at least $1 - TH\lceil\log_2 T\rceil\left(3\times 4\sqrt{d}T/\ell_j\right)^{2d}\delta' \geq 1 - HT^2(12dT^2)^{2d}\delta'$, Eq. (35) holds for all possible t, h, j, and ξ, w_{h+1} in the $\frac{\ell_j}{2T}$-cover.

Therefore, for all possible t, h, j, ξ, and w_{h+1}, with probability at least $1 - HT^2(12dT^2)^{2d}\delta'$, the left-hand side of Eq. (35) is upper bounded by

$$\ell_j\left(\sum_{\tau=1}^{t-1}c_\tau + 2\sqrt{t\log(T/\delta')}\right) + \frac{\ell_j}{2T}\times t\times 2 + \ell_j \times \frac{\ell_j}{2T}\times t \times 2 \leq \ell_j\left(\sum_{\tau=1}^{t-1}c_\tau + 4\sqrt{t\log(T/\delta')}\right)$$

where we use the fact that $|\mathrm{clip}_j(\phi_h^{\tau\top}\xi) - \mathrm{clip}_j(\phi_h^{\tau\top}\xi')| \leq |\phi_h^\tau(\xi-\xi')| \leq \|\xi-\xi'\|$ and $|V(w_{h+1})(s) - V(w'_{h+1})(s)| = |\max_a w_{h+1}^\top\phi(s,a) - \max_a w'^\top_{h+1}\phi(s,a)| \leq \|w_{h+1} - w'_{h+1}\|$. Choosing $\delta' = \delta\Big/\left(T^2H(12dT^2)^{2d}\right)$ finishes the proof. ■

Corollary 36 *With probability at least $1-\delta$, $w^\star \in \mathcal{W}^t$ for all t.*

Proof It suffices to show that with probability at least $1-\delta$,

$$\left|\sum_{\tau=1}^{t-1}\mathrm{clip}_j\left(\phi_h^{\tau\top}\xi\right)\left(\phi_h^{\tau\top}w_h^\star - \left(\sigma_h^\tau + V(w_{h+1}^\star)(s_{h+1}^\tau)\right)\right)\right| \leq 200\cdot\ell_j\left(\sum_{\tau=1}^{t-1}c_\tau + \sqrt{dt\log(dHT/\delta)}\right)$$

for all t, h, ξ, j. This can be obtained by Lemma 35 with the fact that $w_h^\star = \rho + \int \nu(s')V_{h+1}^\star(s')\mathrm{d}s' = \rho + \int \nu(s')V(w_{h+1}^\star)(s')\mathrm{d}s'$. ■

Definition 37 $\xi_h^t \triangleq w_h^t - \left(\rho + \int \nu(s')V_{h+1}^t(s')\mathrm{d}s'\right)$.

Lemma 38 *With probability at least $1-\delta$, the following holds for all t, h, and j:*

$$\sum_{\tau=1}^{t-1} \mathit{clip}_j\left(\phi_h^{\tau\top}\xi_h^t\right)\phi_h^{\tau\top}\xi_h^t \le 400\cdot\ell_j\left(\sqrt{dt\log(dTH/\delta)}+C^{\mathsf{a}}\right).$$

Proof By the definition of ξ_h^t,

$$\begin{aligned}
&\sum_{\tau=1}^{t-1}\mathrm{clip}_j\left(\phi_h^{\tau\top}\xi_h^t\right)\phi_h^{\tau\top}\xi_h^t \\
&=\sum_{\tau=1}^{t-1}\mathrm{clip}_j\left(\phi_h^{\tau\top}\xi_h^t\right)\left(\phi_h^{\tau\top}w_h^t-\sigma_h^\tau-V_{h+1}^t(s_{h+1}^\tau)\right) \\
&\qquad+\sum_{\tau=1}^{t-1}\mathrm{clip}_j\left(\phi_h^{\tau\top}\xi_h^t\right)\left(\sigma_h^\tau+V_{h+1}^t(s_{h+1}^\tau)-\phi_h^{\tau\top}\left(\rho+\int\nu(s')V_{h+1}^t(s')\mathrm{d}s'\right)\right) \\
&\le 400\ell_j\left(C^{\mathsf{a}}+\sqrt{dt\log(dTH/\delta)}\right)
\end{aligned}$$

where we use the fact that $w^t\in\mathcal{W}^t$, and Lemma 35 with $w=w^t$. ■

Lemma 39 *With probability at least $1-\mathcal{O}(\delta)$,*

$$\sum_{\tau=1}^{t}\left(r_\tau^{\pi^\star}-r_\tau\right)\le\widetilde{\mathcal{O}}\left(Hd^{4.5}\sqrt{t}+Hd^4C^{\mathsf{a}}\right).$$

Proof Notice that

$$V_1^\star(s_1^t)-V^{\pi_t}(s_1^t)\le V_1^t(s_1^t)-V^{\pi_t}(s_1^t)=\phi_1^{t\top}(w_1^t-w_1^{\pi_t}). \tag{36}$$

where in the first equality we use the optimism of w_1^t. For any h,

$$\begin{aligned}
&\phi_h^{t\top}(w_h^t-w_h^{\pi_t}) \\
&=\phi_h^{t\top}w_h^t-\phi_h^{t\top}\rho-\phi_h^{t\top}\int\nu(s')V_{h+1}^{\pi_t}(s')\mathrm{d}s' \\
&=\phi_h^{t\top}\xi_h^t+\phi_h^{t\top}\int\nu(s')\left(V_{h+1}^t(s')-V_{h+1}^{\pi_t}(s')\right)\mathrm{d}s' \\
&=\phi_h^{t\top}\xi_h^t+\mathbb{E}_{s'\sim p(\cdot|s_h^t,a_h^t)}\left[V_{h+1}^t(s')-V_{h+1}^{\pi_t}(s')\right] \\
&\le\phi_h^{t\top}\xi_h^t+\mathbb{E}_{s'\sim p_t(\cdot|s_h^t,a_h^t)}\left[V_{h+1}^t(s')-V_{h+1}^{\pi_t}(s')\right]+\frac{2}{H}c_t \\
&=\phi_h^{t\top}\xi_h^t+\mathbb{E}\left[V_{h+1}^t(s_{h+1}^t)-V_{h+1}^{\pi_t}(s_{h+1}^t)\;\middle|\;s_h^t,a_h^t\right]+\frac{2}{H}c_t \\
&=\phi_h^{t\top}\xi_h^t+\frac{2}{H}c_t+\mathbb{E}\left[\phi_{h+1}^t(w_{h+1}^t-w_{h+1}^{\pi_t})\;\middle|\;s_h^t,a_h^t\right]
\end{aligned} \tag{37}$$

where in the inequality we use

$$\mathbb{E}_{s'\sim p(\cdot|s_h^t,a_h^t)}\left[V_{h+1}^t(s')-V_{h+1}^{\pi_t}(s')\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{s' \sim p_t(\cdot|s_h^t, a_h^t)} \left[V_{h+1}^t(s') - V_{h+1}^{\pi_t}(s') \right] \\
&\quad + \left(\sigma(s_h^t, a_h^t) + \mathbb{E}_{s' \sim p(\cdot|s_h^t, a_h^t)} \left[V_{h+1}^t(s') \right] - \sigma_t(s_h^t, a_h^t) - \mathbb{E}_{s' \sim p_t(\cdot|s_h^t, a_h^t)} \left[V_{h+1}^t(s') \right] \right) \\
&\quad - \left(\sigma(s_h^t, a_h^t) + \mathbb{E}_{s' \sim p(\cdot|s_h^t, a_h^t)} \left[V_{h+1}^{\pi}(s') \right] - \sigma_t(s_h^t, a_h^t) - \mathbb{E}_{s' \sim p_t(\cdot|s_h^t, a_h^t)} \left[V_{h+1}^{\pi}(s') \right] \right) \\
&\le \mathbb{E}_{s' \sim p_t(\cdot|s_h^t, a_h^t)} \left[V_{h+1}^t(s') - V_{h+1}^{\pi_t}(s') \right] + 2 \sup_{s', a'} \sup_{V \in [0,1]^{\mathcal{S}}} |\mathcal{T}_t V(s', a') - \mathcal{T} V(s', a')| \\
&= \mathbb{E}_{s' \sim p_t(\cdot|s_h^t, a_h^t)} \left[V_{h+1}^t(s') - V_{h+1}^{\pi_t}(s') \right] + \frac{2}{H} c_t.
\end{aligned}
$$

Combining Eq. (36) and Eq. (37), we get

$$
\sum_{\tau=1}^{t-1} \mathbb{E}\left[V_1^\star(s_1^\tau) - V^{\pi_\tau}(s_1^\tau) \right] \le \sum_{\tau=1}^{t-1} \sum_{h=1}^{H} {\phi_h^\tau}^\top \xi_h^\tau + 2 \sum_{\tau=1}^{t-1} c_\tau
$$

Applying Azuma-Hoeffding's inequality, we further get that with probability at least $1 - \mathcal{O}(\delta)$,

$$
\sum_{\tau=1}^{t-1} \left(r_\tau^{\pi^\star} - r_\tau \right) \le \sum_{\tau=1}^{t-1} \sum_{h=1}^{H} {\phi_h^\tau}^\top \xi_h^\tau + 2 \sum_{\tau=1}^{t-1} c_\tau + \widetilde{\mathcal{O}}(\sqrt{t}). \tag{38}
$$

It remains to bound $\sum_{\tau=1}^{t-1} {\phi_h^\tau}^\top \xi_h^\tau$ for all h:

$$
\begin{aligned}
&\sum_{\tau=1}^{t-1} {\phi_h^\tau}^\top \xi_h^\tau \\
&= \sum_{\tau=1}^{t-1} {\phi_h^\tau}^\top \xi_h^\tau \mathbf{1}[|{\phi_h^\tau}^\top \xi_h^\tau| \ge 1/(2t)] + \sum_{\tau=1}^{t-1} {\phi_h^\tau}^\top \xi_h^\tau \mathbf{1}[|{\phi_h^\tau}^\top \xi_h^\tau| < 1/(2t)] \\
&\le \sum_{\tau=1}^{t-1} {\phi_h^\tau}^\top \xi_h^\tau \times \frac{\sum_{s=1}^{\tau-1} \text{clip}_{j_\tau}\left({\phi_h^s}^\top \xi_h^\tau \right) {\phi_h^s}^\top \xi_h^\tau + \ell_{j_\tau}}{\sum_{s=1}^{\tau-1} \text{clip}_{j_\tau}\left({\phi_h^s}^\top \xi_h^\tau \right) {\phi_h^s}^\top \xi_h^\tau + \ell_{j_\tau}} + 1 \quad (j_\tau \text{ is such that } \tfrac{1}{2}\ell_{j_\tau} \le |{\phi_h^\tau}^\top \xi_h^\tau| \le \ell_{j_\tau}) \\
&\le \sum_{\tau=1}^{t-1} |{\phi_h^\tau}^\top \xi_h^\tau| \times \frac{\ell_{j_\tau} \times \widetilde{\mathcal{O}}\left(\sqrt{d\tau} + C^{\mathsf{a}} \right)}{\sum_{s=1}^{\tau-1} \text{clip}_{j_\tau}\left({\phi_h^s}^\top \xi_h^\tau \right) {\phi_h^s}^\top \xi_h^\tau + \ell_{j_\tau}} + 1 \qquad \text{(Lemma 38)} \\
&\le \left(\sum_{\tau=1}^{t-1} \frac{2\left(\text{clip}_{j_\tau}\left({\phi_h^\tau}^\top \xi_h^\tau \right) \right)^2}{\sum_{s=1}^{\tau-1} \text{clip}_{j_\tau}\left({\phi_h^s}^\top \xi_h^\tau \right) {\phi_h^s}^\top \xi_h^\tau + \ell_{j_\tau}} \right) \times \widetilde{\mathcal{O}}\left(\sqrt{dt} + C^{\mathsf{a}} \right) + 1 \\
&\le \widetilde{\mathcal{O}}(d^4) \times \widetilde{\mathcal{O}}\left(\sqrt{dt} + C^{\mathsf{a}} \right) \qquad \text{(by Lemma 40)}
\end{aligned}
$$

Combining this with Eq. (38) finishes the proof. ■

Lemma 40 (Lemma 20 of (Zhang et al., 2021c))

$$
\sum_{\tau=1}^{t-1} \frac{2\left(\mathit{clip}_{j_\tau}\left({\phi_h^\tau}^\top \xi_h^\tau \right) \right)^2}{\sum_{s=1}^{\tau-1} \mathit{clip}_{j_\tau}\left({\phi_h^s}^\top \xi_h^\tau \right) {\phi_h^s}^\top \xi_h^\tau + \ell_{j_\tau}} \le \mathcal{O}\left(d^4 \log^3(t) \right).
$$

Theorem 41 *For linear contextual bandits,* COBE *with Robust VOFUL as the base algorithm guarantees* $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(d^{4.5}\sqrt{T} + d^4 C^{\mathsf{a}}\right)$. *For linear MDPs,* COBE *with Robust VARLin as the base algorithm guarantees* $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(Hd^{4.5}\sqrt{T} + Hd^4 C^{\mathsf{a}}\right)$.

Proof By Lemma 33, we see that Robust OFUL satisfies Eq. (5) with $\beta_1 = \widetilde{\Theta}(d^9), \beta_2 = \widetilde{\Theta}(d^4), \beta_3 = \Theta(1)$. Using them in Theorem 3 gives the desired bound for linear contextual bandits. For Robust LSVI-UCB, $\beta_1 = \widetilde{\Theta}(d^9H^2), \beta_2 = \widetilde{\Theta}(d^4H), \beta_3 = \Theta(1)$. Using them in Theorem 3, we get the desired bound for linear MDPs. ■

G.5. Robust GOLF

In this section, we adapt the GOLF algorithm by Jin et al. (2021a) to the corruption setting. For simplicity, we assume that the function class $\mathcal{F}$ is finite (the extension to infinite case is straightforward through a discretization step, as shown in (Jin et al., 2021a)). The algorithm is presented in Algorithm 8.

Algorithm 8 Robust GOLF

input: C^{r}
parameter: $\zeta = 16\log(TH|\mathcal{F}|/\delta)$.
Initialize: $\mathcal{B}^1 \leftarrow \mathcal{F}$
for $t = 1, 2, \ldots, T$ **do**

Choose policy: $\pi^t = \pi_{f^t}$, where $f^t \in \mathrm{argmax}_{f\in\mathcal{B}^t} f(s_1, \pi_f(s_1))$
Collect a trajectory $(s_1^t, a_1^t, \sigma_1^t, \ldots, s_H^t, a_H^t, \sigma_H^t, s_{H+1}^t)$ by following π^t.
Update

$$\mathcal{B}^{t+1} = \left\{ f \in \mathcal{F} \colon \mathcal{L}_h^t(f_h, f_{h+1}) \le \inf_{g\in\mathcal{F}_h} \mathcal{L}_h^t(g, f_{h+1}) + \left(\zeta + \frac{2C^{\mathsf{r}}}{H^2 t}\right) \text{ for all } h \in [H] \right\}, \tag{39}$$

$$\text{where} \quad \mathcal{L}_h^t(f_h, f_{h+1}) = \sum_{\tau=1}^{t} \left(f_h(s_h^\tau, a_h^\tau) - \sigma_h^\tau - \max_{a'} f_{h+1}(s_{h+1}^\tau, a') \right)^2. \tag{40}$$

end

Lemma 42 (*c.f.* Lemma 39 of (Jin et al., 2021a)) *With probability at least* $1-\delta$*, we have*

$$(a) \qquad \sum_{\tau=1}^{t-1} \mathbb{E}\left[\left(f_h^t(s_h, a_h) - (\mathcal{T} f_{h+1}^t)(s_h, a_h) \right)^2 \;\middle|\; s_h, a_h \sim \pi_\tau \right] \le \mathcal{O}\left(\zeta + \frac{C^{\mathsf{r}}}{H^2 t}\right)$$

$$(b) \qquad \sum_{\tau=1}^{t-1} \left(f_h^t(s_h^\tau, a_h^\tau) - (\mathcal{T} f_{h+1}^t)(s_h^\tau, a_h^\tau) \right)^2 \le \mathcal{O}\left(\zeta + \frac{C^{\mathsf{r}}}{H^2 t}\right)$$

Proof Define for any $h \in [H], g \in \mathcal{F}$,

$$X_h^t(g) \triangleq \left(g_h(s_h^t, a_h^t) - \sigma_h^t - g_{h+1}(s_{h+1}^t, \pi_g(s_{h+1}^t))\right)^2 - \left((\mathcal{T} g_{h+1})(s_h^t, a_h^t) - \sigma_h^t - g_{h+1}(s_{h+1}^t, \pi_g(s_{h+1}^t))\right)^2.$$

Then we have

$$
\begin{aligned}
& X_h^t(g) \\
& = (g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t))^2 \\
& \quad + 2\left((\mathcal{T} g_{h+1})(s_h^t, a_h^t) - \sigma_h^t - g_{h+1}(s_{h+1}^t, \pi_g(s_{h+1}^t))\right)\left(g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t)\right) \\
& = (g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t))^2 \\
& \quad + 2\left((\mathcal{T} g_{h+1})(s_h^t, a_h^t) - \sigma_t(s_h^t, a_h^t) - \mathbb{E}_{s' \sim p_t(\cdot|s_h^t, a_h^t)}\left[g_{h+1}(s', \pi_g(s'))\right]\right)\left(g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t)\right) \\
& \quad + \underbrace{2\left(\sigma_t(s_h^t, a_h^t) - \sigma_h^t + \mathbb{E}_{s' \sim p_t(\cdot|s_h^t, a_h^t)}\left[g_{h+1}(s', \pi_g(s'))\right] - g_{h+1}(s_{h+1}^t, \pi_g(s_{h+1}^t))\right)}_{\triangleq \epsilon_h^t}\left(g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t)\right) \\
& \geq \left(g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t)\right)^2 - \frac{2}{H} c_t \left|g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t)\right| + \epsilon_h^t \left(g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t)\right) \\
& \hspace{20em} \text{(by the definition of } c_t\text{)} \\
& \geq \frac{1}{2}\left(g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t)\right)^2 - \frac{2c_t^2}{H^2} + \epsilon_h^t \left(g_h(s_h^t, a_h^t) - (\mathcal{T} g_{h+1})(s_h^t, a_h^t)\right) \quad \text{(AM-GM)}
\end{aligned} \tag{41}
$$

Notice that ϵ_h^t is a zero-mean random variable. By the definition of $\mathcal{B}^t$ and that $f^t \in \mathcal{B}^t$, we have

$$
\begin{aligned}
& \sum_{\tau=1}^{t-1} X_h^\tau(f^t) \\
& = \sum_{\tau=1}^{t-1}\left[\left(f_h^t(s_h^\tau, a_h^\tau) - \sigma_h^\tau - f_{h+1}^t(s_{h+1}^\tau, \pi_{f^t}(s_{h+1}^\tau))\right)^2 - \left(\mathcal{T} f_{h+1}^t(s_h^\tau, a_h^\tau) - \sigma_h^\tau - f_{h+1}^t(s_{h+1}^\tau, \pi_{f^t}(s_{h+1}^\tau))\right)^2\right] \\
& \leq \sum_{\tau=1}^{t-1}\left[\left(f_h^t(s_h^\tau, a_h^\tau) - \sigma_h^\tau - f_{h+1}^t(s_{h+1}^\tau, \pi_{f^t}(s_{h+1}^\tau))\right)^2 - \min_{g \in \mathcal{F}_h}\left(g(s_h^\tau, a_h^\tau) - \sigma_h^\tau - f_{h+1}^t(s_{h+1}^\tau, \pi_{f^t}(s_{h+1}^\tau))\right)^2\right] \\
& \hspace{20em} \text{(by the closeness of } \mathcal{F}\text{)} \\
& \leq \zeta + \frac{2C^{\mathsf{r}}}{H^2 t}.
\end{aligned}
$$

Combining this with Eq. (41), we get

$$
\begin{aligned}
& \frac{1}{2}\sum_{\tau=1}^{t-1}\left(f_h^t(s_h^\tau, a_h^\tau) - (\mathcal{T} f_{h+1}^t)(s_h^\tau, a_h^\tau)\right)^2 \\
& \leq \sum_{\tau=1}^{t-1} X_h^\tau(f^t) + \frac{2}{H^2}\sum_{\tau=1}^{t-1} c_\tau^2 - \sum_{\tau=1}^{t-1} \epsilon_h^\tau \left(f_h^t(s_h^\tau, a_h^\tau) - (\mathcal{T} f_{h+1}^t)(s_h^\tau, a_h^\tau)\right) \\
& \leq \sum_{\tau=1}^{t-1} X_h^\tau(f^t) + \frac{2C^{\mathsf{r}}}{H^2 t} + 2\sqrt{\sum_{\tau=1}^{t-1}\left(f_h^t(s_h^\tau, a_h^\tau) - (\mathcal{T} f_{h+1}^t)(s_h^\tau, a_h^\tau)\right)^2 \log(TH|\mathcal{F}|/\delta)} \\
& \hspace{20em} \text{(Freedman's inequality)} \\
& \leq \zeta + \frac{4C^{\mathsf{r}}}{H^2 t} + \frac{1}{4}\sum_{\tau=1}^{t-1}\left(f_h^t(s_h^\tau, a_h^\tau) - (\mathcal{T} f_{h+1}^t)(s_h^\tau, a_h^\tau)\right)^2 + 4\log(TH|\mathcal{F}|/\delta) \qquad \text{(AM-GM)}
\end{aligned}
$$

The above inequality implies

$$\sum_{\tau=1}^{t-1}\left(f_h^t(s_h^\tau,a_h^\tau)-(\mathcal{T}f_{h+1}^t)(s_h^\tau,a_h^\tau)\right)^2 \le (4\zeta+16\log(TH|\mathcal{F}|/\delta))+\frac{16C^{\mathsf{r}}}{H^2t}=\mathcal{O}\left(\zeta+\frac{C^{\mathsf{r}}}{H^2t}\right),$$

proving (b). (a) can be proven by the same approach (see also (Jin et al., 2021a)). ■

Lemma 43 *With probability at least $1-\delta$, the optimal Q-function of the uncorrupted MDP is always feasible, that is, $Q^\star \in \mathcal{B}^t$ for all $t\in[T]$.*

Proof Following the proof of Lemma 40 in Jin et al. (2021a), we define for any $h\in[H]$, $t\in[T]$ and $g\in\mathcal{F}$,

$$\begin{aligned}
W_h^t(g) &\triangleq \left(g_h(s_h^t,a_h^t)-\sigma_h^t-V^\star(s_{h+1}^t)\right)^2-\left(Q^\star(s_h^t,a_h^t)-\sigma_h^t-V^\star(s_{h+1}^t)\right)^2\\
&= \left(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t)\right)^2+2\left(Q^\star(s_h^t,a_h^t)-\sigma_h^t-V^\star(s_{h+1}^t)\right)\left(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t)\right)\\
&= (g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t))^2\\
&\quad +2\left(Q^\star(s_h^t,a_h^t)-\sigma_t(s_h^t,a_h^t)-\mathbb{E}_{s'\sim p_t(s'|s_h^t,a_h^t)}[V^\star(s')]\right)\left(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t)\right)\\
&\quad +\underbrace{2\left(\sigma_t(s_h^t,a_h^t)-\sigma_h^t+\mathbb{E}_{s'\sim p_t(s'|s_h^t,a_h^t)}[V^\star(s')]-V^\star(s_{h+1}^t)\right)}_{\epsilon_h^t}(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t))\\
&\ge (g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t))^2-\frac{2}{H}c_t|g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t)|+\epsilon_h^t(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t))\\
&\ge \frac{1}{2}(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t))^2-\frac{2}{H^2}c_t^2+\epsilon_h^t(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t)) \qquad \text{(AM-GM)}
\end{aligned}$$

Let $\mathfrak{F}_h^t$ be the sigma-field induced by all samples up to s_h^t, a_h^t (but not σ_h^t or s_{h+1}^t). Then

$$\mathbb{E}[W_h^t(g)\mid \mathfrak{F}_h^t] \ge \frac{1}{2}(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t))^2-\frac{2}{H^2}c_t^2\,. \tag{42}$$

and by the definition of $W_h^t(g)$, the variance is bounded by

$$\mathbb{E}[W_h^t(g)^2\mid \mathfrak{F}_{t,h}] \le 4(g_h(s_h^t,a_h^t)-Q^\star(s_h^t,a_h^t))^2\,. \tag{43}$$

By Freedman's inequality, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\frac{1}{2}\sum_{\tau=1}^{t-1}(g_h(s_h^\tau,a_h^\tau)-Q_h^\star(s_h^\tau,a_h^\tau))^2-\sum_{\tau=1}^{t-1}W_h^\tau(g)\\
&\le \frac{2}{H^2}\sum_{\tau=1}^{t-1}c_\tau^2+2\sqrt{4\sum_{\tau=1}^{t-1}(g_h(s_h^\tau,a_h^\tau)-Q_h^\star(s_h^\tau,a_h^\tau))^2\log(TH|\mathcal{F}|/\delta)} \quad \text{(by Eq. (42) and Eq. (43))}\\
&\le \frac{2C^{\mathsf{r}}}{H^2t}+\frac{1}{4}\sum_{\tau=1}^{t-1}(g_h(s_h^\tau,a_h^\tau)-Q_h^\star(s_h^\tau,a_h^\tau))^2+16\log(TH|\mathcal{F}|/\delta) \qquad \text{(AM-GM)}
\end{aligned}$$

which implies

$$-\sum_{\tau=1}^{t-1} W_h^\tau(g) \leq \frac{2C^{\mathsf{r}}}{H^2 t} + 16\log(TH|\mathcal{F}|/\delta).$$

This implies that $Q^\star \in \mathcal{B}^t$. ■

Lemma 44 *With probability at least* $1 - \mathcal{O}(\delta)$,

$$\sum_{\tau=1}^{t}(r_\tau^{\pi^\star} - r_\tau) = \widetilde{\mathcal{O}}\left(H\sqrt{\zeta \mathrm{dim}_{\mathrm{DE}} \cdot t} + \sqrt{\mathrm{dim}_{\mathrm{DE}}}C^{\mathsf{r}}\right).$$

where ζ is defined in Algorithm 8, and $\mathrm{dim}_{\mathrm{DE}}$ is the Bellman eluder dimension. We refer the reader to (Jin et al., 2021a) for the precise definition of the Bellman eluder dimension.

Proof

$$\begin{aligned}
&\sum_{\tau=1}^{t}(r_\tau^{\pi^\star} - r_\tau) \\
&\leq \sum_{\tau=1}^{t}(V^{\pi^\star}(s_1^\tau) - V^{\pi_\tau}(s_1^\tau)) + \widetilde{\mathcal{O}}\left(\sqrt{t} + C^{\mathsf{a}}\right) && \text{(Azuma's inequality)} \\
&\leq \sum_{\tau=1}^{t}(\max_a f^\tau(s_1^\tau, a) - V^{\pi_\tau}(s_1^\tau)) + \widetilde{\mathcal{O}}\left(\sqrt{t} + C^{\mathsf{a}}\right) \\
&\leq \sum_{\tau=1}^{t}\sum_{h=1}^{H}\mathbb{E}\left[f_h^\tau(s_h, a_h) - \mathcal{T} f_{h+1}^\tau(s_h, a_h) \,\middle|\, (s_h, a_h) \sim \pi_\tau\right] + \widetilde{\mathcal{O}}\left(\sqrt{t} + C^{\mathsf{a}}\right) \\
& && \text{(by (Jin et al., 2021a, Eq.(4)))} \\
&\leq \sum_{h=1}^{H}\widetilde{\mathcal{O}}\left(\sqrt{\mathrm{dim}_{\mathrm{DE}} \cdot t}\sqrt{\zeta + \frac{1}{H^2}\frac{(C^{\mathsf{r}})^2}{t}}\right) + \widetilde{\mathcal{O}}\left(\sqrt{t} + C^{\mathsf{a}}\right) \\
& && \text{(using Lemma 42 together with (Jin et al., 2021a, Lemma 17))} \\
&= \widetilde{\mathcal{O}}\left(H\sqrt{\zeta \mathrm{dim}_{\mathrm{DE}} \cdot t} + \sqrt{\mathrm{dim}_{\mathrm{DE}}}C^{\mathsf{r}}\right).
\end{aligned}$$

■

Theorem 45 *For MDPs with low Bellman-eluder dimension,* COBE *with Robust GOLF as the base algorithm guarantees* $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(H\sqrt{\zeta \mathrm{dim}_{\mathrm{DE}} \cdot T} + \sqrt{\mathrm{dim}_{\mathrm{DE}}}C^{\mathsf{r}}\right)$.

Proof By Lemma 44, we see that Robust GOLF satisfies Eq. (5) with $\beta_1 = \widetilde{\Theta}(H^2\zeta\mathrm{dim}_{\mathrm{DE}}), \beta_2 = \widetilde{\Theta}(\sqrt{\mathrm{dim}_{\mathrm{DE}}}), \beta_3 = \Theta(1)$. Using them in Theorem 3 gives the desired bound. ■