Maximize to Explore: One Objective Function Fusing Estimation, Planning, and Exploration

Zhihan Liu[1][*]   **Miao Lu**[2][*]   **Wei Xiong**[3][*]   **Han Zhong**[4]   **Hao Hu**[5]
Shenao Zhang[1]   **Sirui Zheng**[1]   **Zhuoran Yang**[6]   **Zhaoran Wang**[1]
[1]Northwestern University   [2]Stanford University   [3]University of Illinois Urbana-Champaign
[4]Peking University   [5]Tsinghua University   [6]Yale University
{zhihanliu2027,shenaozhang2028,siruizheng2025}@u.northwestern.edu
miaolu@stanford.edu, wx13@illinois.edu, hanzhong@stu.pku.edu.cn
huh22@mails.tsinghua.edu.cn, zhuoran.yang@yale.edu, zhaoranwang@gmail.com

Abstract

In reinforcement learning (RL), balancing exploration and exploitation is crucial for achieving an optimal policy in a sample-efficient way. To this end, existing sample-efficient algorithms typically consist of three components: estimation, planning, and exploration. However, to cope with general function approximators, most of them involve impractical algorithmic components to incentivize exploration, such as data-dependent level-set constraints or complicated sampling procedures. To address this challenge, we propose an easy-to-implement RL framework called *Maximize to Explore* (MEX), which only needs to optimize *unconstrainedly* a single objective that integrates the estimation and planning components while balancing exploration and exploitation automatically. Theoretically, we prove that the MEX achieves a sublinear regret with general function approximators and is extendable to the zero-sum Markov game setting. Meanwhile, we adapt deep RL baselines to design practical versions of MEX in both the model-based and model-free settings, which outperform baselines in various MuJoCo environments with sparse reward by a stable margin. Compared with existing sample-efficient algorithms with general function approximators, MEX achieves similar sample efficiency while also enjoying a lower computational cost and is more compatible with modern deep RL methods. Our codes are available at https://github.com/agentification/MEX.

1 Introduction

The crux of online reinforcement learning (online RL) lies in maintaining a balance between exploiting the current knowledge of the agent about the environment and exploring unfamiliar areas [69]. To fulfill this, agents in existing sample-efficient RL algorithms predominantly undertake three tasks: i) *estimate* a hypothesis using historical data to encapsulate their understanding of the environment; ii) perform *planning* based on the estimated hypothesis to exploit their current knowledge; iii) further *explore* the unknown environment via carefully designed exploration strategies.

There exists a long line of research on integrating the aforementioned three components harmoniously, to find optimal policies in a sample-efficient manner. From theoretical perspectives, existing theories aim to minimize the notion of *online external regret* which measures the cumulative suboptimality gap of the policies learned during online learning. It is well studied that one can design both statistically and computationally efficient algorithms (e.g., upper confidence bound (UCB), [6, 39, 12, 91]) with sublinear online regret for tabular and linear Markov decision processes (MDPs). But when it

---

[*]Equal contribution.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

comes to MDPs with general function approximations, most of them involve impractical algorithmic components to incentivize exploration. Usually, to cope with general function approximations, agents need to solve constrained optimization problems within data-dependent level-sets [37, 20], or sample from complicated posterior distributions over the space of hypotheses [19, 2, 89], both of which pose considerable challenges for implementation. From a practical perspective, a prevalent approach in deep RL for balancing exploration and exploitation is to use an ensemble of neural networks [75, 59, 14, 53, 44, 18, 45], which serves as an empirical approximation of the UCB method. However, such an ensemble method suffers from high computational cost and lacks theoretical guarantee when the underly MDP is neither linear nor tabular. As for other deep RL algorithms for exploration [29, 3, 11, 9, 17, 65], such as curiosity-driven method [60], it also remains unknown in theory whether they are provably sample-efficient in the context of general function approximations.

Hence, in this paper, we are aimed at tackling these issues and answering the following question:

Under general function approximation, can we design a sample-efficient and
easy-to-implement RL framework to trade off between exploration and exploitation?

To this end, we propose an easy-to-implement RL framework, *Maximize to Explore* (MEX), as an affirmative answer to the question. To strike a balance between exploration and exploitation, MEX propose to maximize a weighted sum of two objectives: (i) the optimal expected total return associated with a given hypothesis and (ii) the negative estimation error of that hypothesis. Consequently, MEX naturally combines planning and estimation components in just a single objective. By choosing the hypothesis that maximizes the weighted sum and executing the optimal policy with respect to the chosen hypothesis, MEX automatically balances between exploration and exploitation.

We highlight that the objective of MEX is *not* obtained by taking the Lagrange dual of the constrained optimization objective within data-dependent level-sets [37, 20, 15].This is because the coefficient of the weighted sum, which remains fixed, is data-independent and predetermined for all episodes. Contrary to Lagrangian methods, MEX does not necessitate an inner loop of optimization for dual variables, thereby circumventing the complications associated with minimax optimization. As a maximization-only framework, MEX is friendly to implementations with neural networks and does not rely on sampling or ensemble.

In the theory part, we prove that MEX achieves a sublinear regret $\widetilde{\mathcal{O}}(\mathrm{Poly}(H) \cdot d_{\mathrm{GEC}}^{1/2}(1/\sqrt{HK}) \cdot K^{1/2})$ under mild assumptions and is thus sample-efficient, where K is the number of episodes and H is the horizon length. Here $d_{\mathrm{GEC}}(\cdot)$ is the Generalized Eluder Coefficient (GEC) [89] that characterizes the complexity of learning the underlying MDP under general function approximations. Because the class of low-GEC MDPs includes almost all known theoretically tractable MDP instances, our proved result can be tailored to a multitude of specific settings with either a model-free or a model-based hypothesis, such as MDPs with low Bellman eluder dimension [37], MDPs of bilinear class [20], and MDPs with low witness rank [67]. Besides, thanks to the flexibility of the MEX framework, we further extend it to online RL in two-player zero-sum Markov games (MGs), for which we further extend the definition of GEC to two-player zero-sum MGs and establish the sample efficiency with general function approximations. Moving beyond theory and into practice, we adapt famous RL baselines TD3 [27] and MBPO [34] to design practical versions of MEX in model-free and model-based fashion, respectively. On various MuJoCo environments [71] with sparse rewards, experimental results show that MEX outperforms baselines steadily and significantly. Compared with other deep RL algorithms, MEX has low computational overhead and straightforward implementation while maintaining a theoretical guarantee.

Contributions. We conclude our contributions from three perspectives.

1. We propose an easy-to-implement RL algorithm framework MEX that *unconstrainedly* maximizes a single objective to fuse estimation and planning, automatically trading off between exploration and exploitation. Under mild structural assumptions, we prove that MEX achieves a sublinear regret $\widetilde{\mathcal{O}}(\mathrm{Poly}(H) \cdot d_{\mathrm{GEC}}^{1/2}(1/\sqrt{HK})K^{1/2})$ with general function approximators, and thus is sample-efficient. Here $d_{\mathrm{GEC}}(\cdot)$ is the generalized Eluder Coefficient (GEC) of the underlying MDP.
2. We instantiate the generic MEX framework to several model-based and model-free examples and establish corresponding theoretical results. Further, we extend the MEX framework to two-player zero-sum MGs and also prove the sample efficiency with an extended definition of GEC.
3. We design practical implementations of MEX for MDPs in both model-based and model-free styles. Experiments on various MuJoCo environments with sparse reward demonstrate the effectiveness of our proposed MEX framework.

2

1.1 Related work

Sample-efficient RL with function approximation. The success of DRL methods has motivated a line of work focused on function approximation scenarios. This work originated in the linear case [74, 81, 12, 39, 84, 5, 82, 57, 91, 90] and is later extended to general function approximation. Wang et al. [73] first study the general function approximation using the notion of eluder dimension [63], which takes the linear MDP [39] as a special case but with inferior results. Zanette et al. [85] consider a different type of framework based on Bellman completeness, which assumes that the class used for approximating the optimal Q-functions is closed in terms of the Bellman operator and improves the results for linear MDP. After this, Jin et al. [37] consider the eluder dimension of the class of Bellman residual associated with the RL problems, which captures more solvable problems. Another line of works focuses on the low-rank structures of the problems, where Jiang et al. [35] propose the Bellman rank for model-free RL and Sun et al. [67] propose the witness rank for model-based RL. Following these two works, Du et al. [20] propose the bilinear class, which contains more MDP models with low-rank structures [6, 67, 39, 57, 12, 91] by allowing a flexible choice of discrepancy function class. However, it is known that neither BE nor bilinear class captures each other. Dann et al. [19] first consider eluder-coefficient-type complexity measure on the Q-type model-free RL. It was later extended by Zhong et al. [89] to cover all the above-known solvable problems in both model-free and model-based manners. Foster et al. [25, 23] study another notion of complexity measure, the decision-estimation coefficient (DEC), which also unifies the Bellman eluder dimension and bilinear class. The DEC framework is appealing due to the matching lower bound in some decision-making problems, where all other complexity measures do not have. However, due to the presence of a minimax subroutine in its definition, they require a much more complicated minimax optimization oracle and cannot apply to the classical optimism-based or sampling-based methods. Chen et al. [13], Foster et al. [24] extend the vanilla DEC (to the model-free case) by incorporating an optimistic modification, which was originally referred to as the feel-good modification in Zhang [87]. Chen et al. [15] study Admissible Bellman Characterization (ABC) class to generalize BE. They also extend the GOLF algorithm and Bellman completeness in model-free RL to the model-based case by considering more general (vector-form) discrepancy loss functions to construct sharper in-sample error estimators and obtain sharper bounds compared to Sun et al. [67]. Xie et al. [79] connect the online RL with the coverage condition in the offline RL, and also study the GOLF algorithm proposed in Jin et al. [37].

Algorithmic design in sample-efficient RL with function approximation. The most prominent approach in this area is based on the principle of "Optimism in the Face of Uncertainty" (OFU), which dates back to Auer et al. [4]. For instance, for linear function approximation, Jin et al. [39] propose an optimistic variant of Least-Squares Value Iteration (LSVI), which achieves optimism by adding a bonus at each step. For the general case, Jiang et al. [36] first propose an elimination-based algorithm with optimism in model-free RL and is extended to model-based RL by [67]. After these, Du et al. [20], Jin et al. [37] propose two OFU-based algorithms, which are more similar to the lin-UCB algorithm [1] studied in the linear contextual bandit literature. The model-based counterpart (Optimistic Maximum Likelihood Estimation) is studied in Liu et al. [46], Chen et al. [13]. Specifically, these algorithms explicitly maintain a confidence set that contains the ground truth with high probability and conducts a constraint optimization step at each iteration to select the most optimistic hypothesis in the confidence set. The other line of work studies another powerful algorithmic framework based on posterior sampling. For instance, Zanette et al. [84] study randomized least-squares value iteration (RLSVI), which can be interpreted as a sampling-based algorithm and achieves an order-optimal result for linear MDP. For general function approximation, the works mainly follow the idea of the "feel-good" modification proposed in Zhang [87]. These algorithms start from some prior distribution over the hypothesis space and update the posterior distribution according to the collected samples but with certain optimistic modifications in either the prior or the loglikelihood function. Then the hypothesis for each iteration is sampled from the posterior and guides data collection. In particular, Dann et al. [19] studies the model-free Q-type problem, and Agarwal and Zhang [2] studies the model-based problems, but under different notions of complexity measures. Zhong et al. [89] further utilize the idea in Zhang [87] and extend the posterior sampling algorithm in Dann et al. [19] to be a unified sampling-based framework to solve both model-free and model-based RL problems, which is also shown to apply to the more challenging partially observable setting. In addition to the OFU-based algorithm and the sampling-based framework, Foster et al. [25] propose the Estimation-to-Decisions (E2D) algorithm, which can solve problems with low DEC but requires solving a complicated minimax subroutine to fit in the framework of DEC.

3

Due to the limitation of the page, we defer the remaining discussions of **Relationship with reward-biased maximum likelihood estimation**, **Exploration of DRL**, and **Two-player Zero-Sum Markov Game** to Appendix B.

2 Preliminaries

2.1 Episodic Markov Decision Processes and Online Reinforcement Learning

We consider an episodic MDP defined by a tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, H is a finite horizon, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ with $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ the transition kernel at the h-th timestep, and $r = \{r_h\}_{h=1}^H$ with $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ the reward function at the h-th timestep. Without loss of generality, we assume that the reward function r_h is both deterministic and known.

In the episodic MDP, the agent interacts with the environment by the following *online* protocol. At the beginning of the k-th episode, the agent selects a policy $\pi^k = \{\pi_h^k : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h=1}^H$. It takes an action $a_h^k \sim \pi_h^k(\cdot \mid x_h^k)$ at timestep h and state x_h^k. After receiving the reward $r_h^k = r_h(x_h^k, a_h^k)$ from the environment, it transits to the next state $x_{h+1}^k \sim \mathbb{P}_h(\cdot \mid x_h^k, a_h^k)$. When the agent reaches the state x_{H+1}^k, it ends the k-th episode. Without loss of generality, we assume that the initial state $x_1 = \underline{x}$ is fixed. Our analysis can be generalized to the setting where x_1 is sampled from a distribution on $\mathcal{S}$.

Policy and value functions. For a policy $\pi = \{\pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h=1}^H$, we denote by $V_h^\pi : \mathcal{S} \mapsto \mathbb{R}$ and $Q_h^\pi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ its value function and state-action value function at the h-th timestep, which characterizes the expected total rewards received by the agent under policy π afterward, starting from some $x_h = x \in \mathcal{S}$ (or $x_h = x \in \mathcal{S}, a_h = a \in \mathcal{A}$, resp.), till the end of the episode. Specifically,

$$V_h^\pi(x) := \mathbb{E}_\pi \left[\sum_{h'=h}^H r_{h'}(x_{h'}, a_{h'}) \middle| x_h = x \right], \quad \forall x \in \mathcal{S}, \tag{2.1}$$

$$Q_h^\pi(x, a) := \mathbb{E}_\pi \left[\sum_{h'=h}^H r_{h'}(x_{h'}, a_{h'}) \middle| x_h = x, a_h = a \right], \quad \forall (x, a) \in \mathcal{S} \times \mathcal{A}. \tag{2.2}$$

We know there exists an optimal policy π^* which has the optimal value function for all initial states [61], that is, $V_h^{\pi^*}(s) = \sup_\pi V_h^\pi(x)$ for all $h \in [H]$ and $x \in \mathcal{S}$. For simplicity, we abbreviate V^{π^*} as V^* and the optimal state-action value function Q^{π^*} as Q^*. Moreover, the optimal value functions Q^* and V^* satisfy the following Bellman optimality equation [61], given by

$$V_h^*(x) = \max_{a \in \mathcal{A}} Q_h^*(x, a), \qquad V_{H+1}^*(x) = 0, \tag{2.3}$$

$$Q_h^*(x, a) = \left(\mathcal{T}_h Q_{h+1}^* \right)(x, a) := r_h(x, a) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot \mid x, a)}[V_{h+1}^*(x')], \tag{2.4}$$

for all $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. We call $\mathcal{T}_h$ the Bellman optimality operator at timestep h. Also, for any two functions Q_h and Q_{h+1} on $\mathcal{S} \times \mathcal{A}$, we define

$$\mathcal{E}_h(Q_h, Q_{h+1}; x, a) := Q_h(x, a) - \mathcal{T}_h Q_{h+1}(x, a), \quad \forall (x, a) \in \mathcal{S} \times \mathcal{A}, \tag{2.5}$$

as the Bellman residual at timestep h of (Q_h, Q_{h+1}).

Performance metric. We measure the online performance of an agent after K episodes by *regret*. We assume that the learner predicts the optimal policy π^* via π^k in the k-th episode for each $k \in [K]$. Then the regret of K episodes is defined as the cumulative suboptimality gap of $\{\pi^k\}_{k \in [K]}$[2],

$$\text{Regret}(K) = \sum_{k=1}^K V_1^*(x_1) - V_1^{\pi^k}(x_1). \tag{2.6}$$

The target of sample-efficient online RL is to achieve sublinear regret (2.6) with respect to K.

2.2 Function Approximation: Model-Free and Model-Based Hypothesis

To handle MDPs with large or even infinite state space, we introduce a family of function approximators. In specific, we consider a hypothesis class $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_H$, which can be specified to model-based and model-free settings respectively. Also, we denote by $\Pi = \Pi_1 \times \cdots \times \Pi_H$ as the space of Markovian policies. We now specify $\mathcal{H}$ for model-free and model-based settings.

---

[2]We allow the agent to predict the optimal policy via π^k while executing some other exploration policy π_{exp}^k to interact with the environment and collect data, as is considered in the related literature [67, 20, 89]

4

Algorithm 1 Maximize to Explore (`MEX`)

1: **Input**: Hypothesis class $\mathcal{H}$, parameter $\eta > 0$.
2: **for** $k = 1, \cdots, K$ **do**
3:     Solve $f^k \in \mathcal{H}$ via

$$f^k = \underset{f \in \mathcal{H}}{\arg\sup} \left\{ V_{1,f}(x_1) - \eta \cdot \sum_{h=1}^{H} L_h^{k-1}(f) \right\}. \tag{3.1}$$

4:     Execute $\pi_{\exp}(f^k)$ to collect data $\mathcal{D}^k = \{\mathcal{D}_h^k\}_{h \in [H]}$ with $\mathcal{D}_h^k = (x_h^k, a_h^k, r_h^k, x_{h+1}^k)$.
5:     Predict the optimal policy via π_{f^k}.
6: **end for**

Example 2.1 (Model-free hypothesis). *For model-free hypothesis class, $\mathcal{H}$ contains state-action value functions of the MDP, i.e., $\mathcal{H}_h \subseteq \{f_h : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}\}$. Specifically, for any $f = (f_1, \cdots, f_H) \in \mathcal{H}$, we denote $Q_f = \{Q_{h,f}\}_{h \in [H]}$ with $Q_{h,f} = f_h$. Also, we denote the corresponding optimal state-value function $V_f = \{V_{h,f}\}_{h \in [H]}$ with $V_{h,f}(\cdot) = \max_{a \in \mathcal{A}} Q_{h,f}(\cdot, a)$ and denote the corresponding optimal policy by $\pi_f = \{\pi_{h,f}\}_{h \in [H]}$ with $\pi_{h,f}(\cdot) = \arg\max_{a \in \mathcal{A}} Q_{h,f}(\cdot, a)$. Finally, we denote the optimal state-action value function under the true model, i.e., Q^*, by f^*.*

Example 2.2 (Model-based hypothesis). *For model-based hypothesis class, $\mathcal{H}$ contains models of the MDP, i.e., the transition kernel. Specifically, we denote $f = \mathbb{P}_f = (\mathbb{P}_{1,f}, \cdots, \mathbb{P}_{H,f}) \in \mathcal{H}$. For any $(f, \pi) \in \mathcal{H} \times \Pi$, we define $V_f^\pi = \{V_{h,f}^\pi\}_{h \in [H]}$ as the state-value function induced by model $\mathbb{P}_f$ and policy π. We use $V_f = \{V_{h,f}\}_{h \in [H]}$ to denote the corresponding optimal state-value function, i.e., $V_{h,f} = \sup_{\pi \in \Pi} V_{h,f}^\pi$. The corresponding optimal policy is denoted by $\pi_f = \{\pi_{h,f}\}_{h \in [H]}$, where $\pi_{h,f} = \arg\sup_{\pi \in \Pi} V_{h,f}^\pi$. Finally, we denote the true model $\mathbb{P}$ of the MDP as f^*.*

We remark that the main difference between the model-based hypothesis (Example 2.2) and the model-free hypothesis (Example 2.1) is that model-based RL directly learns the transition kernel of the underlying MDP, while model-free RL learns the optimal state-action value function. Since we do not add any specific structural form to the hypothesis class, e.g., linear function or kernel function, we are in the context of *general function approximations* [67, 37, 20, 89, 15].

3   Algorithm: Maximize to Explore for Online RL

In this section, we propose **Maximize to Explore** (`MEX`, Algorithm 1) for online RL in MDPs with general function approximations. With a novel single objective, `MEX` automatically balances the goal of exploration and exploitation. We first give a generic algorithm framework and then instantiate it to model-free (Example 2.1) and model-based (Example 2.2) hypotheses respectively.

Generic algorithm design. In each episode $k \in [K]$, the agent first estimates a hypothesis $f^k \in \mathcal{H}$ using historical data $\{\mathcal{D}^s\}_{s=1}^{k-1}$ by maximizing a composite objective (3.1). Specifically, in order to achieve exploiting history knowledge while encouraging exploration, the agent considers a single objective that sums: **(a)** the negative loss $-L_h^{k-1}(f)$ induced by the hypothesis f, which represents the exploitation of the agent's current knowledge; **(b)** the expected total return of the optimal policy associated with this hypothesis, i.e., $V_{1,f}$, which represents exploration for a higher return. With a tuning parameter $\eta > 0$, the agent balances the weight put on the tasks of exploitation and exploration. The agent then predicts π^* via the optimal policy with respect to the hypothesis that maximizes the composite exploration-exploitation objective function, i.e., π_{f^k}. Also, the agent executes certain exploration policy $\pi_{\exp}(f^k)$ to collect data $\mathcal{D}^k = \{(x_h^k, a_h^k, r_h^k, x_{h+1}^k)\}_{h=1}^{H}$ and updates the loss function $L_h^k(f)$. The choice of $\pi_{\exp}(f^k)$ depends on the specific MDP structure, and we refer to examples in Section 5 for detailed discussions.

We highlight that `MEX` is **not** a Lagrangian duality of constrained optimization objectives within data-dependent level-sets [37, 20, 15]. In fact, `MEX` only needs to fix the parameter η across each episode. Thus η is independent of data and predetermined, which contrasts Lagrangian methods that involve an inner loop of optimization for the dual variables. We also remark that we can rewrite (3.1) as a joint optimization $(f, \pi) = \arg\sup_{f \in \mathcal{H}, \pi \in \Pi} V_{1,f}^\pi(x_1) - \eta \sum_{h=1}^{H} L_h^{k-1}(f)$. When η tends to infinity, `MEX` can be reduced to vanilla Actor-Critic framework [41], where critic f minimizes estimation error and actor π conducts greedy policy following the critic f. In the following two parts,

5

we instantiate Algorithm 1 to model-based and mode-free hypotheses by specifying the loss function $L_h^k(f)$.

Model-free algorithm. For model-free hypothesis (Example 2.1), (3.1) becomes

$$f^k = \underset{f \in \mathcal{H}}{\text{argsup}} \left\{ \max_{a_1 \in \mathcal{A}} Q_{1,f}(x_1, a_1) - \eta \cdot \sum_{h=1}^{H} L_h^{k-1}(f) \right\}. \tag{3.2}$$

Regarding the choice of the loss function, for seek of theoretical analysis, to deal with MDPs with low Bellman eluder dimension [37] and MDPs of bilinear class [20], we assume the existence of certain function l which generalizes the notion of Bellman residual.

Assumption 3.1. *Suppose the function* $l : \mathcal{H} \times \mathcal{H}_h \times \mathcal{H}_{h+1} \times (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}) \mapsto \mathbb{R}$ *satisfies:*
*i) (**Generalized Bellman completeness**) [89, 15] there exists an operator* $\mathcal{P}_h : \mathcal{H}_{h+1} \mapsto \mathcal{H}_h$ *such that for any* $(f', f_h, f_{h+1}) \in \mathcal{H} \times \mathcal{H}_h \times \mathcal{H}_{h+1}$ *and* $\mathcal{D}_h = (x_h, a_h, r_h, x_{h+1}) \in \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$, *it holds*

$$l_{f'}\big((f_h, f_{h+1}), \mathcal{D}_h\big) - l_{f'}\big((\mathcal{P}_h f_{h+1}, f_{h+1}), \mathcal{D}_h\big) = \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h, a_h)}\big[l_{f'}\big((f_h, f_{h+1}), \mathcal{D}_h\big)\big],$$

where we require that $\mathcal{P}_h f_{h+1}^* = f_h^*$ *and that* $\mathcal{P}_h f_{h+1} \in \mathcal{H}_h$ *for any* $f_{h+1} \in \mathcal{H}_{h+1}$ *and step* $h \in [H]$;
*ii) (**Boundedness**) it holds that* $\sup_{f' \in \mathcal{H}} \|l_{f'}((f_h, f_{h+1}), \mathcal{D}_h)\|_\infty \leq B_l$ *for some constant* $B_l > 0$.

We then set the loss function L_h^k as an empirical estimation of the generalized squared Bellman error $|\mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h, a_h)}[l_{f^s}((f_h, f_{h+1}), \mathcal{D}_h^s)]|^2$, given by

$$L_h^k(f) = \sum_{s=1}^{k} l_{f^s}\big((f_h, f_{h+1}), \mathcal{D}_h^s\big)^2 - \inf_{f_h' \in \mathcal{H}_h} \sum_{s=1}^{k} l_{f^s}\big((f_h', f_{h+1}), \mathcal{D}_h^s\big)^2. \tag{3.3}$$

We remark that the subtracted infimum term in (3.3) is to handle the variance terms in the estimation to achieve a fast theoretical rate. Similar essential ideas are also adopted by [37, 78, 19, 38, 52, 2, 89, 15].

Model-based algorithm. For model-based hypothesis (Example 2.2), (3.1) becomes

$$f^k = \underset{f \in \mathcal{H}}{\text{argsup}} \left\{ \sup_{\pi \in \Pi} V_{1,\mathbb{P}_f}^\pi(x_1) - \eta \sum_{h=1}^{H} L_h^{k-1}(f) \right\}, \tag{3.4}$$

which is a joint optimization over the model $\mathbb{P}_f$ and the policy π. In the model-based algorithm, we choose the loss function L_h^k as the negative log-likelihood loss, defined as

$$L_h^k(f) = -\sum_{s=1}^{k} \log \mathbb{P}_{h,f}(x_{h+1}^s | x_h^s, a_h^s). \tag{3.5}$$

4 Regret Analysis for MEX Framework

In this section, we establish a regret analysis for the MEX framework (Algorithm 1). We give a generic theoretical guarantee which holds for both model-free and model-based settings. We first present three key assumptions needed for sample-efficient learning with MEX. In Section 5, we specify the generic theory to specific examples of MDPs and hypothesis classes satisfying these assumptions.

Firstly, we assume that the hypothesis class $\mathcal{H}$ is well-specified, containing the true hypothesis f^*.

Assumption 4.1 (Realizablity). *We assume that the true hypothesis* $f^* \in \mathcal{H}$.

Then we need to make a structural assumption on the MDP to ensure sample-efficient online learning. Inspired by Zhong et al. [89], we require the MDP to have low *Generalized Eluder Coefficient* (GEC). A low GEC indicates that the agent can effectively mitigate out-of-sample prediction errors by minimizing in-sample errors derived from historical data. To introduce, we define a discrepancy function $\ell_{f'}(\hat{f}; \xi_h) : \mathcal{H} \times \mathcal{H} \times (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}) \mapsto \mathbb{R}$ which characterizes the error of a hypothesis $f \in \mathcal{H}$ on data $\xi_h = (x_h, a_h, r_h, x_{h+1})$. Specific choices of ℓ are given in Section 5 for concrete model-free and model-based examples.

Assumption 4.2 (Low Generalized Eluder Coefficient [89]). *We assume that given an* $\epsilon > 0$, *there exists* $d(\epsilon) \in \mathbb{R}_+$, *such that for any sequence of* $\{f^k\}_{k \in [K]} \subset \mathcal{H}$, $\{\pi_{\exp}(f^k)\}_{k \in [K]} \subset \Pi$,

$$\sum_{k=1}^{K} V_{1,f^k} - V_1^{\pi_{f^k}} \leq \inf_{\mu > 0} \left\{ \frac{\mu}{2} \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f^k; \xi_h)] + \frac{d(\epsilon)}{2\mu} + \sqrt{d(\epsilon)HK} + \epsilon HK \right\}.$$

We denote the smallest number $d(\epsilon) \in \mathbb{R}_+$ *satisfying this condition as* $d_{\text{GEC}}(\epsilon)$.

6

As is shown by Zhong et al. [89], the low-GEC MDP class covers almost all known theoretically tractable MDP instances, such as linear MDP [81, 39], linear mixture MDP [5, 57, 12], MDPs of low witness rank [67], MDPs of low Bellman eluder dimension [37], and MDPs of bilinear class [20].

Finally, we make a concentration-style assumption which characterizes how the loss function L_h^k is related to the expectation of the discrepancy function $\mathbb{E}[\ell]$ appearing in the definition of GEC. For ease of presentation, we assume that $\mathcal{H}$ is finite, i.e., $|\mathcal{H}| < \infty$, but our result can be directly extended to an infinite $\mathcal{H}$ using covering number arguments [72, 37, 49, 38].

Assumption 4.3 (Generalization). *We assume that $\mathcal{H}$ is finite, i.e., $|\mathcal{H}| < +\infty$, and that with probability at least $1 - \delta$, for any episode $k \in [K]$ and hypothesis $f \in \mathcal{H}$, it holds that*

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) \lesssim - \sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f; \xi_h)] + B\big(H \log(HK/\delta) + \log(|\mathcal{H}|)\big),$$

where $B = B_l^2$ for model-free hypothesis (Assumption 3.1) and $B = 1$ for model-based hypothesis.

Such a concentration style inequality is well known in the literature of online RL with general function approximation and similar analysis is also adopted by [37, 15]. With Assumptions 4.1, 4.2, and 4.3, we can present our main result (see Appendix D.1 for a proof).

Theorem 4.4 (Online regret of MEX (Algorithm 1)). *Under Assumptions 4.1, 4.2, and 4.3, by setting*

$$\eta = \sqrt{\frac{d_{\mathrm{GEC}}(1/\sqrt{HK})}{(H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B \cdot K}},$$

then the regret of Algorithm 1 after K episodes is upper bounded by

$$\mathrm{Regret}(K) \lesssim \sqrt{d_{\mathrm{GEC}}(1/\sqrt{HK}) \cdot (H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B \cdot K},$$

with probability at least $1 - \delta$. Here $d_{\mathrm{GEC}}(\cdot)$ is defined in Assumption 4.2.

Theorem 4.4 shows that the regret of Algorithm 1 scales with the square root of the number of episodes K and the polynomials of horizon H, GEC $d_{\mathrm{GEC}}(1/\sqrt{K})$, and log covering number $\log \mathcal{N}(\mathcal{H}, 1/K)$. When the number of episodes K tends to infinity, the average regret vanishes, meaning that the output policy of Algorithm 1 achieves global optimality. Since the regret of Algorithm 1 is sublinear with respect to the number of episodes K, Algorithm 1 is proved to be sample-efficient. In Appendix C, we extend the algorithm framework and the analysis to the two-player zero-sum Markov game (MG) setting, for which we also extend the definition of GEC to two-player zero-sum MGs.

Besides, as we can see from Theorem 4.4 and its specifications in Section 5, MEX matches existing theoretical results in the literature of online RL with general function approximations [89, 67, 20, 37, 19, 2]. But in the meanwhile, MEX does not require explicitly solving a constrained optimization problem within data-dependent level-sets or performing a complex sampling procedure. This advantage makes MEX a principled approach with easier practical implementation. We conduct deep RL experiments for MEX in Section 6 to demonstrate its power in complicated online problems.

5 Examples of MEX Framework

In this section, we specify Algorithm 1 to model-based and model-free hypothesis classes for various examples of MDPs of low GEC (Assumption 4.2), including MDPs with low witness rank [67], MDPs with low Bellman eluder dimension [37], and MDPs of bilinear class [20]. For ease of presentation, we assume that $|\mathcal{H}| < \infty$, but our result can be directly extended to infinite $\mathcal{H}$ using covering number arguments [72, 37, 49]. All the proofs of the propositions in this section are in Appendix E.

We note that another important step in specifying Theorem 4.4 to concrete hypothesis classes is to check Assumption 4.3 (supervised learning guarantee). It is worth highlighting that, in our analysis, for both model-free and model-based hypotheses, we provide supervised learning guarantees in a neat and unified manner, independent of specific MDP structures.

5.1 Model-free online RL in Markov Decision Processes

In this subsection, we specify Algorithm 1 for model-free hypothesis class $\mathcal{H}$ (Example 2.1). For a model-free hypothesis class, we choose the discrepancy function ℓ as, given $\mathcal{D}_h = (x_h, a_h, r_h, x_{h+1})$,

$$\ell_{f'}(f; \mathcal{D}_h) = \big(\mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h, a_h)}[l_{f'}((f_h, f_{h+1}), \mathcal{D}_h)]\big)^2. \tag{5.1}$$

7

where the function $l : \mathcal{H} \times \mathcal{H}_h \times \mathcal{H}_{h+1} \times (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}) \mapsto \mathbb{R}$ satisfies Assumption 3.1. We specify the choice of l in concrete examples of MDPs later. In the following, we check and specify Assumptions 4.2 and 4.3 in Section 4 for model-free hypothesis classes.

Proposition 5.1 (Generalization: model-free RL). *We assume that $\mathcal{H}$ is finite, i.e., $|\mathcal{H}| < +\infty$. Then under Assumption 3.1, with probability at least $1 - \delta$, for any $k \in [K]$ and $f \in \mathcal{H}$, it holds that*

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) \lesssim -\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f; \xi_h)] + B_l^2\big(H\log(HK/\delta) + \log(|\mathcal{H}|)\big),$$

where L and ℓ are defined in (3.3) *and* (5.1) *respectively. Here B_l is specified in Assumption 3.1.*

Proposition 5.1 specifies Assumption 4.3 for model-free hypothesis classes. For Assumption 4.2, we need structural assumptions on the MDP. Given an MDP with generalized eluder dimension d_{GEC}, we have the following corollary of our main theoretical result (Theorem 4.4).

Corollary 5.2 (Online regret of MEX: model-free hypothesis). *Given an MDP with generalized eluder coefficient $d_{\text{GEC}}(\cdot)$ and a finite model-free hypothesis class $\mathcal{H}$ with $f^* \in \mathcal{H}$, under Assumption 3.1, setting*

$$\eta = \sqrt{\frac{d_{\text{GEC}}(1/\sqrt{HK})}{(H\log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B_l^2 \cdot K}}, \tag{5.2}$$

then the regret of Algorithm 1 after K episodes is upper bounded by

$$\text{Regret}(T) \lesssim B_l \cdot \sqrt{d_{\text{GEC}}(1/\sqrt{HK}) \cdot (H\log(HK/\delta) + \log(|\mathcal{H}|)) \cdot K}, \tag{5.3}$$

with probability at least $1 - \delta$. Here B_l is specified in Assumption 3.1.

Corollary 5.2 can be directly specified to MDPs with low GEC, including MDPs with low Bellman eluder dimension [37] and MDPs of bilinear class [20]. See Appendix E.1 for a detailed discussion.

5.2 Model-based online RL in Markov Decision Processes

In this subsection, we specify Algorithm 1 for model-based hypothesis class $\mathcal{H}$ (Example 2.2). For model-based hypothesis class, we choose the discrepancy function ℓ in Assumption 4.2 and 4.3 as the hellinger distance. Given data $\mathcal{D}_h = (x_h, a_h, r_h, x_{h+1})$, we let

$$\ell_{f'}(f; \mathcal{D}_h) = D_{\text{H}}(\mathbb{P}_{h,f}(\cdot|x_h, a_h)\|\mathbb{P}_{h,f^*}(\cdot|x_h, a_h)), \tag{5.4}$$

where $D_{\text{H}}(\cdot\|\cdot)$ denotes the Hellinger distance. We note that by (5.4), the discrepancy function ℓ does not depend on the input $f' \in \mathcal{H}$. In the following, we check and specify Assumption 4.2 and 4.3.

Proposition 5.3 (Generalization: model-based RL). *We assume that $\mathcal{H}$ is finite, i.e., $|\mathcal{H}| < +\infty$. Then with probability at least $1 - \delta$, for any $k \in [K]$, $f \in \mathcal{H}$, it holds that*

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) \lesssim -\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f; \xi_h)] + H\log(H/\delta) + \log(|\mathcal{H}|),$$

where L and ℓ are defined in (3.5) *and* (5.4) *respectively.*

Proposition 5.3 specifies Assumption 4.3 for model-based hypothesis classes. For Assumption 4.2, we also need structural assumptions on the MDP. Given an MDP with generalized eluder dimension d_{GEC}, we have the following corollary of our main theoretical result (Theorem 4.4).

Corollary 5.4 (Online regret of MEX: model-based hypothesis). *Given an MDP with generalized eluder coefficient $d_{\text{GEC}}(\cdot)$ and a finite model-based hypothesis class $\mathcal{H}$ with $f^* \in \mathcal{H}$, by setting*

$$\eta = \sqrt{\frac{d_{\text{GEC}}(1/\sqrt{HK})}{(H\log(H/\delta) + \log(|\mathcal{H}|)) \cdot K}}, \tag{}$$

then the regret of Algorithm 1 after K episodes is upper bounded by, with probability at least $1 - \delta$,

$$\text{Regret}(K) \lesssim \sqrt{d_{\text{GEC}}(1/\sqrt{HK}) \cdot (H\log(H/\delta) + \log(|\mathcal{H}|)) \cdot K}, \tag{5.5}$$

Corollary 5.4 can be directly specified to MDPs with low GEC, including MDPs with low witness rank [67]. We refer to Appendix E.2 for a detailed discussion.

8

6 Experiments

In this section, we aim to answer the following two questions: **(a)** What are the practical approaches to implementing MEX in both model-based (MEX-MB) and model-free (MEX-MF) settings? **(b)** Can MEX handle challenging exploration tasks, especially in sparse reward scenarios?

Experimental setups. We evaluate the effectiveness of MEX by assessing its performance in both standard gym locomotion tasks and sparse reward locomotion and navigation tasks within the MuJoCo [71] environment. For sparse reward tasks, we select cheetah-vel, walker-vel, hopper-vel, ant-vel, and ant-goal adapted from Yu et al. [83], where the agent receives a reward only when it successfully attains the desired velocity or goal. To adapt to deep RL settings, we consider infinite-horizon γ-discounted MDPs and MEX variants. We report the results averaged over five random seeds. The full experimental settings are in Appendix H.

Implementation details. For the model-based variant MEX-MB, we use the following objective:

$$\max_\phi \max_\pi \mathbb{E}_{(x,a,r,x')\sim\beta} \left[\log \mathbb{P}_\phi(x',r\,|\,x,a) \right] + \eta' \cdot \mathbb{E}_{x\sim\sigma} \left[V^\pi_{\mathbb{P}_\phi}(x) \right], \quad (6.1)$$

where we denote by σ the initial state distribution, β the replay buffer, and η' corresponds to $1/\eta$ in the previous theory sections. We leverage the *score function* to obtain the model value gradient $\nabla_\phi V^\pi_{\mathbb{P}_\phi}$ in a similar way to likelihood ratio policy gradient [70], with the gradient of action log-likelihood replaced by the gradient of state and reward log-likelihood in the model. Specifically,

$$\nabla_\phi \mathbb{E}_{x\sim\sigma} \left[V^\pi_{\mathbb{P}_\phi}(x) \right] = \mathbb{E}_{\tau^\pi_\phi} \left[\left(r + \gamma V^\pi_{\mathbb{P}_\phi}(x') - Q^\pi_{\mathbb{P}_\phi}(x,a) \right) \cdot \nabla_\phi \log \mathbb{P}_\phi(x',r\,|\,x,a) \right], \quad (6.2)$$

where τ^π_ϕ is the trajectory under policy π and transition $\mathbb{P}_\phi$, starting from σ. We refer the readers to previous works [62, 76] for a derivation of (6.2). The model ϕ and policy π in (6.1) are updated iteratively in a Dyna [68] style, where model-free policy updates are performed on model-generated data. Particularly, we adopt SAC [30] to update the policy π and estimate the value $Q^\pi_{\mathbb{P}_\phi}$ by performing temporal difference on the model data generated by $\mathbb{P}_\phi$. We also follow [62] to update the model using mini-batches from β and normalize the advantage $r_h + \gamma V^\pi_{\mathbb{P}_\phi} - Q^\pi_{\mathbb{P}_\phi}$ within each mini-batch.

For the model-free variant MEX-MF, we observe from (3.2) that adding a maximization bias term to the standard TD error is sufficient for exploration. However, this may lead to instabilities as the bias term only involves the state-action value function of the current policy, and thus the policy may be ever-changing. To address this issue, we adopt a similar treatment as in CQL [42] by subtracting a baseline state-action value from random policy $\mu = \text{Unif}(\mathcal{A})$ and obtain the following objective:

$$\max_\theta \max_\pi \mathbb{E}_\beta \left[\left(r + \gamma Q_\theta(x',a') - Q_\theta(x,a) \right)^2 \right] + \eta' \cdot \mathbb{E}_\beta \left[\mathbb{E}_{a\sim\pi} Q_\theta(x,a) - \mathbb{E}_{a\sim\mu} Q_\theta(x,a) \right]. \quad (6.3)$$

We update θ and π in (6.3) iteratively in an actor-critic fashion. Due to space limits, we refer the readers to Appendix H for more implementation details of MEX-MF.

Results. We report the performance of MEX-MB and MEX-MF in Figures 1 and 2, respectively. We compare MEX-MB with MBPO [34], where our method differs from MBPO *only* in the inclusion of the value gradient in (6.2) during model updates. We find that MEX-MB offers an easy implementation with minimal computational overhead and yet remains highly effective across sparse and standard MuJoCo tasks. Notably, in the sparse reward settings, MEX-MB excels at achieving the goal velocity and outperforms MBPO by a stable margin. In standard gym tasks, MEX-MB showcases greater sample efficiency in challenging high-dimensional tasks with higher asymptotic returns.

We then compare MEX-MF with the model-free baseline TD3 [27]. We observe that TD3 fails in many sparse reward tasks, while MEX-MF significantly boosts the performance. In standard MuJoCo gym tasks, MEX-MF also steadily outperforms TD3 with faster convergence and higher final returns.

7 Conclusions

In this paper, we introduce a novel RL algorithm framework—*Maximize to Explore* (MEX)—aimed at striking a balance between exploration and exploitation in online learning scenarios. MEX is provably sample-efficient under general function approximations and is easy to implement. Theoretically, we prove that under mild structural assumptions (low generalized eluder coefficient (GEC)), MEX achieves

9
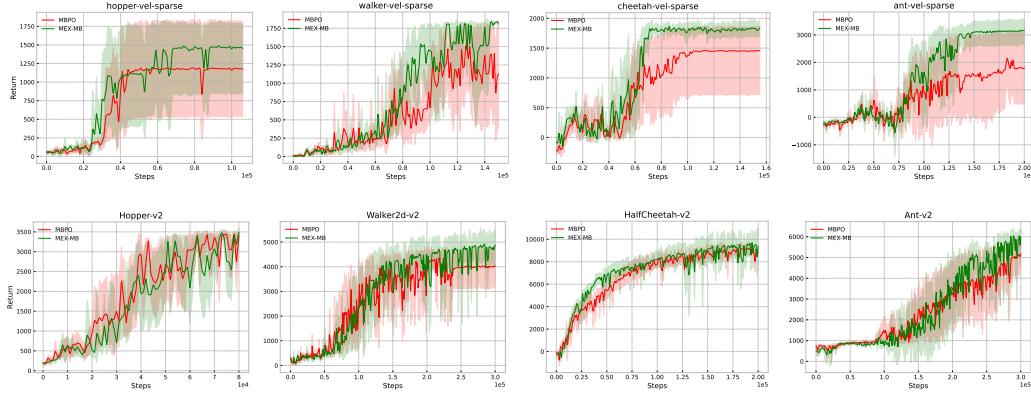
Figure 1: Model-based `MEX-MB` in sparse and standard MuJoCo locomotion tasks.
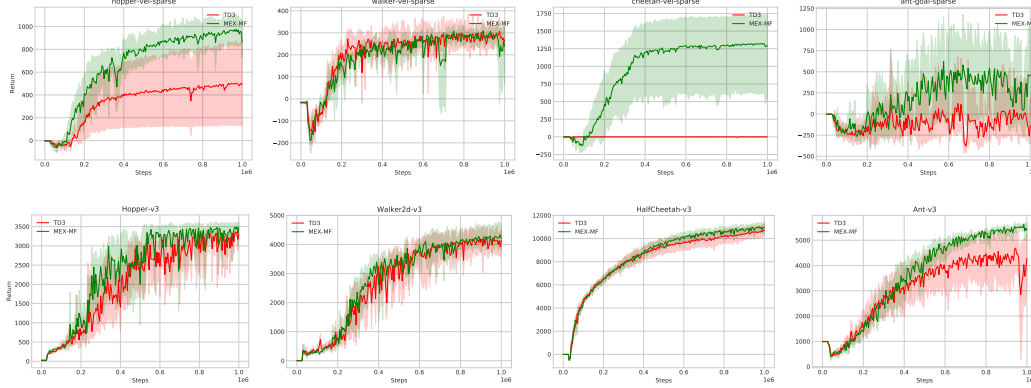
Figure 2: Model-free `MEX-MF` in sparse and standard MuJoCo locomotion tasks.

$\widetilde{\mathcal{O}}(\sqrt{K})$-online regret for MDPs. We further extend the definition of GEC and `MEX` framework to two-player zero-sum Markov games (see Appendix C) and also prove the $\widetilde{\mathcal{O}}(\sqrt{K})$-online regret. In practice, we adapt `MEX` to deep RL methods in both model-based and model-free styles and apply them to sparse-reward MuJoCo environments, outperforming baselines significantly. We hope our work can shed light on future research of designing both statistically efficient and practically effective RL algorithms with powerful function approximation.

Acknowledgement

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports.

References

[1] ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24**.

[2] AGARWAL, A. and ZHANG, T. (2022). Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659* .

[3] AUBRET, A., MATIGNON, L. and HASSAS, S. (2019). A survey on intrinsic motivation in reinforcement learning.

[4] AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning* **47** 235–256.

10

[5] AYOUB, A., JIA, Z., SZEPESVARI, C., WANG, M. and YANG, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*. PMLR.

[6] AZAR, M. G., OSBAND, I. and MUNOS, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*. PMLR.

[7] BAI, Y. and JIN, C. (2020). Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*. PMLR.

[8] BAI, Y., JIN, C. and YU, T. (2020). Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007* .

[9] BELLEMARE, M., SRINIVASAN, S., OSTROVSKI, G., SCHAUL, T., SAXTON, D. and MUNOS, R. (2016). Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems* **29**.

[10] BLANCHET, J., LU, M., ZHANG, T. and ZHONG, H. (2023). Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659* .

[11] BURDA, Y., EDWARDS, H., STORKEY, A. and KLIMOV, O. (2018). Exploration by random network distillation. *arXiv preprint arXiv:1810.12894* .

[12] CAI, Q., YANG, Z., JIN, C. and WANG, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*. PMLR.

[13] CHEN, F., MEI, S. and BAI, Y. (2022). Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745* .

[14] CHEN, R. Y., SIDOR, S., ABBEEL, P. and SCHULMAN, J. (2017). Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502* .

[15] CHEN, Z., LI, C. J., YUAN, A., GU, Q. and JORDAN, M. I. (2022). A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634* .

[16] CHEN, Z., ZHOU, D. and GU, Q. (2021). Almost optimal algorithms for two-player Markov games with linear function approximation. *arXiv preprint arXiv:2102.07404* .

[17] CHOI, J., GUO, Y., MOCZULSKI, M., OH, J., WU, N., NOROUZI, M. and LEE, H. (2018). Contingency-aware exploration in reinforcement learning. *arXiv preprint arXiv:1811.01483* .

[18] CHUA, K., CALANDRA, R., MCALLISTER, R. and LEVINE, S. (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems* **31**.

[19] DANN, C., MOHRI, M., ZHANG, T. and ZIMMERT, J. (2021). A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems* **34** 12040–12051.

[20] DU, S., KAKADE, S., LEE, J., LOVETT, S., MAHAJAN, G., SUN, W. and WANG, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*. PMLR.

[21] EYSENBACH, B., KHAZATSKY, A., LEVINE, S. and SALAKHUTDINOV, R. R. (2022). Mismatched no more: Joint model-policy optimization for model-based rl. *Advances in Neural Information Processing Systems* **35** 23230–23243.

[22] FILAR, J. and VRIEZE, K. (2012). *Competitive Markov decision processes*. Springer Science & Business Media.

[23] FOSTER, D. J., GOLOWICH, N. and HAN, Y. (2023). Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215* .

11

[24] FOSTER, D. J., GOLOWICH, N., QIAN, J., RAKHLIN, A. and SEKHARI, A. (2022). A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv preprint arXiv:2211.14250* .

[25] FOSTER, D. J., KAKADE, S. M., QIAN, J. and RAKHLIN, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487* .

[26] FREEDMAN, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability* 100–118.

[27] FUJIMOTO, S., HOOF, H. and MEGER, D. (2018). Addressing function approximation error in actor-critic methods. In *International conference on machine learning*. PMLR.

[28] GUO, H., CAI, Q., ZHANG, Y., YANG, Z. and WANG, Z. (2022). Provably efficient offline reinforcement learning for partially observable markov decision processes. In *International Conference on Machine Learning*. PMLR.

[29] HAARNOJA, T., ZHOU, A., ABBEEL, P. and LEVINE, S. (2018). Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on machine learning (ICML-18)*.

[30] HAARNOJA, T., ZHOU, A., ABBEEL, P. and LEVINE, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR.

[31] HOUTHOOFT, R., CHEN, X., DUAN, Y., SCHULMAN, J., DE TURCK, F. and ABBEEL, P. (2016). Vime: Variational information maximizing exploration. *Advances in neural information processing systems* **29**.

[32] HUANG, B., LEE, J. D., WANG, Z. and YANG, Z. (2021). Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702* .

[33] HUNG, Y.-H., HSIEH, P.-C., LIU, X. and KUMAR, P. (2021). Reward-biased maximum likelihood estimation for linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35.

[34] JANNER, M., FU, J., ZHANG, M. and LEVINE, S. (2019). When to trust your model: Model-based policy optimization. *Advances in neural information processing systems* **32**.

[35] JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*. PMLR.

[36] JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*. PMLR.

[37] JIN, C., LIU, Q. and MIRYOOSEFI, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems* **34**.

[38] JIN, C., LIU, Q. and YU, T. (2022). The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*. PMLR.

[39] JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR.

[40] JIN, Y., YANG, Z. and WANG, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*. PMLR.

[41] KONDA, V. and TSITSIKLIS, J. (1999). Actor-critic algorithms. In *Advances in Neural Information Processing Systems* (S. Solla, T. Leen and K. Müller, eds.), vol. 12. MIT Press.

12

[42] KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* **33** 1179–1191.

[43] KUMAR, P. and BECKER, A. (1982). A new family of optimal adaptive controllers for markov chains. *IEEE Transactions on Automatic Control* **27** 137–146.

[44] KURUTACH, T., CLAVERA, I., DUAN, Y., TAMAR, A. and ABBEEL, P. (2018). Model-ensemble trust-region policy optimization. *arXiv preprint arXiv:1802.10592* .

[45] LEE, K., LASKIN, M., SRINIVAS, A. and ABBEEL, P. (2021). Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*. PMLR.

[46] LIU, Q., CHUNG, A., SZEPESVÁRI, C. and JIN, C. (2022). When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967* .

[47] LIU, Q., YU, T., BAI, Y. and JIN, C. (2020). A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604* .

[48] LIU, X., HSIEH, P.-C., HUNG, Y. H., BHATTACHARYA, A. and KUMAR, P. (2020). Exploration through reward biasing: Reward-biased maximum likelihood estimation for stochastic multi-armed bandits. In *International Conference on Machine Learning*. PMLR.

[49] LIU, Z., LU, M., WANG, Z., JORDAN, M. and YANG, Z. (2022). Welfare maximization in competitive equilibrium: Reinforcement learning for markov exchange economy. In *International Conference on Machine Learning*. PMLR.

[50] LIU, Z., ZHANG, Y., FU, Z., YANG, Z. and WANG, Z. (2022). Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *International Conference on Machine Learning*. PMLR.

[51] LOPES, M., LANG, T., TOUSSAINT, M. and OUDEYER, P.-Y. (2012). Exploration in model-based reinforcement learning by empirically estimating learning progress. *Advances in neural information processing systems* **25**.

[52] LU, M., MIN, Y., WANG, Z. and YANG, Z. (2022). Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. *arXiv preprint arXiv:2205.13589* .

[53] LU, X. and VAN ROY, B. (2017). Ensemble sampling. *Advances in neural information processing systems* **30**.

[54] METE, A., SINGH, R. and KUMAR, P. (2022). Augmented rbmle-ucb approach for adaptive control of linear quadratic systems. *Advances in Neural Information Processing Systems* **35** 9302–9314.

[55] METE, A., SINGH, R. and KUMAR, P. (2022). The rbmle method for reinforcement learning. In *2022 56th Annual Conference on Information Sciences and Systems (CISS)*. IEEE.

[56] METE, A., SINGH, R., LIU, X. and KUMAR, P. (2021). Reward biased maximum likelihood estimation for reinforcement learning. In *Learning for Dynamics and Control*. PMLR.

[57] MODI, A., JIANG, N., TEWARI, A. and SINGH, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

[58] MOHAMED, S. and JIMENEZ REZENDE, D. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems* **28**.

[59] OSBAND, I., BLUNDELL, C., PRITZEL, A. and VAN ROY, B. (2016). Deep exploration via bootstrapped dqn. *Advances in neural information processing systems* **29**.

13

[60] PATHAK, D., AGRAWAL, P., EFROS, A. A. and DARRELL, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*. PMLR.

[61] PUTERMAN, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[62] RIGTER, M., LACERDA, B. and HAWES, N. (2022). Rambo-rl: Robust adversarial model-based offline reinforcement learning. *arXiv preprint arXiv:2204.12581* .

[63] RUSSO, D. and VAN ROY, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems* **26**.

[64] STADIE, B. C., LEVINE, S. and ABBEEL, P. (2015). Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814* .

[65] SUN, H., HAN, L., YANG, R., MA, X., GUO, J. and ZHOU, B. (2022). Exploit reward shifting in value-based deep-rl: Optimistic curiosity-based exploration and conservative exploitation via linear reward shaping. *Advances in Neural Information Processing Systems* **35** 37719–37734.

[66] SUN, H., XU, Z., FANG, M. and ZHOU, B. (2022). Supervised q-learning can be a strong baseline for continuous control. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

[67] SUN, W., JIANG, N., KRISHNAMURTHY, A., AGARWAL, A. and LANGFORD, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*. PMLR.

[68] SUTTON, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*. Elsevier, 216–224.

[69] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement learning: An introduction*.

[70] SUTTON, R. S., MCALLESTER, D., SINGH, S. and MANSOUR, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* **12**.

[71] TODOROV, E., EREZ, T. and TASSA, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE.

[72] WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge university press.

[73] WANG, R., SALAKHUTDINOV, R. R. and YANG, L. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems* **33** 6123–6135.

[74] WANG, Y., WANG, R., DU, S. S. and KRISHNAMURTHY, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136* .

[75] WIERING, M. A. and VAN HASSELT, H. (2008). Ensemble algorithms in reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **38** 930–936.

[76] WU, C., LI, T., ZHANG, Z. and YU, Y. (2022). Bayesian optimistic optimization: Optimistic exploration for model-based reinforcement learning. *Advances in Neural Information Processing Systems* **35** 14210–14223.

[77] XIE, Q., CHEN, Y., WANG, Z. and YANG, Z. (2020). Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*. PMLR.

14

[78] XIE, T., CHENG, C.-A., JIANG, N., MINEIRO, P. and AGARWAL, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems* **34** 6683–6694.

[79] XIE, T., FOSTER, D. J., BAI, Y., JIANG, N. and KAKADE, S. M. (2022). The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157* .

[80] XIONG, W., ZHONG, H., SHI, C., SHEN, C. and ZHANG, T. (2022). A self-play posterior sampling algorithm for zero-sum Markov games. In *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*. PMLR.

[81] YANG, L. and WANG, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*. PMLR.

[82] YANG, Z., JIN, C., WANG, Z., WANG, M. and JORDAN, M. (2020). Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems* **33** 13903–13916.

[83] YU, T., QUILLEN, D., HE, Z., JULIAN, R., HAUSMAN, K., FINN, C. and LEVINE, S. (2020). Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*. PMLR.

[84] ZANETTE, A., BRANDFONBRENER, D., BRUNSKILL, E., PIROTTA, M. and LAZARIC, A. (2020). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

[85] ZANETTE, A., LAZARIC, A., KOCHENDERFER, M. and BRUNSKILL, E. (2020). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*. PMLR.

[86] ZHA, D., MA, W., YUAN, L., HU, X. and LIU, J. (2021). Rank the episodes: A simple approach for exploration in procedurally-generated environments. *arXiv preprint arXiv:2101.08152* .

[87] ZHANG, T. (2022). Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science* **4** 834–857.

[88] ZHANG, T. (2022). Mathematical analysis of machine learning algorithms .

[89] ZHONG, H., XIONG, W., ZHENG, S., WANG, L., WANG, Z., YANG, Z. and ZHANG, T. (2022). A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962* .

[90] ZHONG, H. and ZHANG, T. (2023). A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *arXiv preprint arXiv:2305.08841* .

[91] ZHOU, D., GU, Q. and SZEPESVARI, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

15

Contents

1 Introduction   **1**

    1.1 Related work   3

2 Preliminaries   **4**

    2.1 Episodic Markov Decision Processes and Online Reinforcement Learning   4

    2.2 Function Approximation: Model-Free and Model-Based Hypothesis   4

3 Algorithm: Maximize to Explore for Online RL   **5**

4 Regret Analysis for MEX Framework   **6**

5 Examples of MEX Framework   **7**

    5.1 Model-free online RL in Markov Decision Processes   7

    5.2 Model-based online RL in Markov Decision Processes   8

6 Experiments   **9**

7 Conclusions   **9**

A Limitations   **17**

B Additional Related Works   **17**

C Extensions to Two-player Zero-sum Markov Games   **18**

    C.1 Online Reinforcement Learning in Two-player Zero-sum Markov Games   18

    C.2 Function Approximation   20

    C.3 Algorithm Framework: Maximize to Explore (MEX-MG)   20

        C.3.1 Generic algorithm   20

        C.3.2 Model-free algorithm   21

        C.3.3 Model-based algorithm.   22

    C.4 Regret Analysis for MEX-MG Framework   22

    C.5 Examples of MEX-MG Framework   24

        C.5.1 Model-free online RL in Two-player Zero-sum Markov Games   24

        C.5.2 Model-based online RL in Two-player Zero-sum Markov Games   25

D Proof of Main Theoretical Results   **26**

    D.1 Proof of Theorem 4.4   26

    D.2 Proof of Theorem C.7   28

E Examples of Model-based and Model-free Online RL in MDPs   **29**

    E.1 Examples of Model-free Online RL in MDPs   30

    E.2 Examples of Model-based Online RL in MDPs   31

16

    E.3   Proof of Proposition 5.1 32

    E.4   Proof of Proposition 5.3 35

F   Proofs for Model-free and Model-based Online RL in Two-player Zero-sum MGs   36

    F.1   Proof of Proposition C.11 36

    F.2   Proof of Proposition C.16 39

    F.3   Proof of Proposition C.8 40

G Technical Lemmas   45

H Experimental Settings   46

    H.1   Environment Setup 46

    H.2   Implementation Details of `MEX-MF` 46

    H.3   Implementation Details of `MEX-MB` 46

    H.4   Tabular Experiments 46

A   Limitations

Though `MEX` framework is very flexible and can even be adapted for handling POMDP with low GEC [89], it remains unclear if we could adapt `MEX` to an offline setting where no well-explored dataset is provided. The adaptation is non-trivial as we would face an issue of minimax optimization to apply the pessimism principle as many existing works do [40, 78, 50, 28, 10].

B   Additional Related Works

Relationship with reward-biased maximum likelihood estimation. Our work is also related to a line of work in reward-biased maximum likelihood estimation. While [43] firstly proposed an estimation criterion that biases maximum likelihood estimation (RBMLE) with the cost or value, their algorithm is actually different from ours, by their Equation (6) and (8) in Section 3, their algorithm performs the estimation of model and policy optimization separately, for which they only obtained asymptotic convergence guarantees. Also, how well their decision rule explores remains unknown in theory. In contrast, MEX adopts a single optimization objective that combines estimation with policy optimization, which also ensures sample-efficient online exploration. [48, 33, 56, 55, 54] study RBMLE in Multi-arm bandit [48], Linear Stochastic Bandits [33], tabular RL [56], and Linear Quadratic Regulator settings (linear parameterized models of MDPs, [55, 54]) and also obtain the theoretical guarantees. While these settings are special cases for our proposed algorithms, our proven theoretical guarantee can also be generalized to these concrete cases. As we claim in this paper, our main contribution is to address the exploration-exploitation trade-off issue under general function approximation, which makes our work differ from these papers. [76] consider an algorithm similar to MEX, but our theory differs from theirs in both techniques and results. Our theory is based upon a unified framework of online RL with general function approximations, which covers their setup for the model-based hypothesis with kernel function approximation (RKHS). More importantly, they derived asymptotic regret of their algorithm based upon certain uniform boundedness and asymptotic normality assumptions, which are relatively strong conditions. In contrast, we derive finite sample regret upper bound for MEX, and the only fundamental assumption needed is a lower Generalized Eluder Coefficient (GEC) MDP, which contains almost all known theoretically tractable MDP classes (therefore covers their RKHS model). Finally, our paper further extends MEX to two-player zero-sum Markov games where similar algorithms and theories are previously unknown to the best of our knowledge. Moreover, the works mentioned above do not design experiments in deep RL environments, while we propose deep RL implementations and demonstrate their effectiveness in several MuJoco tasks.

17

Exploration in DRL. There has also been a long line of research that studies the exploration-exploitation trade-off from a practical perspective, where a prominent approach is referred to as the curiosity-driven method [60]. Curiosity-driven method focuses on the intrinsic rewards [60] (to handle the sparse extrinsic reward case) when making decisions, whose formulation can be largely grouped into either encouraging the algorithm to explore "novel" states [9, 51] or encouraging the algorithm to pick actions that reduce the uncertainty in its knowledge of the environment [31, 58, 64, 66]. These methods share the same theoretical motivation as the OFU principle in principle. In particular, one popular approach in this area is to use ensemble methods, which combine multiple neural networks of the value function and (or) policy ([75, 59, 14, 53, 44, 18, 45] and reference therein). For instance, Chen et al. [14] leverage the idea of upper confidence bound by estimating the uncertainty via the ensembles to improve the sample efficiency. However, the uncertainty estimation via ensembles is more computationally inefficient as compared to the vanilla algorithm. Meanwhile, these methods lack theoretical guarantees beyond the tabular and linear settings. It remains unknown in theory whether they are provably sample-efficient in the context of general function approximations. There is a rich body of literature, and we refer interested readers to Section 4 of Zha et al. [86] for a comprehensive review.

Two-player zero-sum Markov game. There have been numerous works on designing provably efficient algorithms for zero-sum Markov games (MGs). In the tabular case, Bai et al. [8], Bai and Jin [7], Liu et al. [47] propose algorithms with regret guarantees polynomial in the number of states and actions. Xie et al. [77], Chen et al. [16] then study the MGs in the linear function approximation case and design algorithms with $\widetilde{\mathcal{O}}(\mathrm{poly}(d, H)\sqrt{K})$ regret, where d is the dimension of the linear features. These approaches are later extended to the general function approximation by Jin et al. [38], Huang et al. [32], Xiong et al. [80], where the former two works studied OFU-based algorithms and the last one studied posterior sampling. However, these works focused only on the model-free case.

C Extensions to Two-player Zero-sum Markov Games

C.1 Online Reinforcement Learning in Two-player Zero-sum Markov Games

Markov games (MGs) generalize the standard Markov decision process to the multi-agent setting. We consider the episodic two-player zero-sum MG, which is denoted as $(H, \mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r)$. Here $\mathcal{S}$ is the state space shared by both players, $\mathcal{A}$ and $\mathcal{B}$ are the action spaces of the two players (referred to as the max-player and the min-player) respectively, $H \in \mathbb{N}_+$ denotes the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h \in [H]}$ with $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto \Delta(\mathcal{S})$ the transition kernel of the next state given the current state and two actions from the two players at timestep h, and $r = \{r_h\}_{h \in [H]}$ with $r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto [0, 1]$ the reward function at timestep h.

We consider *online* reinforcement learning in the episodic two-player zero-sum MG, where the two players interact with the MG for $K \in \mathbb{N}_+$ episodes through the following protocal. Each episode k starts from an initial state x_1^k. At each timestep h, two players observe the current state x_h^k, take joint actions (a_h^k, b_h^k) individually, and observe the next state $x_{h+1}^k \sim \mathbb{P}_h(\cdot \mid x_h^k, a_h^k, b_h^k)$. The k-th episode ends after step H and then a new episode starts. Without loss of generality, we assume each episode has a common fixed initial state $x_1^k = \underline{x}_1$, which can be easily generalized to having x_1 sampled from a fixed but unknown distribution.

Policies and value functions. We consider Markovian policies for both the max-player and the min-player. A Markovian policy of the max-player is denoted by $\mu = \{\mu_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h \in [H]}$. Similarly, a Markovian policy of the min-player is denoted by $\nu = \{\nu_h : \mathcal{X} \mapsto \Delta(\mathcal{B})\}_{h \in [H]}$. Given a joint policy $\boldsymbol{\pi} = (\mu, \nu)$, its state-value function $V_h^{\mu,\nu} : \mathcal{S} \mapsto \mathbb{R}_+$ and state-action value function $Q_h^{\mu,\nu} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}_+$ at timestep h are defined as

$$V_h^{\mu,\nu}(x) := \mathbb{E}_{\mathbb{P},(\mu,\nu)}\left[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}, b_{h'}) \,\middle|\, x_h = x\right], \tag{C.1}$$

$$Q_h^{\mu,\nu}(x, a, b) := \mathbb{E}_{\mathbb{P},(\mu,\nu)}\left[\sum_{h=h}^{H} r_{h'}(x_{h'}, a_{h'}, b_{h'}) \,\middle|\, (x_h, a_h, b_h) = (x, a, b)\right], \tag{C.2}$$

18

where the expectations are taken over the randomness of the transition kernel and the policies. In the zero-sum game, the max-player wants to maximize the value functions, while the min-layer aims at minimizing the value functions.

Best response, Nash equilibrium, and Bellman equations. Given a max-player's policy μ, the *best response policy* of the min-player, denoted by $\nu^\dagger(\mu)$, is the policy that minimizes the total rewards given that the max-player uses μ. According to this definition, and for notational simplicity, we denote

$$V_h^{\mu,\dagger}(x) := V_h^{\mu,\nu^\dagger(\mu)}(x) = \inf_\nu V_h^{\mu,\nu}(x),$$

$$Q_h^{\mu,\dagger}(x,a,b) := Q_h^{\mu,\nu^\dagger(\mu)}(x,a,b) = \inf_\nu Q_h^{\mu,\nu}(x,a,b), \tag{C.3}$$

for any $(x,a,b,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$. Similarly, given a min-player's policy ν, there is a *best response policy* $\mu^\dagger(\nu)$ for the max-player that maximizes the total rewards given ν. According to the definition, we denote

$$V_h^{\dagger,\nu}(x) := V_h^{\mu^\dagger(\nu),\nu}(x) = \sup_\mu V_h^{\mu,\nu}(x),$$

$$Q_h^{\dagger,\nu}(x,a,b) := Q_h^{\mu^\dagger(\nu),\nu}(x,a,b) = \sup_\mu Q_h^{\mu,\nu}(x,a,b), \tag{C.4}$$

for any $(x,a,b,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$. Furthermore, there exists a *Nash equilibrium* (NE) joint policy (μ^*, ν^*) [22] such that both players are optimal against their best responses. That is,

$$V_h^{\mu^*,\dagger}(x) = \sup_\mu V_h^{\mu,\dagger}(x), \quad V_h^{\dagger,\nu^*}(x) = \inf_\nu V_h^{\dagger,\nu}(x), \tag{C.5}$$

for any $(x,h) \in \mathcal{S} \times [H]$. For the NE joint policy, we have the following minimax equation,

$$\sup_\mu \inf_\nu V_h^{\mu,\nu}(x) = V_h^{\mu^*,\nu^*}(x) = \inf_\nu \sup_\mu V_h^{\mu,\nu}(x). \tag{C.6}$$

for any $(x,h) \in \mathcal{S} \times [H]$. This shows that: i) the for two-player zero-sum MG, the sup and the inf exchanges; ii) the NE policy has a unique state-value (state-action value) function, which we denote as V^* and Q^* respectively. Finally, we introduce two sets of Bellman equations for best response value functions and NE value functions. In specific, for the min-player's best response value functions given max-player policy μ, i.e., (C.3), we have the following Bellman equation,[3]

$$Q_h^{\mu,\dagger}(x,a,b) = (\mathcal{T}_h^\mu Q_{h+1}^{\mu,\dagger})(x,a,b)$$

$$:= r_h(x,a,b) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a,b)} \left[\inf_{\nu_{h+1}} \mathbb{D}_{(\mu_{h+1},\nu_{h+1})} Q_{h+1}^{\mu,\dagger}(x') \right], \tag{C.7}$$

for any $(x,a,b,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$. We name $\mathcal{T}_h^\mu$ as the *min-player best response Bellman operator* given max-player policy μ, and we define

$$\mathcal{E}_h^\mu(Q_h, Q_{h+1}; x,a,b) := Q_h(x,a,b) - \mathcal{T}_h^\mu Q_{h+1}(x,a,b), \tag{C.8}$$

as the *min-player best response Bellman residual* given max-player policy μ at timestep h of any functions (Q_h, Q_{h+1}). Also, for the NE value functions, i.e., (C.1), we also have the following NE Bellman equation,

$$Q_h^*(x,a,b) = (\mathcal{T}_h^{\text{NE}} Q_{h+1}^*)(x,a,b)$$

$$:= r_h(x,a,b) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a,b)} \left[\sup_{\mu_{h+1}} \inf_{\nu_{h+1}} \mathbb{D}_{(\mu_{h+1},\nu_{h+1})} Q_{h+1}^*(x') \right], \tag{C.9}$$

for any $(x,a,b,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$. We call $\mathcal{T}_h^{\text{NE}}$ the NE Bellman operator, and we define

$$\mathcal{E}_h^{\text{NE}}(Q_h, Q_{h+1}; x,a,b) := Q_h(x,a,b) - \mathcal{T}_h^{\text{NE}} Q_{h+1}(x,a,b), \tag{C.10}$$

as the *NE Bellman residual* at timestep h of any functions (Q_h, Q_{h+1}).

---

[3]For simplicity, we define $\mathbb{D}_{(\mu_h,\nu_h)} := \mathbb{E}_{a \sim \mu_h(\cdot|x), b \sim \nu_h(\cdot|x)}[Q(x,a,b)]$ for any μ_h, ν_h, and function Q.

19

Performance metric. We say a max-player's policy μ is ϵ-close to Nash equilibrium if $V_1^*(x_1) - V_1^{\mu,\dagger}(x_1) < \epsilon$. The goal of this section is to design a sample-efficient online learning algorithm to find such a max-player policy. The corresponding regret after K episodes is defined as,

$$\mathrm{Regret}_{\mathrm{MG}}(K) = \sum_{k=1}^K V_1^*(x_1) - V_1^{\mu^k,\dagger}(x_1), \qquad (C.11)$$

where μ^k is the policy used by the max-player for the k-th episode. Such a problem setting is also studied by Jin et al. [38], Huang et al. [32], Xiong et al. [80]. Actually, the roles of the two players can be exchanged, so that the goal turns to learning a min-player policy ν ϵ-close to the Nash equilibrium, to which the our algorithm can also apply.

C.2 Function Approximation

Function Approximation. To handle two-player zero-sum MGs with large or even infinite state spaces, we consider a function approximation approach. In specific, we have access to an abstract hypothesis class $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_H$, which can be specified to model-based and model-free settings respectively. Also, we denote by $\mathbf{\Pi} = \mathbf{M} \times \mathbf{N}$ with $\mathbf{M} = \mathbf{M}_1 \times \cdots \times \mathbf{M}_H$ and $\mathbf{N} = \mathbf{N}_1 \times \cdots \times \mathbf{N}_H$ as the space of Markovian joint policies. We specify this notation for model-free and model-based hypotheses respectively in the following.

Example C.1 (Model-free hypothesis). *For model-free hypothesis class, $\mathcal{H}$ contains state-actions value functions of the MG, i.e., $\mathcal{H}_h \subseteq \{f_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}\}$. Specifically, for any $f = (f_1, \cdots, f_H) \in \mathcal{H}$, we denote $Q_f = \{Q_{h,f}\}_{h \in [H]}$ with $Q_{h,f} = f_h$. Also:*

1. *Given $f \in \mathcal{H}$, we denote the corresponding NE state-value function by $V_f = \{V_{h,f}\}_{h \in [H]}$ with $V_{h,f}(\cdot) = \max_{\mu \in \mathbf{M}} \min_{\nu \in \mathbf{N}} \mathbb{D}_{(\mu_h, \nu_h)} Q_{h,f}(\cdot, \cdot, \cdot)$ and denote the corresponding NE max-player policy by $\mu_f = \{\mu_{h,f}\}_{h \in [H]}$ with $\mu_{h,f}(\cdot) = \arg\max_{\mu \in \mathbf{M}} \min_{\nu \in \mathbf{N}} \mathbb{D}_{(\mu_h, \nu_h)} Q_{h,f}(\cdot, \cdot, \cdot)$.*
2. *Given the policy of the max-player $\mu \in \mathbf{M}$, we define $V_f^{\mu,\dagger} = \{V_{h,f}^{\mu,\dagger}\}_{h \in [H]}$ as the state-value function induced by Q_f, μ, and its best response, i.e., $V_{h,f}^{\mu,\dagger} = \min_{\nu \in \mathbf{N}} \mathbb{D}_{(\mu_h, \nu_h)} Q_{h,f}(\cdot, \cdot, \cdot)$, and we denote the corresponding best response min-player policy as $\nu_{f,\mu} = \{\nu_{h,f,\mu}\}_{h \in [H]}$.*
3. *Finally, we denote the NE state-action value function under the true model Q^* by f^*.*

Example C.2 (Model-based hypothesis). *For model-based hypothesis class, $\mathcal{H}$ contains models of the MG, i.e., transition kernel $\mathbb{P}$. Specifically, we denote $f = \mathbb{P}_f = (\mathbb{P}_{1,f}, \cdots, \mathbb{P}_{H,f}) \in \mathcal{H}$. Also:*

1. *For any $(f, (\mu, \nu)) \in \mathcal{H} \times \mathbf{\Pi}$, we define $V_f^{\mu,\nu} = \{V_{h,f}^{\mu,\nu}\}_{h \in [H]}$ as the state-value function induced by model $\mathbb{P}_f$ and joint policy (μ, ν).*
2. *We define $V_f = \{V_{h,f}\}_{h = \in [H]}$ as the NE state-value function induced by model $\mathbb{P}_f$, and we denote the corresponding NE max-player policy as $\mu_f = \{\mu_{h,f}\}_{h \in [H]}$.*
3. *Given the policy of the max-player $\mu \in \mathbf{M}$, we define $V_f^{\mu,\dagger} = \{V_{h,f}^{\mu,\dagger}\}_{h \in [H]}$ as the state-value function induced by $\mathbb{P}_f$, μ, and its best response, and we denote the corresponding best response min-player policy as $\nu_{f,\mu} = \{\nu_{h,f,\mu}\}_{h \in [H]}$.*
4. *Finally, we denote the true model $\mathbb{P}$ of the MG as f^*.*

C.3 Algorithm Framework: Maximize to Explore (MEX-MG)

In this section, we extend the *Maximize to Explore* framework (MEX, Algorithm 1) to the two-player zero-sum MG setting, resulting in MEX-MG (Algorithm 2). MEX-MG controls the max-player and the min-player in a centralized manner. The min-player is aimed at assisting the max-player to achieve low regret. This kind of *self-play* algorithm framework has received considerable attention recently in theoretical study of two-player zero-sum MGs [38, 32, 80].

We first give a generic algorithm framework and then instantiate it to model-free (Example C.1) and model- based (Example C.2) hypotheses respectively.

C.3.1 Generic algorithm

MEX-MG leverages the asymmetric structure between the max-player and min-player to achieve sample-efficient learning. In specific, it picks two different hypotheses for the two players respectively, so

20

Algorithm 2 Maximize to Explore for two-player zero-sum Markov Game (`MEX-MG`)

1: **Input**: Hypothesis class $\mathcal{H}$, parameter $\eta > 0$.
2: **for** $k = 1, \cdots, K$ **do**
3:     Solve $f^k \in \mathcal{H}$ via

$$f^k = \operatorname*{argsup}_{f \in \mathcal{H}} \left\{ V_{1,f}(x_1) - \eta \cdot \sum_{h=1}^{H} L_h^{k-1}(f) \right\}. \tag{C.12}$$

4:     Set the max-player policy as $\mu^k = \mu_{f^k}$.
5:     Solve $g^k \in \mathcal{H}$ via

$$g^k = \operatorname*{argsup}_{g \in \mathcal{H}} \left\{ -V_{1,g}^{\mu^k,\dagger}(x_1) - \eta \cdot \sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(g) \right\}. \tag{C.13}$$

6:     Set the min-player policy as $\nu^k = \nu_{g^k,\mu^k}$.
7:     Execute $\boldsymbol{\pi}^k = (\mu^k, \nu^k)$ to collect data $\mathcal{D}^k = \{\mathcal{D}_h^k\}_{h \in [H]}$ with $\mathcal{D}_h^k = (x_h^k, a_h^k, b_h^k, r_h^k, x_{h+1}^k)$.
8: **end for**

---

that the max-player is aimed at approximating the NE max-player policy and the min-player is aimed at approximating the best response of the max-player, assisting its regret minimization.

Max-player. At each episode $k \in [K]$, `MEX-MG` first estimates a hypothesis $f^k \in \mathcal{H}$ for the max-player using historical data $\{\mathcal{D}^s\}_{s=1}^{k-1}$ by maximizing objective (C.12). Parallel to MEX, to achieve the goal of exploiting history knowledge while encouraging exploration, the composite objective (C.12) sums: **(a)** the negative loss $-L_h^{k-1}(f)$ induced by the hypothesis f; **(b)** the Nash equilibrium value associated with the current hypothesis, i.e., $V_{1,f}$. `MEX-MG` balances exploration and exploitation via a tuning parameter $\eta > 0$. With the hypothesis f^k, `MEX-MG` sets the max-player's policy μ^k as the Nash equilibrium max-player policy associated with f^k, i.e., μ_{f^k}.

Min-player. After obtaining the max-player policy μ^k, `MEX-MG` goes to estimate another hypothesis for the min-player in order to approximate the best response of the max-player. In specific, `MEX-MG` estimates $g^k \in \mathcal{H}$ using historical data $\{\mathcal{D}^s\}_{s=1}^{k-1}$ by maximizing objective (C.13), which also sums two objectives: **(a)** the negative loss $-L_{h,\mu^k}^{k-1}(g)$ induced by the hypothesis g. Here the loss function depends on μ^k since we aim to approximate the best response of μ^k; **(b)** the negative best response min-player value associated with the current hypothesis g and μ^k, i.e., $-V_{1,g}^{\mu^k,\dagger}$. The negative sign is due to the goal of min-player, i.e., minimization of the total rewards. With g^k, `MEX-MG` sets the min-player's policy ν^k as the best response policy of μ^k under g^k, i.e., ν_{g^k,μ^k}.

Data collection. Finally, the two agents execute the joint policy $\boldsymbol{\pi}^k = (\mu^k, \nu^k)$ to collect new data $\mathcal{D}^k = \{(x_h^k, a_h^k, b_h^k, r_h^k, x_{h+1}^k)\}_{h=1}^{H}$ and update their loss functions $L(\cdot)$. The choice of the loss functions varies between model-free and model-based hypotheses, which we specify in the following.

C.3.2 Model-free algorithm

For model-free hypothesis (Example C.1), the composite objectives (C.12) and (C.13) becomes

$$f^k = \operatorname*{argsup}_{f \in \mathcal{H}} \left\{ \sup_{\mu_1 \in \mathbf{M}_1} \inf_{\nu_1 \in \mathbf{N}_1} \mathbb{D}_{(\mu_1, \nu_1)} Q_{1,f}(x_1) - \eta \cdot \sum_{h=1}^{H} L_h^{k-1}(f) \right\}, \tag{C.14}$$

$$g^k = \operatorname*{argsup}_{g \in \mathcal{H}} \left\{ - \inf_{\nu_1 \in \mathbf{N}_1} \mathbb{D}_{(\mu_1^k, \nu_1)} Q_{1,g}(x_1) - \eta \cdot \sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(g) \right\}. \tag{C.15}$$

In the model-free algorithm, we choose the loss functions as empirical estimates of squared Bellman residuals. For the max-player who wants to approximate the NE max-player policy, we choose the

21

loss function $L_h^k(f)$ as an estimation of the squared NE Bellman residual, given by

$$L_h^k(f) = \sum_{s=1}^{k} \left(Q_{h,f}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s) \right)^2$$

$$- \inf_{f_h' \in \mathcal{H}_h} \sum_{s=1}^{k} \left(Q_{h,f'}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s) \right)^2. \qquad \text{(C.16)}$$

For the min-player who aims at approximating the best response policy of μ^k, we set the loss function $L_{h,\mu}^k(g)$ as an estimation of the squared best-response Bellman residual given max-player policy μ,

$$L_{h,\mu}^k(g) = \sum_{s=1}^{k} \left(Q_{h,g}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,g}^{\mu,\dagger}(x_{h+1}^s) \right)^2$$

$$- \inf_{g_h' \in \mathcal{H}_h} \sum_{s=1}^{k} \left(Q_{h,g'}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,g}^{\mu,\dagger}(x_{h+1}^s) \right)^2. \qquad \text{(C.17)}$$

We remark again that the subtracted infimum term in both (C.16) and (C.17) is for handling the variance terms in the estimation to achieve a fast theoretical rate, as we do for MEX with model-free hypothesis in Section 3.

C.3.3 Model-based algorithm.

For model-based hypothesis (Example C.2), the composite objectives (C.12) and (C.13) becomes

$$f^k = \underset{f \in \mathcal{H}}{\text{argsup}} \left\{ \sup_{\mu \in \mathbf{M}} \inf_{\nu \in \mathbf{N}} V_{1,\mathbb{P}_f}^{\mu,\nu}(x_1) - \eta \cdot \sum_{h=1}^{H} L_h^{k-1}(f) \right\}, \qquad \text{(C.18)}$$

$$g^k = \underset{g \in \mathcal{H}}{\text{argsup}} \left\{ - \inf_{\nu \in \mathbf{N}} V_{1,\mathbb{P}_g}^{\mu^k,\nu}(x_1) - \eta \cdot \sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(g) \right\}, \qquad \text{(C.19)}$$

which can be understood as a joint optimization over model $\mathbb{P}_f$ and the joint policy policy $\boldsymbol{\pi} = (\mu, \nu)$. In the model-based algorithm, we choose the loss function $L_h^k(f)$ as the negative log-likelihood loss,

$$L_h^k(f) = - \sum_{s=1}^{k} \log \mathbb{P}_{h,f}(x_{h+1}^s | x_h^s, a_h^s, b_h^s). \qquad \text{(C.20)}$$

Meanwhile, we choose the loss function $L_{h,\mu}^k(g) = L_h^k(g)$, i.e., (C.20), regardless of the max-player policy μ. But we remark that despite $L_h^k = L_{h,\mu}^k$, f^k and g^k are still different since the exploitation component in (C.18) and (C.19) are not the same due to the different targets of the two players.

C.4 Regret Analysis for MEX-MG Framework

In this section, we analyze the regret of the MEX-MG framework (Algorithm 2). Specifically, we give an upper bound of its regret which holds for the both model-free (Example C.1) and model-based (Example C.2) settings. We first present several key assumptions needed for the main result.

We first assume that the hypothesis class $\mathcal{H}$ is well-specified, containing certain true hypotheses.

Assumption C.3 (Realizablity). *We make the following realizability assumptions for the model-free and model-based hypotheses respectively:*

- *For model-free hypothesis (Example C.1), we assume that the true Nash equilibrium value $f^* \in \mathcal{H}$. Moreover, for any $f \in \mathcal{F}$, it holds that $Q^{\mu_f,\dagger} \in \mathcal{H}$.*
- *For model-based hypothesis (Example C.2), we assume that the true transition $f^* \in \mathcal{H}$.*

Also, we make the following completeness and boundedness assumption on $\mathcal{H}$.

Assumption C.4 (Completeness and Boundedness). *For model-free hypothesis (Example C.1), we assume that for any $f, g \in \mathcal{H}$, it holds that $\mathcal{T}_h^{\mu_f} g_h \in \mathcal{H}_h$, for any timestep $h \in [H]$. Also, we assume that there exists $B_f \geq 1$ such that for any $f_h \in \mathcal{H}_h$, it holds that $f_h(x, a, b) \in [0, B_f]$ for any $(x, a, b, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$.*

22

Assumptions C.3 and C.4 are standard assumptions in studying two-player zero-sum MGs [38, 32, 80]. Moreover, we make a structural assumption on the underlying MG to ensure sample-efficient online RL. Inspired by the single-agent setting, we require the MG to have low **Two-player Generalized Eluder Coefficient** (TGEC), which generalizes the GEC defined in Section 4. We provide specific examples of MGs of low TGEC, either model-free or model-based, in Section C.5.

To define TGEC, we introduce two discrepancy functions

$$\ell_{f'}(f; \xi_h) : \mathcal{H} \times \mathcal{H} \times (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}) \mapsto \mathbb{R}, \tag{C.21}$$

$$\ell_{f',\mu}(f; \xi_h) : \mathcal{H} \times \mathbf{N} \times \mathcal{H} \times (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}) \mapsto \mathbb{R}, \tag{C.22}$$

which characterizes the error incurred by a hypothesis $f \in \mathcal{H}$ on data $\xi_h = (x_h, a_h, r_h, x_{h+1})$. Intuitively, ℓ aims at characterizing the NE Bellman residual (C.10) while ℓ_μ aims at characterizing the best response Bellman residual given max-player μ (C.8). Specific choices of ℓ are given in Section C.5 for concrete model-free and model-based examples.

Assumption C.5 (Low Two-Player Generalized Eluder Coefficient (TGEC))**.** *We assume that given an $\epsilon > 0$, there exists a finite $d(\epsilon) \in \mathbb{R}_+$, such that for any sequence of hypotheses $\{(f^k, g^k)\}_{k \in [K]} \subset \mathcal{H}$ and policies $\{\boldsymbol{\pi}^k = (\mu_{f^k}, \nu_{g^k, \mu_{f^k}})\}_{k \in [K]} \subset \boldsymbol{\Pi}$, it holds that*

$$\sum_{k=1}^{K} V_{1,f^k} - V_1^{\boldsymbol{\pi}^k} \leq \inf_{\mu > 0} \left\{ \frac{\mu}{2} \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k} [\ell_{f^s}(f^k; \xi_h)] + \frac{d(\epsilon)}{2\mu} + \sqrt{d(\epsilon)HK} + \epsilon HK \right\},$$

and it also holds that

$$\sum_{k=1}^{K} V_1^{\boldsymbol{\pi}^k} - V_{1,g^k}^{\mu^k,\dagger} \leq \inf_{\mu > 0} \left\{ \frac{\mu}{2} \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k} [\ell_{g^s,\mu^k}(g^k; \xi_h)] + \frac{d(\epsilon)}{2\mu} + \sqrt{d(\epsilon)HK} + \epsilon HK \right\},$$

where $\mu_k = \mu_{f^k}$. We denote the smallest $d(\epsilon) \in \mathbb{R}_+$ satisfying this condition as $d_{\mathrm{TGEC}}(\epsilon)$.

Finally, we make a concentration-style assumption on loss functions, parallel to Assumption 4.3 for MDPs. For ease of presentation, we also assume that the hypothesis class $\mathcal{H}$ is finite.

Assumption C.6 (Generalization)**.** *We assume that $\mathcal{H}$ is finite, i.e., $|\mathcal{H}| < +\infty$, and that with probability at least $1 - \delta$, for any episode $k \in [K]$ and hypotheses $f, g \in \mathcal{H}$, it holds that*

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) \lesssim - \sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k} [\ell_{f^s}(f; \xi_h)] + B \cdot \big(H \log(HK/\delta) + \log(|\mathcal{H}|) \big).$$

and, with $\star = Q^{\mu^k,\dagger}$ for model-free hypothesis and $\star = f^*$ for model-based hypothesis, it holds that

$$\sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(\star) - L_{h,\mu^k}^{k-1}(g) \lesssim - \sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k} [\ell_{g^s,\mu^k}(g; \xi_h)] + B \cdot \big(H \log(HK/\delta) + \log(|\mathcal{H}|) \big),$$

Here $B = B_f^2$ for model-free hypothesis (see Assumption C.4) and $B = 1$ for model-based hypothesis.

As we show in Proposition C.8 and Proposition C.13, Assumption C.6 holds for both model-free and model-based settings. With Assumptions C.3, C.4 (model-free only), C.5, and C.6, we can present our main theoretical result.

Theorem C.7 (Online regret of `MEX-MG` (Algorithm 2))**.** *Under Assumptions C.3, C.4 (model-free only), C.5, and C.6, by setting*

$$\eta = \sqrt{\frac{d_{\mathrm{TGEC}}(1/\sqrt{HK})}{(H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B \cdot K}},$$

the regret of Algorithm 2 after K episodes is upper bounded by

$$\mathrm{Regret}(K) \lesssim \sqrt{d_{\mathrm{TGEC}}(1/\sqrt{HK}) \cdot (H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B \cdot K},$$

with probability at least $1 - \delta$. Here $d_{\mathrm{TGEC}}(\cdot)$ is given by Assumption C.5.

Proof of Theorem C.7. See Appendix D.2 for detailed proof. □

23

C.5 Examples of MEX-MG Framework

C.5.1 Model-free online RL in Two-player Zero-sum Markov Games

In this subsection, we specify MEX-MG (Algorithm 2) for model-free hypothesis class (Example C.1). In specific, we choose the discrepancy functions ℓ and ℓ_μ as, given $\xi_h = (x_h, a_h, b_h, r_h, x_{h+1})$,

$$\ell_{f'}(f; \xi_h) = \Big(Q_{h,f}(x_h, a_h, b_h) - r_h - \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot \,|\, x_h, a_h, b_h)}[V_{h+1,f}(x_{h+1})] \Big)^2, \qquad \text{(C.23)}$$

$$\ell_{f',\mu}(g; \xi_h) = \Big(Q_{h,g}(x_h, a_h, b_h) - r_h - \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot \,|\, x_h, a_h, b_h)}[V_{h+1,g}^{\mu,\dagger}(x_{h+1})] \Big)^2. \qquad \text{(C.24)}$$

By (C.23) and (C.24), both $\ell_{f'}$ and $\ell_{f',\mu}$ do not depend on the input f'. In the following, we check and specify Assumptions C.5 and C.6 in Section C.4 for model-free hypothesis class.

Proposition C.8 (Generalization: model-free RL)**.** *We assume that $\mathcal{H}$ is finite, i.e., $|\mathcal{H}| < +\infty$. Then with probability at least $1 - \delta$, for any $k \in [K]$ and $f, g \in \mathcal{H}$, it holds simultaneously that*

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) \lesssim -\sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}[\ell_{f^s}(f; \xi_h)] + B_f^2 \big(H \log(HK/\delta) + \log(|\mathcal{H}|) \big),$$

$$\sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(Q^{\mu^k,\dagger}) - L_{h,\mu^k}^{k-1}(g) \lesssim -\sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}[\ell_{g^s,\mu^k}(g; \xi_h)] + B_f^2 \big(H \log(HK/\delta) + \log(|\mathcal{H}|) \big),$$

where L, L_μ, ℓ, and ℓ_μ are defined in (C.15), (C.16), (C.23), and (C.24), respectively.

Proof of Proposition C.8. See Appendix F.3 for a detailed proof. □

Proposition C.8 specifies Assumption C.6 for abstract model-free hypothesis. Now given a two-player zero-sum MG with TGEC d_{TGEC}, we have the following corollary of Theorem C.7.

Corollary C.9 (Online regret of MEX-MG: model-free hypothesis)**.** *Given a two-player zero-sum MG with two-player generalized eluder coefficient $d_{\mathrm{TGEC}}(\cdot)$ and a finite model-free hypothesis class $\mathcal{H}$ satisfying Assumptions C.3 and C.4, by setting*

$$\eta = \sqrt{\frac{d_{\mathrm{TGEC}}(1/\sqrt{HK})}{(H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B_f^2 \cdot K}}, \qquad \text{(C.25)}$$

then the regret of Algorithm 2 after K episodes is upper bounded by

$$\mathrm{Regret}(T) \lesssim B_f \cdot \sqrt{d_{\mathrm{TGEC}}(1/\sqrt{HK}) \cdot (H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot K}, \qquad \text{(C.26)}$$

with probability at least $1 - \delta$. Here B is specified in Assumption C.4.

Linear two-player zero-sum Markov game. Next, we introduce the linear two-player zero-sum MG [77] as a concrete model-free example, for which we can explicitly specify its TGEC. Linear two-player zero-sum MG is a natural extension of linear MDPs [39] to the two-player zero-sum MG setting, whose reward and transition kernels are modeled by linear functions.

Definition C.10 (Linear two-player zero-sum Markov game)**.** *A d-dimensional two-player zero-sum linear Markov game satisfies that*

$$r_h(x, a, b) = \phi_h(x, a, b)^\top \alpha_h, \quad \mathbb{P}_h(x' \,|\, x, a, b) = \phi_h(x, a, b)^\top \psi_h^\star(x'), \qquad \text{(C.27)}$$

for some known feature mapping $\phi_h(x, a, b) \in \mathbb{R}^d$ and some unknown vector $\alpha_h \in \mathbb{R}^d$ and unknown function $\psi_h(x') \in \mathbb{R}^d$ satisfying $\|\phi_h(x, a, b)\|_2 \leq 1$ and $\max\{\|\alpha_h\|_2, \|\psi_h^\star(x')\|_2\} \leq \sqrt{d}$ for any $(x, a, b, x', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathcal{S} \times [H]$.

For a linear two-player zero-sum MG, we choose the model-free hypothesis class as, for each $h \in [H]$,

$$\mathcal{H}_h = \Big\{ \phi_h(\cdot, \cdot, \cdot)^\top \theta_h : \|\theta_h\|_2 \leq (H + 1 - h)\sqrt{d} \Big\}. \qquad \text{(C.28)}$$

The following proposition gives the TGEC of a linear two-player zero-sum MG with hypothesis class (C.28).

24

Proposition C.11 (TGEC of linear two-player zero-sum MG). *For a linear two-player zero-sum MG, with model-free hypothesis* (C.28)*, it holds that*

$$d_{\mathrm{TGEC}}(1/\sqrt{HK}) \lesssim d\log(HK), \quad \log\left(\mathcal{N}(\mathcal{H}, 1/K, \|\cdot\|_\infty)\right) \lesssim dH\log(dK), \qquad \text{(C.29)}$$

where $\mathcal{N}(\mathcal{H}, 1/K, \|\cdot\|_\infty)$ *denotes the* $1/K$*-covering number of* $\mathcal{H}$ *under* $\|\cdot\|_\infty$*-norm.*

Proof of Proposition C.11. See Appendix F.1 for a detailed proof. □

A linear two-player zero-sum MG with the model-free hypothesis class (C.28) also satisfies the realizability and completeness assumption (Assumptions C.3 and C.4), which are proved by Huang et al. [32]. Then we can specify Theorem C.7 for linear two-player zero-sum MGs as follows.

Corollary C.12 (Online regret of MEX-MG: linear two-player zero-sum MG). *With* $\eta = \widetilde{\Theta}(\sqrt{1/H^3K})$*, the regret of Algorithm 2 for linear two-player zero-sum MG after* K *episodes is upper bounded by*

$$\mathrm{Regret}_{\mathrm{MG}}(K) \lesssim dH^{3/2}K^{1/2}\log(HKd/\delta),$$

with probability at least $1 - \delta$*, where* d *is the dimension of the linear two-player zero-sum MG.*

Proof of Corollary C.12. This is a corollary of Theorem C.7, Propositions C.8, C.11, together with a covering number argument. □

C.5.2 Model-based online RL in Two-player Zero-sum Markov Games

In this subsection, we specify Algorithm 2 for model-based hypothesis class $\mathcal{H}$ (Example C.2). In specific, we choose the discrepancy function ℓ as the Hellinger distance. Given data $\xi_h = (x_h, a_h, b_h, x_{h+1})$, we let

$$\ell_{f'}(f; \xi_h) = \ell_{f',\mu}(f; \xi_h) = D_{\mathrm{H}}(\mathbb{P}_{h,f}(\cdot|x_h, a_h, b_h)\|\mathbb{P}_{h,f^*}(\cdot|x_h, a_h, b_h)), \qquad \text{(C.30)}$$

where $D_{\mathrm{H}}(\cdot\|\cdot)$ denotes the Hellinger distance. We note that due to (C.30), the discrepancy function ℓ does not depend on the input $f' \in \mathcal{H}$ and the max-player policy μ. In the following, we check and specify Assumptions C.5 and C.6 in Section C.4 for model-based hypothesis classes.

Proposition C.13 (Generalization: model-based RL). *We assume that* $\mathcal{H}$ *is finite, i.e.,* $|\mathcal{H}| < +\infty$*. Then with probability at least* $1 - \delta$*, for any* $k \in [K]$*,* $f \in \mathcal{H}$*, it holds that*

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) \lesssim -\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h \sim \pi^k}[\ell_{f^s}(f; \xi_h)] + H\log(H/\delta) + \log(|\mathcal{H}|),$$

where L *and* ℓ *are defined in* (C.20) *and* (C.30) *respectively.*

Proof of Proposition C.13. This proposition follows from the same proof of Proposition 5.3. □

Since $L_h^k = L_{h,\mu}^k$ and $\ell_f = \ell_{f,\mu}$, Proposition C.13 means that Assumption C.6 holds. Now given a two-player zero-sum MG with TGEC d_{TGEC}, we have the following corollary of Theorem C.7.

Corollary C.14 (Online regret of MEX-MG: model-based hypothesis). *Given a two-player zero-sum MG with two-player generalized eluder coefficient* $d_{\mathrm{TGEC}}(\cdot)$ *and a finite model-based hypothesis class* $\mathcal{H}$ *with* $f^* \in \mathcal{H}$*, by setting*

$$\eta = \sqrt{\frac{d_{\mathrm{TGEC}}(1/\sqrt{HK})}{(H\log(HK/\delta) + \log(|\mathcal{H}|)) \cdot K}}, \qquad \text{(C.31)}$$

then the regret of Algorithm 2 after K *episodes is upper bounded by*

$$\mathrm{Regret}(T) \lesssim \sqrt{d_{\mathrm{TGEC}}(1/\sqrt{HK}) \cdot (H\log(HK/\delta) + \log(|\mathcal{H}|)) \cdot K}, \qquad \text{(C.32)}$$

with probability at least $1 - \delta$*.*

25

Linear mixture two-player zero-sum Markov game. In the following, we introduce the linear mixture two-player zero-sum MG as a concrete model-based example, for which we can explicitly specify its TGEC. Linear mixture MG is a natural extension of linear mixture MDPs [5, 57, 12] to the two-player zero-sum MG setting, whose transition kernels are modeled by linear kernels. But just as the single-agent setting, the linear mixture MG and the linear MG (Definition C.10) do not cover each other as special cases [12].

Definition C.15 (Linear mixture two-player zero-sum Markov game). *A d-dimensional two-player zero-sum linear mixture Markov game satisfies that*

$$\mathbb{P}_h(x' \mid x, a, b) = \phi_h(x, a, b, x')^\top \theta_h^\star \tag{C.33}$$

for some known feature mapping $\phi_h(x, a, b, x') \in \mathbb{R}^d$ and some unknown vector $\theta_h^\star \in \mathbb{R}^d$ satisfying $\|\phi_h(x, a, b, x')\|_2 \le 1$ and $\|\theta_h\|_2 \le \sqrt{d}$ for any $(x, a, b, x', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathcal{S} \times [H]$.

Linear mixture two-player zero-sum MG also covers the tabular two-player zero-sum MG as a special case. For a linear mixture two-player zero-sum MG, we choose the model-based hypothesis class as, for each h,

$$\mathcal{H}_h = \left\{ \phi_h(\cdot, \cdot, \cdot, \cdot)^\top \theta_h : \|\theta_h\|_2 \le \sqrt{d} \right\}. \tag{C.34}$$

The following proposition gives the TGEC of a linear mixture two-player zero-sum MG.

Proposition C.16 (TGEC of linear mixture two-player zero-sum MG). *For a linear mixture two-player zero-sum MG, with model-free hypothesis (C.28), it holds that*

$$d_{\mathrm{TGEC}}(1/\sqrt{HK}) \lesssim dH^2 \log(HK), \quad \log\left(\mathcal{N}(\mathcal{H}, 1/K, \|\cdot\|_\infty)\right) \lesssim dH \log(dK). \tag{C.35}$$

where $\mathcal{N}(\mathcal{H}, 1/K, \|\cdot\|_\infty)$ denotes the $1/K$-covering number of $\mathcal{H}$ under $\|\cdot\|_\infty$-norm.

Proof of Proposition C.16. See Appendix F.2 for a detailed proof. $\qquad\square$

Then we can specify Theorem C.7 for linear mixture two-player zero-sum MGs as follows.

Corollary C.17 (Online regret of MEX-MG: linear mixture two-player zero-sum MG). *By setting $\eta = \widetilde{\Theta}(\sqrt{H/K})$, the regret of Algorithm 2 for linear mixture two-player zero-sum MG after K episodes is upper bounded by*

$$\mathrm{Regret}_{\mathrm{MG}}(K) \lesssim dH^{3/2} K^{1/2} \log(HKd/\delta),$$

with probability at least $1 - \delta$, where d is the dimension of the linear mixture two-player zero-sum MG.

Proof of Corollary C.17. This is a corollary of Theorem C.7, Propositions C.13, C.16, together with a covering number argument. $\qquad\square$

D Proof of Main Theoretical Results

D.1 Proof of Theorem 4.4

Proof of Theorem 4.4. Consider the following decomposition of the regret,

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} V_1^*(x_1) - V_1^{\pi_{f^k}}(x_1)$$

$$= \underbrace{\sum_{k=1}^{K} V_1^*(x_1) - V_{1,f^k}(x_1)}_{\text{Term (i)}} + \underbrace{\sum_{k=1}^{K} V_{1,f^k}(x_1) - V_1^{\pi_{f^k}}(x_1)}_{\text{Term (ii)}} \tag{D.1}$$

26

Term (i). Note that by our definition in both Example 2.2 and 2.1, we have that $V_1^* = V_{1,f^*}$. Thus we can rewrite the term (i) as

$$\text{Term (i)} = \sum_{k=1}^{K} V_{1,f^*}(x_1) - V_{1,f^k}(x_1). \tag{D.2}$$

Then by our choice of f^k in (3.1) and the fact that $f^* \in \mathcal{H}$, we have that for each $k \in [K]$,

$$V_{1,f^*}(x_1) - \eta \sum_{h=1}^{H} L_h^{k-1}(f^*)(x_1) \leq V_{1,f^k}(x_1) - \eta \sum_{h=1}^{H} L_h^{k-1}(f^k)(x_1) \tag{D.3}$$

By combining (D.2) and (D.3), we can derive that

$$\text{Term (i)} \leq \eta \sum_{k=1}^{K} \sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f^k) \tag{D.4}$$

Now by applying Assumption 4.3 to (D.4), we can further derive that with probability at least $1 - \delta$,

$$\text{Term (i)} \leq -c_{(\text{i})} \cdot \eta \sum_{k=1}^{K} \sum_{s=1}^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f^k; \xi_h)] + c_{(\text{i})} \cdot \eta BK \big(H \log(HK/\delta) + \log(|\mathcal{H}|)\big). \tag{D.5}$$

where $c_{(\text{i})} > 0$ is some absolute constant (recall the definition of $\lesssim$).

Term (ii). For term (ii) of (D.2), we apply Assumption 4.2 and obtain that, for any $\epsilon > 0$,

$$\text{Term (ii)} \leq \inf_{\mu > 0} \left\{ \frac{\mu}{2} \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f^k; \xi_h)] + \frac{d_{\text{GEC}}(\epsilon)}{2\mu} + \sqrt{d_{\text{GEC}}(\epsilon)HK} + \epsilon HK \right\}.$$

By taking $\mu/2 = c_{(\text{i})} \cdot \eta$, we can further derive that,

$$\text{Term (ii)} \leq c_{(\text{i})} \cdot \eta \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f^k; \xi_h)] + \frac{d_{\text{GEC}}(\epsilon)}{4c_{(\text{i})}\eta} + \sqrt{d_{\text{GEC}}(\epsilon)HK} + \epsilon HK. \tag{D.6}$$

Combining Term (i) and Term (ii). Now by combining (D.5) and (D.6), we can obtain that with probability at least $1 - \delta$,

$$\text{Regret}(T) = \text{Term (i)} + \text{Term (ii)}$$

$$\leq -c_{(\text{i})} \cdot \eta \sum_{k=1}^{K} \sum_{s=1}^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f^k; \xi_h)] + c_{(\text{i})} \cdot \eta BK \big(H \log(HK/\delta) + \log(|\mathcal{H}|)\big),$$

$$+ c_{(\text{i})} \cdot \eta \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f^k; \xi_h)] + \frac{d_{\text{GEC}}(\epsilon)}{4c_{(\text{i})}\eta} + \sqrt{d_{\text{GEC}}(\epsilon)HK} + \epsilon HK$$

$$= c_{(\text{i})} \cdot \eta BK \big(H \log(HK/\delta) + \log(|\mathcal{H}|)\big) + \frac{d_{\text{GEC}}(\epsilon)}{4c_{(\text{i})}\eta} + \sqrt{d_{\text{GEC}}(\epsilon)HK} + \epsilon HK. \tag{D.7}$$

By taking $\epsilon = 1/\sqrt{HK}$, $\eta = \sqrt{d_{\text{GEC}}(\epsilon)/(H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B \cdot K)}$, we can derive from (D.7) that, with probability at least $1 - \delta$, it holds that

$$\text{Regret}(K) \lesssim \sqrt{d_{\text{GEC}}(1/\sqrt{HK}) \cdot (H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B \cdot K}. \tag{D.8}$$

This finishes the proof of Theorem 4.4. $\qquad\square$

27

D.2 Proof of Theorem C.7

Proof of Theorem C.7. Consider the following decomposition of the regret,

$$
\begin{aligned}
\mathrm{Regret}(K) &= \sum_{k=1}^{K} V_1^*(x_1) - V_1^{\mu^k,\dagger}(x_1) \\
&= \sum_{k=1}^{K} V_1^*(x_1) - V_1^{\mu^k,\nu^k}(x_1) + \sum_{k=1}^{K} V_1^{\mu^k,\nu^k}(x_1) - V_1^{\mu^k,\dagger}(x_1) \\
&= \underbrace{\sum_{k=1}^{K} V_1^\star - V_{1,f^k}}_{\text{Term (Max.i)}} + \underbrace{\sum_{k=1}^{K} V_{1,f^k} - V_1^{\mu^k,\nu^k}}_{\text{Term (Max.ii)}} + \underbrace{\sum_{k=1}^{K} V_{1,g^k}^{\mu^k,\dagger} - V_1^{\mu^k,\dagger}}_{\text{Term (Min.i)}} + \underbrace{\sum_{k=1}^{K} V_1^{\mu^k,\nu^k} - V_{1,g^k}^{\mu^k,\dagger}}_{\text{Term (Min.ii)}},
\end{aligned}
$$

$$\tag{D.9}$$

where in the last equality we omit the dependence on x_1 for simplicity.

Term (Max.i). Note that by our definition in both Example C.1 and Example C.2, we have that $V_1^* = V_{1,f^*}$. Thus we can rewrite the term (Max.i) as

$$
\text{Term (Max.i)} = \sum_{k=1}^{K} V_{1,f^*}(x_1) - V_{1,f^k}(x_1). \tag{D.10}
$$

Then by our choice of f^k in (C.12) and the fact that $f^* \in \mathcal{H}$, we have that for each $k \in [K]$,

$$
V_{1,f^*}(x_1) - \eta \sum_{h=1}^{H} L_h^{k-1}(f^*) \leq V_{1,f^k}(x_1) - \eta \sum_{h=1}^{H} L_h^{k-1}(f^k). \tag{D.11}
$$

By combining (D.10) and (D.11), we can derive that

$$
\text{Term (Max.i)} \leq \eta \sum_{k=1}^{K} \sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f^k). \tag{D.12}
$$

Now applying Assumption C.6 to (D.12), we can further derive that with probability at least $1 - \delta$,

Term (Max.i) $\tag{D.13}$

$$
\leq -c_{(\mathrm{max.i})} \cdot \eta \sum_{k=1}^{K} \sum_{s=1}^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\xi_h \sim \pi^k}[\ell_{f^s}(f;\xi_h)] + c_{(\mathrm{max.i})} \cdot BK\big(H \log(HK/\delta) + \log(|\mathcal{H}|)\big),
$$

$$\tag{D.14}$$

for some absolute constant $c_{(\mathrm{max.i})} > 0$.

Term (Max.ii). For term (Max.ii), we apply Assumption C.5 and obtain that, for any $\epsilon > 0$,

$$
\text{Term (Max.ii)} \leq \inf_{\zeta > 0} \left\{ \frac{\zeta}{2} \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^k}[\ell_{f^s}(f^k;\xi_h)] + \frac{d_{\mathrm{TGEC}}(\epsilon)}{2\zeta} + \sqrt{d_{\mathrm{TGEC}}(\epsilon)HK} + \epsilon HK \right\}.
$$

By taking $\zeta/2 = c_{(\mathrm{max.i})} \cdot \eta$, we can further derive that

$$
\text{Term (Max.ii)} \leq c_{(\mathrm{max.i})} \cdot \eta \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^k}[\ell_{f^s}(f^k;\xi_h)] + \frac{d_{\mathrm{TGEC}}(\epsilon)}{4c_{(\mathrm{max.i})}\eta} + \sqrt{d_{\mathrm{TGEC}}(\epsilon)HK} + \epsilon HK.
$$

$$\tag{D.15}$$

28

Term (Min.i). For either model-free or model-based hypothsis, by our definition in Example C.1 and Example C.2 respectively, we both have that $V_1^{\mu^k,\dagger} = V_{1,\star}^{\mu^k,\dagger}$. Here $\star = Q^{\mu^k,\dagger}$ for model-free hypothesis and $\star = f^*$ for model-based hypothesis. Thus we can rewrite the term (Min.i) as[4].

$$\text{Term (Min.i)} = \sum_{k=1}^{K} V_{1,g^k}^{\mu^k,\dagger}(x_1) - V_{1,\star}^{\mu^k,\dagger}(x_1). \tag{D.16}$$

Then by our choice of g^k in (C.13) and the fact that $\star \in \mathcal{H}$ (Assumption C.3), we have that for each $k \in [K]$,

$$-V_{1,\star}^{\mu^k,\dagger}(x_1) - \eta \sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(\star) \leq -V_{1,g^k}^{\mu^k,\dagger}(x_1) - \eta \sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(g^k) \tag{D.17}$$

By combining (D.16) and (D.17), we can derive that

$$\text{Term (Min.i)} \leq \eta \sum_{k=1}^{K} \sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(\star) - L_{h,\mu^k}^{k-1}(g^k) \tag{D.18}$$

Now applying Assumption C.6 to (D.18), we can further derive that with probability at least $1 - \delta$,

$$\text{Term (Min.i)} \leq -c_{(\text{min.i})} \cdot \eta \sum_{k=1}^{K} \sum_{s=1}^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}[\ell_{g^s,\mu^k}(g;\xi_h)] + c_{(\text{min.i})} \cdot \eta B K \big(H \log(HK/\delta) + \log(|\mathcal{H}|) \big). \tag{D.19}$$

Term (Min.ii). For term (Min.ii), we apply Assumption C.5 and obtain that, for any $\epsilon > 0$,

$$\text{Term (Min.ii)} \leq \inf_{\zeta > 0} \left\{ \frac{\zeta}{2} \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}[\ell_{g^s,\mu^k}(g^k;\xi_h)] + \frac{d_{\text{TGEC}}(\epsilon)}{2\zeta} + \sqrt{d_{\text{TGEC}}(\epsilon)HK} + \epsilon HK \right\}.$$

By taking $\zeta/2 = c_{(\text{min.i})} \cdot \eta$, we can further derive that

$$\text{Term (Max.ii)} \leq c_{(\text{min.i})} \cdot \eta \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}[\ell_{g^s,\mu^k}(g^k;\xi_h)] + \frac{d_{\text{TGEC}}(\epsilon)}{4c_{(\text{min.i})}\eta} + \sqrt{d_{\text{TGEC}}(\epsilon)HK} + \epsilon HK. \tag{D.20}$$

Combining Term (Max.i), Term (Max.ii), Term (Min.i), and Term (Min.ii). Now combining (D.13), (D.15), (D.19), and (D.15), taking $\epsilon = 1/\sqrt{HK}$ and

$$\eta = \sqrt{\frac{d_{\text{TGEC}}(1/\sqrt{HK})}{(H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B \cdot K}}, \tag{D.21}$$

we can finally derive that with probability at least $1 - 2\delta$,

$$\text{Regret}(K) \lesssim \sqrt{d_{\text{TGEC}}(1/\sqrt{HK}) \cdot (H \log(HK/\delta) + \log(|\mathcal{H}|)) \cdot B \cdot K}.$$

This finishes the proof of Theorem C.7. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

E Examples of Model-based and Model-free Online RL in MDPs

In this section, we specify Corollaries 5.2 and 5.4 to various examples of MDPs with low generalized eluder coefficient (GEC [89]). Sections E.1 and E.2 consider model-free hypothesis and model-based hypothesis, respectively. After, we give proof of the generalization guarantees involved in Section 5. Section E.3 provides proof of Proposition 5.1 and Section E.4 provides proof of Proposition 5.3.

---

[4]We remark that this notation is well-defined, since $Q^{\mu_f,\dagger} \in \mathcal{H}$ for any $f \in \mathcal{H}$ by Assumption 4.1.

29

E.1 Examples of Model-free Online RL in MDPs

MDPs with low Bellman eluder dimension. In this part, we study MDPs with low Bellman eluder (BE) dimension [37]. To introduce, we define the notion of ϵ-independence between distributions and the notion of distributional eluder dimension.

Definition E.1 (ϵ-independence between distributions)**.** *Let $\mathcal{G}$ be a function class on the space $\mathcal{X}$, and let $\nu, \mu_1, \cdots, \mu_n$ be probability measures on $\mathcal{X}$. We say ν is ϵ-independent of $\{\mu_1, \cdots, \mu_n\}$ with respect to $\mathcal{G}$ if there exists a $g \in \mathcal{G}$ such that $\sqrt{\sum_{i=1}^{n} (\mathbb{E}_{\mu_i}[g])^2} \leq \epsilon$ but $|\mathbb{E}_{\nu}[g]| > \epsilon$.*

Definition E.2 (Distributional Eluder (DE) dimension)**.** *Let $\mathcal{G}$ be a function class on space $\mathcal{X}$, and let Π be a family of probability measures on $\mathcal{X}$. The distributional eluder dimension $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ is defined as the length of the longest sequence $\{\rho_1, \cdots, \rho_n\} \subset \Pi$ such that there exists $\epsilon' \geq \epsilon$ with ρ_i being ϵ'-independent of $\{\rho_1, \cdots, \rho_{i-1}\}$ for each $i \in [n]$.*

The Bellman eluder dimension is based upon the notion of distributional eluder dimsntion. For a model-free hypothesis class $\mathcal{H}$, we the Bellman operator $\mathcal{T}_h$ defined in Section 2 becomes,

$$(\mathcal{T}_h f_{h+1})(x, a) = R_h(x, a) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)}[V_{h+1,f}(x')], \tag{E.1}$$

for any $f \in \mathcal{H}$. Then we define the Q-type/V-type Bellman eluder dimension as the following.

Definition E.3 (Q-type Bellman eluder (BE) dimension [37, 89])**.** *We define $(I - \mathcal{T}_h)\mathcal{H} = \{(x, a) \mapsto (f_h - \mathcal{T}_h f_{h+1})(x, a) : f \in \mathcal{H}\}$ as the set of Bellman residuals induced by $\mathcal{H}$ at step h, and let $\Pi = \{\Pi_h\}_{h=1}^{H}$ be a collection of H families of probability measure over $\mathcal{S} \times \mathcal{A}$. The Q-type ϵ-Bellman eluder dimension of $\mathcal{H}$ with respect to Π is defined as*

$$\dim_{\mathrm{BE}}(\mathcal{H}, \Pi, \epsilon) = \max_{h \in [H]} \{\dim_{\mathrm{DE}}((I - \mathcal{T}_h)\mathcal{H}, \Pi_h, \epsilon)\}.$$

Definition E.4 (V-type Bellman eluder (BE) dimension [37, 89])**.** *We define $(I - \mathcal{T}_h)V_{\mathcal{H}} = \{x \mapsto (f_h - \mathcal{T}_h f_{h+1})(x, \pi_{h,f}(x)) : f \in \mathcal{H}\}$ as the set of V-type Bellman residuals induced by $\mathcal{H}$ at step h, and let $\Pi = \{\Pi_h\}_{h=1}^{H}$ be a collection of H families of probability measure over $\mathcal{S}$. The V-type ϵ-Bellman eluder dimension of $\mathcal{H}$ with respect to Π is defined as*

$$\dim_{\mathrm{VBE}}(\mathcal{H}, \Pi, \epsilon) = \max_{h \in [H]} \{\dim_{\mathrm{DE}}((I - \mathcal{T}_h)V_{\mathcal{H}}, \Pi_h, \epsilon)\}.$$

For MDPs with low Bellman eluder dimension, we choose the function l in Assumption 3.1 as

$$l_{f'}((f_h, f_{h+1}); \mathcal{D}_h) = Q_{h,f}(x_h, a_h) - r_h - V_{h+1,f}(x_{h+1}). \tag{E.2}$$

and we choose the operator $\mathcal{P}_h = \mathcal{T}_h$ defined in (E.1). One can check that such a choice satisfies Assumption 3.1. By further choosing the exploration policy as $\pi_{\exp}(f) = \pi_f$ for Q-type problems and $\pi_{\exp}(f) = \pi_f \circ_h \mathrm{Unif}(\mathcal{A})$ for V-type problems[5], we can bound the GEC for MDPs with low BE dimension by the following lemma.

Lemma E.5 (GEC for low Bellman eluder dimension, Lemma 3.16 in [89])**.** *Let the discrepancy ℓ function be chosen as (5.1) with l defined in (E.2). Define $\Pi_{\mathcal{H}}$ as the distributions induced by following some $f \in \mathcal{H}$ greedily. For Q-type problems, by choosing $\pi_{\exp}(f) = \pi_f$, we have that*

$$d_{\mathrm{GEC}}(\epsilon) \leq 2 \dim_{\mathrm{BE}}(\mathcal{H}, \Pi_{\mathcal{H}}, \epsilon) H \cdot \log(K),$$

For V-type problems, by choosing $\pi_{\exp}(f) = \pi_f \circ_h \mathrm{Unif}(\mathcal{A})$, we have that

$$d_{\mathrm{GEC}}(\epsilon) \leq 2 \dim_{\mathrm{VBE}}(\mathcal{H}, \Pi_{\mathcal{H}}, \epsilon) |\mathcal{A}| H \cdot \log(K).$$

Proof of Lemma E.5. See Lemma 3.16 in [89] for a detailed proof. $\qquad\square$

By combining Lemma E.5 and Corollary 5.2, we can obtain that for Q-type low Bellman eluder dimension problem, it holds that with probability at least $1 - \delta$,

$$\mathrm{Regret}(T) \lesssim B_l^2 \cdot \sqrt{\dim_{\mathrm{BE}}(\mathcal{H}, \Pi_{\mathcal{H}}, 1/\sqrt{HK}) \cdot \log(HK|\mathcal{H}|/\delta) \cdot H^2 K}, \tag{E.3}$$

and for V-type Bellman eluder dimension problem, it holds that with probability at least $1 - \delta$,

$$\mathrm{Regret}(T) \lesssim B_l^2 \cdot \sqrt{\dim_{\mathrm{VBE}}(\mathcal{H}, \Pi_{\mathcal{H}}, 1/\sqrt{HK}) \cdot |\mathcal{A}| \cdot \log(HK|\mathcal{H}|/\delta) \cdot H^2 K}. \tag{E.4}$$

---

[5]The policy $\pi_f \circ_h \mathrm{Unif}(\mathcal{A})$ means that when executing the exploration policy to collect data $\mathcal{D}_h$ at timestep h, the agent first executes policy π_f for the first $h - 1$ steps and then takes an action uniformly sampled from $\mathcal{A}$ at timestep h.

30

MDPs of bilinear class. In this part, we consider MDPs of bilinear class [20].

Definition E.6 (Bilinear class [20, 89]). *Given an MDP, a model-free hypothesis class $\mathcal{H}$, and a function $l_f : \mathcal{H} \times \mathcal{H} \times (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}) \mapsto \mathbb{R}$, we say the corresponding RL problem is in a bilinear class if there exist functions $W_h : \mathcal{H} \mapsto \mathcal{V}$ and $X_h : \mathcal{H} \mapsto \mathcal{V}$ for some Hilbert space $\mathcal{V}$, such that for all $f, g \in \mathcal{H}$ and $h \in [H]$, we have that*

$$\left| \mathbb{E}_{\pi_f} [Q_{h,f}(x_h, a_h) - R_h(x_h, a_h) - V_{h+1,f}(x_{h+1})] \right| \leq \left| \langle W_h(f) - W_h(f^*), X_h(f) \rangle_{\mathcal{V}} \right|,$$

$$\left| \mathbb{E}_{x_h \sim \pi_f, a_h \sim \widetilde{\pi}}[l_f(g; \xi_h)] \right| = \left| \langle W_h(g) - W_h(f^*), X_h(f) \rangle_{\mathcal{V}} \right|,$$

where $\widetilde{\pi}$ is either π_f for Q-type problems or π_g for V-type problems. Meanwhile, we make the assumption that $\sup_{f \in \mathcal{H}, h \in [H]} \|W_h(f)\|_2 \leq 1$ and $\sup_{f \in \mathcal{H}, h \in [H]} \|X_h(f)\|_2 \leq 1$.

For MDPs of bilinear class, we choose the function l as the function introduced in the definition of bilinear class. By choosing the exploration policy as $\pi_{\exp}(f) = \pi_f$ for Q-type problems and $\pi_{\exp}(f) = \pi_f \circ_h \mathrm{Unif}(\mathcal{A})$ for V-type problems, we can bound the generalized eluder coefficient for MDPs of bilinear class using the following lemma. To simplify the notation, we define $\mathcal{X}_h = \{X_h(f) : f \in \mathcal{H}\} \subseteq \mathcal{V}$ and $\mathcal{X} = \{\mathcal{X}_h : h \in [H]\}$.

Lemma E.7 (GEC for bilinear class, Lemma 3.22 in [89]). *Let the discrepancy ℓ function be chosen as (5.1) with l defined in Definition E.6. Define the maximum information gain $\gamma_K(\epsilon, \mathcal{X})$ as*

$$\gamma_K(\epsilon, \mathcal{X}) = \sum_{h=1}^{H} \max_{x_1, \cdots, x_K \in \mathcal{X}_h} \log \det \left(\mathcal{I}(\cdot) + \frac{1}{\epsilon} \sum_{s=1}^{K} x_s \langle x_s, \cdot \rangle_{\mathcal{V}} \right)$$

with $\mathcal{I}$ being the identity mapping. Then for Q-type problems, choosing $\pi_{\exp}(f) = \pi_f$, we have that

$$d_{\mathrm{GEC}}(\epsilon) \leq 2\gamma_K(\epsilon, \mathcal{X}).$$

For V-type problems, by choosing $\pi_{\exp}(f) = \pi_f \circ_h \mathrm{Unif}(\mathcal{A})$, we have that

$$d_{\mathrm{GEC}}(\epsilon) \leq 2|\mathcal{A}|\gamma_K(\epsilon, \mathcal{X}).$$

Proof of Lemma E.5. See Lemma 3.22 in [89] for a detailed proof. □

By combining Lemma E.7 and Corollary 5.2, we know that For Q-type bilinear class problem, it holds that with probability at least $1 - \delta$,

$$\mathrm{Regret}(T) \lesssim \sqrt{\gamma_K(1/\sqrt{HK}, \mathcal{X}) \cdot \log(HK|\mathcal{H}|/\delta) \cdot HK}, \tag{E.5}$$

and for V-type bilinear class problem, it holds that with probability at least $1 - \delta$,

$$\mathrm{Regret}(T) \lesssim \sqrt{\gamma_K(1/\sqrt{HK}, \mathcal{X}) \cdot |\mathcal{A}| \cdot \log(HK|\mathcal{H}|/\delta) \cdot HK}. \tag{E.6}$$

E.2 Examples of Model-based Online RL in MDPs

MDPs with low witness rank. We consider the example of MDPs with low witness rank [67, 2]. To introduce, we define the function class $\mathcal{V} = \{v : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]\}$.

Definition E.8 (Q-type/V-type witness rank [67, 2]). *An MDP is called of witness rank d if for any two models $f, f' \in \mathcal{H}$, there exists mappings $X_h : \mathcal{H} \mapsto \mathbb{R}^d$ and $W_h : \mathcal{H} \mapsto \mathbb{R}^d$ for each timestep h such that,*

$$\max_{v \in \mathcal{V}} \mathbb{E}_{x_h \sim \pi_f, a_h \sim \widetilde{\pi}} \left[\left(\mathbb{E}_{x' \sim \mathbb{P}_{h,f'}(\cdot|x_h, a_h)} - \mathbb{E}_{x' \sim \mathbb{P}_{h,f^*}(\cdot|x_h, a_h)} \right) [v(x_h, a_h, x')] \right] \geq \langle W_h(f'), X_h(f) \rangle,$$

$$\kappa_{\mathrm{wit}} \cdot \mathbb{E}_{x_h \sim \pi_f, a_h \sim \widetilde{\pi}} \left[\left(\mathbb{E}_{x' \sim \mathbb{P}_{h,f'}(\cdot|x_h, a_h)} - \mathbb{E}_{x' \sim \mathbb{P}_{h,f^*}(\cdot|x_h, a_h)} \right) [V_{h+1,f'}(x')] \right] \leq \langle W_h(f'), X_h(f) \rangle,$$

where $\widetilde{\pi}$ is either π_f for Q-type problems or $\pi_{f'}$ for V-type problems and $\kappa_{\mathrm{wit}} \in (0, 1]$ is a constant. Also, we let $\sup_{f \in \mathcal{H}, h \in [H]} \|W_h(f)\| \leq 1$ and $\sup_{f \in \mathcal{H}, h \in [H]} \|X_h(f)\| \leq 1$.

By choosing the exploration policy as $\pi_{\exp}(f) = \pi_f$ for Q-type problems and $\pi_{\exp}(f) = \pi_f \circ_h \mathrm{Unif}(\mathcal{A})$ for V-type problems, we can bound the generalized eluder coefficient by the following lemma.

31

Lemma E.9 (GEC for low witness rank, Lemma 3.22 in [89]). *Let the discrepancy function ℓ be chosen as* (5.4). *For Q-type problems, by choosing $\pi_{\exp}(f) = \pi_f$, we have that*

$$d_{\mathrm{GEC}}(\epsilon) \leq 4dH \cdot \log(1 + K/(\epsilon\kappa_{\mathrm{wit}}^2))/\kappa_{\mathrm{wit}}^2.$$

For V-type problems, by choosing $\pi_{\exp}(f) = \pi_f \circ_h \mathrm{Unif}(\mathcal{A})$, we have that

$$d_{\mathrm{GEC}}(\epsilon) \leq 4d|\mathcal{A}|H \cdot \log(1 + K/(\epsilon\kappa_{\mathrm{wit}}^2))/\kappa_{\mathrm{wit}}^2.$$

Proof of Lemma E.9. See Lemma 3.22 in [89] for a detailed proof. $\square$

By combining Lemma E.9 and Corollary 5.4, we know that For Q-type low witness rank problem, it holds that with probability at least $1 - \delta$,

$$\mathrm{Regret}(K) \lesssim \sqrt{4dH^2K \cdot \log(H|\mathcal{H}|/\delta) \cdot \log(1 + H^{1/2}K^{3/2}/\kappa_{\mathrm{wit}}^2)/\kappa_{\mathrm{wit}}^2}, \quad \text{(E.7)}$$

and for V-type low witness rank problem, it holds that with probability at least $1 - \delta$,

$$\mathrm{Regret}(K) \lesssim \sqrt{4d|\mathcal{A}|H^2K \cdot \log(H|\mathcal{H}|/\delta) \cdot \log(1 + H^{1/2}K^{3/2}/\kappa_{\mathrm{wit}}^2)/\kappa_{\mathrm{wit}}^2}. \quad \text{(E.8)}$$

E.3 Proof of Proposition 5.1

Proof of Proposition 5.1. To prove Proposition 5.1, we define the random variables $X_{h,f}^k$ as

$$X_{h,f}^k = l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)^2 - l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}); \mathcal{D}_h^k)^2, \quad \text{(E.9)}$$

for any $f \in \mathcal{H}$, where the operator $\mathcal{P}_h$ is introduced in Assumption 3.1. We first show that $X_{h,f}^k$ is an unbiased estimator of the discrepancy function $\ell_{f^k}(f)$. Consider that

$$
\begin{aligned}
&l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)^2 \\
&= \left(l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k) - l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}); \mathcal{D}_h^k) + l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}); \mathcal{D}_h^k) \right)^2 \\
&= \left(\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)] + l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}); \mathcal{D}_h^k) \right)^2 \\
&= \left(\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)] \right)^2 + l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}); \mathcal{D}_h^k)^2 \\
&\quad + 2\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)] \cdot l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}); \mathcal{D}_h^k), \quad \text{(E.10)}
\end{aligned}
$$

where in the second equality we apply the generalized Bellman completeness condition in Assumption 3.1. By the generalized Bellman completeness condition again, we also have that in (E.10),

$$
\begin{aligned}
&\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)} \left[\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)] \cdot l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}); \mathcal{D}_h^k) \right] \\
&= \mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)] \cdot \mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}); \mathcal{D}_h^k)] \\
&= \mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)] \\
&\quad \cdot \mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)} \left[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k) - \mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)] \right] \\
&= 0. \quad \text{(E.11)}
\end{aligned}
$$

Thus by combining (E.10) and (E.11), we can derive that

$$\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[X_{h,f}^k] = \left(\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}); \mathcal{D}_h^k)] \right)^2 = \ell_{f^k}(f; \mathcal{D}_h^k), \quad \text{(E.12)}$$

Now for each timestep h, we define a filtration $\{\mathcal{F}_{h,k}\}_{k=1}^K$, with

$$\mathcal{F}_{h,k} = \sigma \left(\bigcup_{s=1}^k \bigcup_{h=1}^H \mathcal{D}_h^s \right), \quad \text{(E.13)}$$

32

where $\mathcal{D}_h^s = \{x_h^s, a_h^s, r_h^s, x_{h+1}^s\}$. From previous arguments, we can derive that

$$\mathbb{E}[X_{h,f}^k | \mathcal{F}_{h,k-1}] = \mathbb{E}\left[\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot | x_h^k, a_h^k)}[X_{h,f^k}] \Big| \mathcal{F}_{h,k-1}\right] = \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^k)}[\ell_{f^k}(f; \xi_h)]. \quad \text{(E.14)}$$

and that

$$\mathbb{V}[X_{h,f}^k | \mathcal{F}_{h,k-1}] \leq \mathbb{E}[(X_{h,f}^k)^2 | \mathcal{F}_{h,k-1}] \leq 4B_l^2 \mathbb{E}[X_{h,f}^k | \mathcal{F}_{h,k-1}] = 4B_l^2 \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^k)}[\ell_{f^k}(f; \xi_h)], \quad \text{(E.15)}$$

where B_l is the upper bound of l defined in Assumption 3.1. By applying Lemma G.2, (E.14), and (E.15), we can obtain that with probability at least $1 - \delta$, for any $(h, k) \in [H] \times [K]$, $(f_h, f_{h+1}) \in \mathcal{H}_h \times \mathcal{H}_{h+1}^6$,

$$\left| \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f; \xi_h)] - \sum_{s=1}^{k-1} X_{h,f}^s \right| \quad \text{(E.16)}$$

$$\lesssim \frac{1}{2} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f; \xi_h)] + 8B_l^2 \log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta). \quad \text{(E.17)}$$

Rearranging terms in (E.16), we can further obtain that

$$-\sum_{s=1}^{k-1} X_{h,f}^s \lesssim -\frac{1}{2} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f; \xi_h)] + 8B_l^2 \log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta). \quad \text{(E.18)}$$

Meanwhile, by the definition of $X_{h,f}^k$ in (E.9) and the loss function L in (3.3), we have that

$$\begin{aligned}
\sum_{s=1}^{k-1} X_{h,f}^s &= \sum_{s=1}^{k-1} l_{f^s}((f_h, f_{h+1}), \mathcal{D}_h^s)^2 - \sum_{s=1}^{k-1} l_{f^k}((\mathcal{P}_h f_{h+1}, f_{h+1}), \mathcal{D}_h^s)^2 \\
&\leq \sum_{s=1}^{k-1} l_{f^s}((f_h, f_{h+1}), \mathcal{D}_h^s)^2 - \inf_{f_h' \in \mathcal{F}} \sum_{s=1}^{k-1} l_{f^s}((f_h', f_{h+1}), \mathcal{D}_h^s)^2 \\
&= L_h^{k-1}(f). \quad \text{(E.19)}
\end{aligned}$$

Thus by (E.18) and (E.19), we can derive that with probability at least $1 - \delta$, for any $f \in \mathcal{H}$, $k \in [K]$,

$$-\sum_{h=1}^{H} L_h^{k-1}(f) \lesssim -\frac{1}{2} \sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f; \xi_h)] + 8HB_l^2 \log(HK/\delta) + 16B_l^2 \log(|\mathcal{H}|). \quad \text{(E.20)}$$

Finally, we deal with the term $L_h^{k-1}(f^*)$. To this end, we invoke the following lemma.

Lemma E.10. *With probability at least $1 - \delta$, it holds that for each $k \in [K]$,*

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) \lesssim 8HB_l^2 \log(HK|\mathcal{H}|/\delta) + 16B_l^2 \log(|\mathcal{H}|).$$

Proof of Lemma E.10. To prove Lemma E.10, we define the random variables $W_{h,f}^k$ as

$$W_{h,f}^k = l_{f^k}((f_h, f_{h+1}^*); \mathcal{D}_h^k)^2 - l_{f^k}((f_h^*, f_{h+1}^*); \mathcal{D}_h^k)^2.$$

Using the same argument as (E.10) and (E.11), together with the condition $\mathcal{P}_h f_{h+1}^* = f_h^*$ in Assumption 3.1, we can show that

$$\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot | x_h^k, a_h^k)}[W_{h,f}^k] = \left(\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot | x_h^k, a_h^k)}[l_{f^k}((f_h, f_{h+1}^*); \mathcal{D}_h)]\right)^2. \quad \text{(E.21)}$$

---

[6]Here $l_{f^s}((f_h, f_{h+1}); \mathcal{D}_h^s)$ and $\ell_{f^s}(f; \xi_h)$ depend on f only through (f_h, f_{h+1}).

33

Under the filtration $\{\mathcal{F}_{h,k}\}_{k=1}^{K}$ defined in the proof of Proposition 5.1, i.e, (E.13), one can derive that

$$\mathbb{E}[W_{h,f}^{k}|\mathcal{F}_{h,k-1}] = \mathbb{E}\left[\mathbb{E}_{x_{h+1}^{k}\sim\mathbb{P}_h(\cdot|x_h^k,a_h^k)}[W_{h,f^k}]\Big|\mathcal{F}_{h,k-1}\right]$$
$$= \mathbb{E}_{\mathcal{D}_h\sim\pi_{\exp}(f^k)}\left[\left(\mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x_h,a_h)}[l_{f^k}((f_h,f_{h+1}^*);\mathcal{D}_h)]\right)^2\right], \qquad \text{(E.22)}$$

and that

$$\mathbb{V}[W_{h,f}^{k}|\mathcal{F}_{h,k-1}] \leq 4B_l^2\mathbb{E}[X_{h,f}^{k}|\mathcal{F}_{h,k-1}]$$
$$= 4B_l^2\mathbb{E}_{\mathcal{D}_h\sim\pi_{\exp}(f^k)}\left[\left(\mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x_h,a_h)}[l_{f^k}((f_h,f_{h+1}^*);\mathcal{D}_h)]\right)^2\right]. \qquad \text{(E.23)}$$

By applying Lemma G.2, (E.22), and (E.23), we obtain that with probability at least $1-\delta$, for any $(h,k)\in[H]\times[K]$ and $(f_h,f_{h+1})\in\mathcal{H}_h\times\mathcal{H}_{h+1}$,

$$\left|\sum_{s=1}^{k-1}W_{h,f}^{s} - \sum_{s=1}^{k-1}\mathbb{E}_{\mathcal{D}_h\sim\pi_{\exp}(f^k)}\left[\left(\mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x_h,a_h)}[l_{f^s}((f_h,f_{h+1}^*);\mathcal{D}_h)]\right)^2\right]\right| \lesssim$$
$$+ \sqrt{\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta)\cdot\sum_{s=1}^{k-1}\mathbb{E}_{\mathcal{D}_h\sim\pi_{\exp}(f^s)}\left[\left(\mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x_h,a_h)}[l_{f^k}((f_h,f_{h+1}^*);\mathcal{D}_h)]\right)^2\right]}$$
$$+ 4B_l^2\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta).$$

Rearranging terms, we have that with probability at least $1-\delta$, for any $f\in\mathcal{H}$, $(h,k)\in[H]\times[K]$,

$$-\sum_{s=1}^{k-1}W_{h,f}^{s}$$
$$\lesssim 4B_l^2\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta) - \sum_{s=1}^{k-1}\mathbb{E}_{\mathcal{D}_h\sim\pi_{\exp}(f^s)}\left[\left(\mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x_h,a_h)}[l_{f^s}((f_h,f_{h+1}^*);\mathcal{D}_h)]\right)^2\right]$$
$$+ \sqrt{\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta)\cdot\sum_{s=1}^{k-1}\mathbb{E}_{\mathcal{D}_h\sim\pi_{\exp}(f^s)}\left[\left(\mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x_h,a_h)}[l_{f^s}((f_h,f_{h+1}^*);\mathcal{D}_h)]\right)^2\right]}$$
$$\lesssim 8B_l^2\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta),$$

where in the second inequality we use the inequality $-x^2+ax\leq a^2/4$. Thus, with probability at least $1-\delta$, for any $k\in[K]$, it holds that

$$\sum_{h=1}^{H}L_h^{k-1}(f^*) = \sum_{h=1}^{H}\left(\sum_{s=1}^{k-1}l_{f^k}((f_h^*,f_{h+1}^*);\mathcal{D}_h^s)^2 - \inf_{f_h\in\mathcal{H}_h}\sum_{s=1}^{k-1}l_{f^k}((f_h,f_{h+1}^*);\mathcal{D}_h^s)^2\right)$$
$$= \sum_{h=1}^{H}\sup_{f_h\in\mathcal{H}_h}\sum_{s=1}^{k-1}-W_{h,f}^{s} \lesssim 8HB_l^2\log(HK/\delta) + 16B_l^2\log(|\mathcal{H}|).$$

This finishes the proof of Lemma E.10. $\qquad\square$

Finally, combining (E.20) and Lemma E.10, with probability at least $1-\delta$, for any $f\in\mathcal{H}$, $k\in[K]$,

$$\sum_{h=1}^{H}L_h^{k-1}(f^*) - L_h^{k-1}(f)$$
$$\lesssim -\frac{1}{2}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\pi_{\exp}(f^s)}[\ell_{f^s}(f;\xi_h)] + 16HB_l^2\log(HK/\delta) + 32B_l^2\log(|\mathcal{H}|).$$

This finishes the proof of Proposition 5.1. $\qquad\square$

34

E.4 Proof of Proposition 5.3

Proof of Proposition 5.3. For notational simplicity, given $f \in \mathcal{H}$, we denote the random variables $X_{h,f}^k$ as

$$X_{h,f}^k = \log\left(\frac{\mathbb{P}_{h,f^*}(x_{h+1}^k|x_h^k, a_h^k)}{\mathbb{P}_{h,f}(x_{h+1}^k|x_h^k, a_h^k)}\right). \tag{E.24}$$

Then by the definition of L_h^k in (3.5), we have that,

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) = -\sum_{h=1}^{H}\sum_{s=1}^{k-1} X_{h,f}^s. \tag{E.25}$$

Now we define a filtration $\{\mathcal{F}_{h,k}\}_{k=1}^{K}$ for each step $h \in [H]$ with

$$\mathcal{F}_{h,k} = \sigma\left(\bigcup_{s=1}^{k}\bigcup_{h=1}^{H}\mathcal{D}_h^s\right). \tag{E.26}$$

Then by (E.24) we know that $X_{h,f}^k \in \mathcal{F}_{h,k}$ for any $(h,k) \in [H] \times [K]$. Therefore, by applying Lemma G.1, we have that with probability at least $1 - \delta$, for any $(h,k) \in [H] \times [K]$ and $f_h \in \mathcal{H}_h$,

$$-\frac{1}{2}\sum_{s=1}^{k-1} X_{h,f}^s \le \sum_{s=1}^{k-1}\log\mathbb{E}\left[\exp\left\{-\frac{1}{2}X_{h,f}^s\right\}\Big|\mathcal{F}_{s-1}\right] + \log(H|\mathcal{H}_h|/\delta). \tag{E.27}$$

Meanwhile, we can calculate that in (E.27), the conditional expectation equals to

$$\mathbb{E}\left[\exp\left\{-\frac{1}{2}X_{h,f}^s\right\}\Big|\mathcal{F}_{s-1}\right]$$

$$= \mathbb{E}\left[\sqrt{\frac{\mathbb{P}_{h,f}(x_{h+1}^s|x_h^s, a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s|x_h^s, a_h^s)}}\Big|\mathcal{F}_{s-1}\right]$$

$$= \mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s), x_{h+1}^s\sim\mathbb{P}_{h,f^*}(\cdot|x_h^s,a_h^s)}\left[\sqrt{\frac{\mathbb{P}_{h,f}(x_{h+1}^s|x_h^s, a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s|x_h^s, a_h^s)}}\right]$$

$$= \mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\left[\int_{\mathcal{S}}\sqrt{\mathbb{P}_{h,f}(x_{h+1}^s|x_h^s, a_h^s)\cdot\mathbb{P}_{h,f^*}(x_{h+1}^s|x_h^s, a_h^s)}dx_{h+1}^s\right]$$

$$= 1 - \frac{1}{2}\mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\left[\int_{\mathcal{S}}\left(\sqrt{\mathbb{P}_{h,f}(x_{h+1}^s|x_h^s, a_h^s)} - \sqrt{\mathbb{P}_{h,f^*}(x_{h+1}^s|x_h^s, a_h^s)}\right)^2 dx_{h+1}^s\right]$$

$$= 1 - \mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\left[D_{\mathrm{H}}(\mathbb{P}_{h,f^*}(\cdot|x_h^s, a_h^s)\|\mathbb{P}_{h,f}(\cdot|x_h^s, a_h^s))\right], \tag{E.28}$$

where the first equality uses the definition of $X_{h,f}^s$ in (E.24), the second equality is due to the fact that $\xi_h^s \sim \pi^s$ and $\pi^s \in \mathcal{F}_{s-1}$, and the last equality uses the definition of Hellinger distance D_{H}. Thus by combining (E.27) and (E.28), we can derive that

$$-\frac{1}{2}\sum_{s=1}^{k-1} X_{h,f}^s \le \sum_{s=1}^{k-1}\mathbb{E}\left[\exp\left\{-\frac{1}{2}X_{h,f}^s\right\}\Big|\mathcal{F}_{s-1}\right] - 1 + \log(H|\mathcal{H}_h|/\delta)$$

$$= -\sum_{s=1}^{k-1}\mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\left[D_{\mathrm{H}}(\mathbb{P}_{h,f^*}(\cdot|x_h^s, a_h^s)\|\mathbb{P}_{h,f}(\cdot|x_h^s, a_h^s))\right] + \log(H|\mathcal{H}_h|/\delta),$$

where in the first inequality we use the fact that $\log(x) \le x - 1$. Finally, by plugging in the definition of $X_{h,f}^s$, summing over $h \in [H]$, we have that with probability at least $1 - \delta$, for any $f \in \mathcal{H}$, any

35

$k \in [K]$, it holds that

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) = -\sum_{h=1}^{H}\sum_{s=1}^{k-1} X_{h,f}^s$$

$$\leq -2\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{(x_h^s, a_h^s) \sim \pi_{\exp}(f^s)} \left[D_{\mathrm{H}}(\mathbb{P}_{h,f^*}(\cdot | x_h^s, a_h^s) \| \mathbb{P}_{h,f}(\cdot | x_h^s, a_h^s)) \right]$$

$$+ 2H\log(H/\delta) + 2\log(|\mathcal{H}|),$$

$$= -2\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f; \xi_h)] + 2H\log(H/\delta) + 2\log(|\mathcal{H}|). \qquad \text{(E.29)}$$

This finishes the proof of Proposition 5.3. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

F Proofs for Model-free and Model-based Online RL in Two-player Zero-sum MGs

F.1 Proof of Proposition C.11

Proof of Proposition C.11. To begin with, we need to introduce the performance difference lemma in two-player zero-sum MG, which are presented in Lemma 1 and Lemma 2 in Xiong et al. [80].

Lemma F.1 (Value decomposition for the max-player). *Let $\mu = \mu_f$ and ν be an arbitrary policy taken by the min-player. It holds that*

$$V_{1,f}(x_1) - V_1^{\mu,\nu}(x_1) \leq \sum_{h=1}^{H} \mathbb{E}_{\xi_h \sim (\mu,\nu)} [\mathcal{E}_h(f_h, f_{h+1}; \xi_h)] \qquad \text{(F.1)}$$

where max-player Bellman error $\mathcal{E}_h(f_h, f_{h+1}; \xi_h)$ is defined as

$$\mathcal{E}_h(f_h, f_{h+1}; \xi_h) = Q_{h,f}(x_h, a_h, b_h) - r_h - (\mathbb{P}_h V_{h+1,f})(x_h, a_h, b_h), \qquad \text{(F.2)}$$

and $\xi_h = (x_h, a_h, b_h, r_h)$. (Actually, this coincides with the NE Bellman error defined in (C.10).)

Lemma F.2 (Value decomposition for the min-player). *Suppose that $\mu = \mu_f$ is taken by the max-player and g is the hypothesis selected by the min-player. Let ν be the policy taken by the min-player. Then, it holds that*

$$V_1^{\mu,\nu}(x_1) - V_{1,g}^{\mu,\dagger}(x_1) = -\sum_{h=1}^{H} \mathbb{E}_{\xi_h \sim (\mu,\nu)} [\mathcal{E}_h^\mu(g_h, g_{h+1}; \xi_h)], \qquad \text{(F.3)}$$

where the min-player Bellman error $\mathcal{E}_h^\mu(g_h, g_{h+1}; \xi_h)$ is defined as

$$\mathcal{E}_h^\mu(g_h, g_{h+1}; \xi_h) = Q_{h,g}^{\mu,\dagger}(x_h, a_h, b_h) - r_h - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x_h, a_h, b_h), \qquad \text{(F.4)}$$

and $\xi_h = (x_h, a_h, b_h, r_h)$.

We note that the value decomposition for the max-player is an inequality because of the property of minimax formulation. Note also that the right side of (F.3) is a general version of the right side of (F.1) when choosing $\mu = \mu_f$. Now we are ready to prove Proposition C.11. The lemmas suggest that we only need to upper-bound the term $\sum_{k=1}^{K}\sum_{h=1}^{H} |\mathbb{E}_{\boldsymbol{\pi}^k}[\mathcal{E}_h^\mu(g_h^k, g_{h+1}^k; \xi_h)]|$ for all admissible max-player policy μ. To this end, we provide a more general result by the following proposition. For simplicity, we denote by $\boldsymbol{\pi}^k = (\mu^k, \nu^k)$.

Proposition F.3. *For a d-dimensional two-player zero-sum Markov game, we assume that its expected min-player bellman error can be decomposed as follows*

$$\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s} [\mathcal{E}_h^\mu(g_h, g_{h+1}; \xi_h)] = \langle W_h(g, \mu), X_h(g, \boldsymbol{\pi}^s, \mu) \rangle, \qquad \text{(F.5)}$$

36

for some $W_h(g,\mu), X_h(g,\boldsymbol{\pi},\mu) \in \mathbb{R}^d$, *and the discrepancy function* $\ell_{g',\mu}(g;\xi_h)$ *can be lower bounded as follows*

$$|\langle W_h(g,\mu), X_h(g',\boldsymbol{\pi},\mu)\rangle|^2 \leq \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}}[\ell_{g',\mu}(g;\xi_h)], \tag{F.6}$$

for all the admissible max-player policy $\mu \in \mathbf{M}$. *Also, we assume that* $\|W_h(\cdot,\cdot)\|_2 \leq B_W$, $\|X_h(\cdot,\cdot,\cdot)\|_2 \leq B_X$ *for some* $B_W, B_X > 0$ *and for all timestep* $h \in [H]$. *Then it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \left| \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}\left[\mathcal{E}_h^\mu(g_h^k, g_{h+1}^k; \xi_h)\right] \right| \tag{F.7}$$

$$\leq \frac{\widetilde{d}(\epsilon)}{4\eta} + \frac{\eta}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\boldsymbol{\pi}^s}[\ell_{g^s,\mu}(g^k;\xi_h)] + 2\min\{HK, 2\widetilde{d}(\epsilon)\} + HKB_W\epsilon,$$

for all admissible max-player policy $\mu \in \mathbf{M}$, $\epsilon \in [0,1]$, $\eta > 0$, *and* $\widetilde{d}(\epsilon) := d\log(1 + KB_X^2/(d\epsilon))$.

Proof of Proposition F.3. We prove this result following a similar procedure as in the proof of Lemma 3.20 in [89], where they prove that the low-GEC class contains the bilinear class. We denote by

$$\Sigma_{h,k} = \epsilon I_d + \sum_{s=1}^{k-1}X_h(g^s,\boldsymbol{\pi}^s,\mu)X_h(g^s,\boldsymbol{\pi}^s,\mu)^\top.$$

By Lemma F.3 in [20] and Lemma G.3, we first have the following equality,

$$\sum_{s=1}^{k}\min\left\{\|X_h(g^s,\boldsymbol{\pi}^s,\mu)\|_{\Sigma_{h,s}^{-1}}, 1\right\} \leq 2\widetilde{d}(\epsilon), \tag{F.8}$$

for all $\epsilon \in [0,1]$. Here $\widetilde{d}(\epsilon)$ is defined in Proposition F.3. Now, since the reward is bounded by $[0,1]$, we have the following inequalities,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left|\mathbb{E}_{\boldsymbol{\pi}^k}\left[\mathcal{E}_h^\mu(g_h^k, g_{h+1}^k; \xi_h)\right]\right|$$

$$= \sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \langle W_h(g^k,\mu), X_h(g^k,\boldsymbol{\pi}^k,\mu)\rangle\}\mathbf{1}\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}} \leq 1\right\}$$

$$+ \sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \langle W_h(g^k,\mu), X_h(g^k,\boldsymbol{\pi}^k,\mu)\rangle\}\mathbf{1}\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}} > 1\right\}$$

$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\langle W_h(g^k,\mu), X_h(g^k,\boldsymbol{\pi}^k,\mu)\rangle\mathbf{1}\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}} \leq 1\right\} + \min\{HK, \widetilde{d}(\epsilon)\}$$

$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\underbrace{\|W_h(g^k,\mu)\|_{\Sigma_{h,k}}\min\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}}, 1\right\}}_{(A)_{h,k}} + \min\{HK, \widetilde{d}(\epsilon)\}, \tag{F.9}$$

where the first equality relies on the assumption in Proposition F.3, the second inequality comes from (F.8), and the last inequality is based on Cauchy Schwarz inequality. Now we expand term $(A)_{h,k}$ in (F.9) as follows.

$$\|W_h(g^k,\mu)\|_{\Sigma_{h,k}} \leq \sqrt{\epsilon}B_W + \left[\sum_{s=1}^{k-1}|\langle W_h(g^k,\mu), X_h(g^s,\boldsymbol{\pi}^s,\mu)\rangle|^2\right]^{1/2},$$

37

where we use the fact that $\|W_h(g^k, \mu)\|_2 \leq B_W$. Thus we have that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} (\mathrm{A})_{h,k} \tag{F.10}$$

$$\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \left(\sqrt{\epsilon} B_W + \left[\sum_{s=1}^{k-1} |\langle W_h(g^k, \mu), X_h(g^s, \boldsymbol{\pi}^s, \mu) \rangle|^2 \right]^{1/2} \right) \cdot \min \left\{ \|X_h(g^k, \boldsymbol{\pi}^k, \mu)\|_{\Sigma_{h,k}^{-1}}, 1 \right\}$$

$$\leq \left[\sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\epsilon} B_W \right]^{1/2} \cdot \left[\sum_{k=1}^{K} \sum_{h=1}^{H} \min \left\{ \|X_h(g^k, \boldsymbol{\pi}^k, \mu)\|_{\Sigma_{h,k}^{-1}}, 1 \right\} \right]^{1/2}$$

$$+ \left[\sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s=1}^{k-1} |\langle W_h(g^k, \mu), X_h(g^s, \boldsymbol{\pi}^s, \mu) \rangle|^2 \right]^{1/2} \cdot \left[\sum_{k=1}^{K} \sum_{h=1}^{H} \min \left\{ \|X_h(g^k, \boldsymbol{\pi}^k, \mu)\|_{\Sigma_{h,k}^{-1}}, 1 \right\} \right]^{1/2}$$

$$\leq \sqrt{H B_W K \epsilon \cdot \min\{2\widetilde{d}(\epsilon), HK\}} + \left[2\widetilde{d}(\epsilon) \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s=1}^{k-1} |\langle W_h(g^k, \mu), X_h(g^s, \boldsymbol{\pi}^s, \mu) \rangle|^2 \right]^{1/2}$$

$$\leq \sqrt{H K B_W \epsilon \cdot \min\{2\widetilde{d}(\epsilon), HK\}} + \left[2\widetilde{d}(\epsilon) \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s} [\ell_{g^s, \mu}(g^k; \xi_h)] \right]^{1/2},$$

where the second inequality is the result of Cauchy-Schwarz inequality, the third inequality comes from (F.8), and the last inequality is derived from (F.6). Back to the analysis for (F.9), we have that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left| \mathbb{E}_{\boldsymbol{\pi}^k} \left[\mathcal{E}_h^\mu(g_h^k, g_{h+1}^k; \xi_h) \right] \right|$$

$$\leq \sqrt{H K B_W \epsilon \cdot \min\{2\widetilde{d}(\epsilon), HK\}}$$

$$+ \left[2\widetilde{d}(\epsilon) \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s} [\ell_{g^s, \mu}(g^k; \xi_h)] \right]^{1/2} + \min\{HK, 2\widetilde{d}(\epsilon)\}$$

$$\leq \left[2\widetilde{d}(\epsilon) \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s} [\ell_{g^s, \mu}(g^k; \xi_h)] \right]^{1/2} + 2\min\{HK, 2\widetilde{d}(\epsilon)\} + H K B_W \epsilon$$

$$\leq \frac{\widetilde{d}(\epsilon)}{4\eta} + \frac{\eta}{2} \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s} [\ell_{g^s, h}(g^k; \xi_h)] + 2\min\{HK, 2\widetilde{d}(\epsilon)\} + H K B_W \epsilon,$$

where the second inequality comes from the AM-GM inequality and the last inequality uses the basic inequality $2ab \leq a^2 + b^2$. Here $\eta > 0$ can be arbitrarily chosen. Then we finish our proof to Proposition F.3. $\qquad \square$

Back to our proof of Proposition C.11, we first check the conditions of Proposition F.3 for linear two-player zero-sum MGs. By Definition C.10 and the choice of model-free hypothesis class (C.28), we know that for any $g \in \mathcal{H}$ and $\mu \in \mathbf{N}$, it holds that

$$Q_{h,g}(x, a, b) - r_h(x, a, b) - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x, a, b)$$

$$= \phi_h(x, a, b)^\top \left(\theta_{h,g} - \alpha_h - \int_{\mathcal{S}} \psi_h^\star(x') V_{h+1,g}^{\mu,\dagger}(x') \mathrm{d}x' \right),$$

where $\theta_{h,g}$ denotes the parameter of $Q_{h,g}$ and α_h is the reward parameter (see Definition C.10). Thus we can define $X_h(g, \boldsymbol{\pi}, \mu) = \mathbb{E}_{\boldsymbol{\pi}}[\phi_h(x, a, b)]$ and

$$W_h(g, \mu) = \theta_{h,g} - \alpha_h - \int_{\mathcal{S}} \psi_h^\star(x') V_{h+1,g}^{\mu,\dagger}(x') \mathrm{d}x'.$$

This specifies condition (F.5) of Proposition F.3. By Jansen inequality and the definition of ℓ_μ in (C.24), it is obvious that the condition (F.6) of Proposition F.3 holds. By the assumptions of linear

38

two-player zero-sum MGs in Definition C.10, we have $B_X \leq 1$ and $B_W \leq 4H\sqrt{d}$. Thus by applying Proposition F.3, we have that

$$\sum_{k=1}^{K} V_1^{\boldsymbol{\pi}^k}(x_1) - V_{1,g^k}^{\mu^k,\dagger}(x_1) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} \left| \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k} \left[\mathcal{E}_h^{\mu}(g_h^k, g_{h+1}^k; \xi_h) \right] \right|$$

$$\leq \frac{\widetilde{d}(\epsilon)}{4\eta} + \frac{\eta}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}[\ell_{g^s,\mu}(g^k; \xi_h)]$$

$$+ 2\min\{HK, 2\widetilde{d}(\epsilon)\} + 4\sqrt{d}H^2 K\epsilon,$$

with $\widetilde{d}(\epsilon) = d\log(1 + K/d\epsilon)$ and any $\eta > 0$. This proves the second inequality of Assumption C.5. For the first inequality in Assumption C.5, we take $g^k = f^k$, $\mu = \mu_{f^k}$, and we can then similarly prove that

$$\sum_{k=1}^{K} V_{1,f^k}(x_1) - V_1^{\boldsymbol{\pi}^k}(x_1) \leq \frac{\widetilde{d}(\epsilon)}{4\eta} + \frac{\eta}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}[\ell_{f^s}(f^k; \xi_h)]$$

$$+ 2\min\{HK, 2\widetilde{d}(\epsilon)\} + 4\sqrt{d}H^2 K\epsilon,$$

with $\widetilde{d}(\epsilon) = d\log(1 + K/d\epsilon)$ and any $\eta > 0$. This proves that $d_{\mathrm{TGEC}}(\epsilon) \leq \widetilde{d}(\epsilon)$.

As for the analysis for covering number, we apply the standard analysis for the covering number of $\mathbb{R}^d$-ball to obtain that

$$\log \mathcal{N}(\mathcal{H}, \epsilon, \|\cdot\|_\infty) \leq d\log\left(\frac{3}{\epsilon}\right) + d\log\left(\frac{\mathbf{Vol}(\mathcal{H})}{\mathbf{Vol}(B_d)}\right),$$

for all $\epsilon \leq 1$ and the unit ball B_d in $\mathbb{R}^d$ space. Selecting $\epsilon = 1/K$, we finish the proof of Proposition C.11. $\qquad\square$

F.2 Proof of Proposition C.16

Proof of Proposition C.16. Similar to the proof of Proposition C.11, we can apply Lemma F.1, Lemma F.2, and Proposition F.3 to obtain the upper bound of TGEC for linear mixture two-player zero-sum MGs. First we need to check the conditions of Proposition F.3. Note that

$$Q_{h,g}^{\mu,\dagger}(x,a,b) - r_h - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x,a,b) = (\mathbb{P}_{h,g} V_{h+1,g}^{\mu,\dagger})(x,a,b) - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x,a,b)$$

$$= \left(\theta_{h,g} - \theta_h^\star\right)^\top \left(\int_{\mathcal{S}} \phi_h(x,a,b,x') V_{h+1,g}^{\mu,\dagger}(x')\mathrm{d}x'\right),$$
(F.11)

where the first equality comes from the Bellman equation, and the second equality is derived from the definition of linear mixture two-player zero-sum MG (Definition C.15). Here $\theta_{h,g}$ denotes the parameter of $\mathbb{P}_{h,g}$. Hence we can define X_h and W_h as

$$X_h(g,\boldsymbol{\pi},\mu) := \mathbb{E}_{\boldsymbol{\pi}}\left[\int_{\mathcal{S}} \phi_h(x,a,b,x') V_{h+1,g}^{\mu,\dagger}(x')\mathrm{d}x'\right], \quad W_h(g,\mu) := \theta_{h,g} - \theta_h^\star. \quad \text{(F.12)}$$

This specifies condition (F.5) of Proposition F.3. By the assumptions of linear mixture two-player zero-sum MGs in Definition C.15, we can obtain that $B_X \leq 1$ and $B_W \leq 4H\sqrt{d}$. As for condition (F.6), different from the proof of Proposition C.11, since we use Hellinger distance as the discrepancy function ℓ for the model-based hypothesis, we propose to connect it to the model-free discrepancy function (C.24). Notice that

$$\left(Q_{h,g}^{\mu,\dagger}(x,a,b) - r_h - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x,a,b)\right)^2 = \left((\mathbb{P}_{h,g} V_{h+1,g}^{\mu,\dagger})(x,a,b) - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x,a,b)\right)^2$$

$$\leq 4\|V_{h+1,g}^{\mu,\dagger}(\cdot)\|_\infty^2 \cdot D_{\mathrm{TV}}(\mathbb{P}_{h,g}(\cdot\,|\,x,a,b)\|\mathbb{P}_h(\cdot\,|\,x,a,b))^2$$

$$\leq 2H^2 D_{\mathrm{H}}(\mathbb{P}_{h,g}(\cdot\,|\,x,a,b)\|\mathbb{P}_h(\cdot\,|\,x,a,b))^2$$

$$\leq 2H^2 D_{\mathrm{H}}(\mathbb{P}_{h,g}(\cdot\,|\,x,a,b)\|\mathbb{P}_h(\cdot\,|\,x,a,b)), \quad \text{(F.13)}$$

39

where the second equality comes from Holder inequality and the fact that the TV distance $D_{\text{TV}}(p\|q) = \|p-q\|_1/2$ for any two distributions p and q, the third inequality follows from the fact that $D_{\text{TV}}(p\|q) \leq \sqrt{2}D_{\text{H}}(p\|q)$, and the last inequality follows from the fact that $D_{\text{H}}(p\|q) \leq 1$. This shows that the model-based discrepancy function defined in (C.30) upper-bounds the model-free discrepancy function up to a factor $2H^2$, that is,

$$
\begin{aligned}
\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}}[\ell_{g',\mu}(g;\xi_h)] &= \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}}[D_{\text{H}}(\mathbb{P}_{h,g}(\cdot|x_h,a_h,b_h)\|\mathbb{P}_h(\cdot|x_h,a_h,b_h))] \\
&\geq \frac{1}{2H^2}\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}}\left[\left(Q_{h,g}^{\mu,\dagger}(x_h,a_h,b_h) - r_h - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x_h,a_h,b_H)\right)^2\right] \\
&= |\langle W_h(g,\mu), X_h(g,\boldsymbol{\pi},\mu)\rangle|^2.
\end{aligned}
\tag{F.14}
$$

Thus by applying Proposition F.3, we have that

$$
\begin{aligned}
\sum_{k=1}^{K} V_1^{\boldsymbol{\pi}^k} - V_{1,g^k}^{\mu^k,\dagger} &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left|\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}\left[\mathcal{E}_h^{\mu}(g_h^k, g_{h+1}^k;\xi_h)\right]\right| \\
&\leq \frac{\widetilde{d}(\epsilon)}{4\eta} + \frac{\eta}{4H^2}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}[\ell_{g^s,\mu}(g^k;\xi_h)] \\
&\quad + 2\min\{HK, 2\widetilde{d}(\epsilon)\} + 4\sqrt{d}H^2K\epsilon \\
&= \frac{\bar{d}(\epsilon)}{4\eta'} + \frac{\eta'}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}[\ell_{g^s,\mu}(g^k;\xi_h)] \\
&\quad + 2\min\{HK, 2\bar{d}(\epsilon)\} + 4\sqrt{d}H^2K\epsilon,
\end{aligned}
$$

with $\bar{d}(\epsilon) = 2H^2\widetilde{d}(\epsilon) = 2H^2 d\log(1+K/d\epsilon)$ and any $\eta > 0$ and $\eta' = \eta/(2H^2)$. This proves the second inequality of Assumption C.5. For the first inequality in Assumption C.5, we take $g^k = f^k$ and let $\mu = \mu_{f^k}$, and we can then also similarly prove that

$$
\sum_{k=1}^{K} V_{1,f^k} - V_1^{\boldsymbol{\pi}^k} \leq \frac{\bar{d}(\epsilon)}{4\eta'} + \frac{\eta'}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}[\ell_{f^s}(f^k;\xi_h)] + 2\min\{HK, 2\bar{d}(\epsilon)\} + 4\sqrt{d}H^2K\epsilon.
$$

This proves that $d_{\text{TGEC}}(\epsilon) \leq \bar{d}(\epsilon)$. As for the analysis of the covering number, it suffices to repeat the same as the proof of Proposition C.11. This finishes the proof of Proposition C.16. $\qquad\square$

F.3 Proof of Proposition C.8

Proof of Proposition C.8. We first prove the *first* inequality of Proposition C.8. To this end, we define the random vairable $X_{h,f}^k$ as

$$
\begin{aligned}
X_{h,f}^k = &\left(Q_{h,f}(x_h^k, a_h^k, b_h^k) - r_h^k - V_{h+1,f}(x_{h+1}^k)\right)^2 \\
&- \left(V_{h+1,f}(x_{h+1}^k) - \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h^k,a_h^k,b_h^k)}[V_{h+1,f}(x_{h+1})]\right)^2.
\end{aligned}
\tag{F.15}
$$

After a calculation similar to (E.10) and (E.11), we can derive that

$$
\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k,a_h^k,b_h^k)}[X_{h,f}^k] = \left(Q_{h,f}(x_h^k, a_h^k, b_h^k) - r_h^k - \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h^k,a_h^k,b_h^k)}[V_{h+1,f}(x_{h+1})]\right)^2.
$$

Now for each timestep h, we define a filtration $\{\mathcal{F}_{h,k}\}_{k=1}^{K}$ with

$$
\mathcal{F}_{h,k} = \sigma\left(\bigcup_{s=1}^{k}\bigcup_{h=1}^{H}\mathcal{D}_h^s\right),
\tag{F.16}
$$

where $\mathcal{D}_h^s = \{x_h^s, a_h^s, b_h^s, r_h^s, x_{h+1}^s\}$. With previous arguments, we can derive that

$$
\mathbb{E}[X_{h,f}^k|\mathcal{F}_{h,k-1}] = \mathbb{E}\left[\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k,a_h^k,b_h^k)}[X_{h,f^k}]|\mathcal{F}_{h,k-1}\right] = \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}[\ell_{f^k}(f;\xi_h)],
\tag{F.17}
$$

40

and that

$$\mathbb{V}[X_{h,f}^k | \mathcal{F}_{h,k-1}] \leq \mathbb{E}[(X_{h,f}^k)^2 | \mathcal{F}_{h,k-1}] \leq 4B_f^2 \mathbb{E}[X_{h,f}^k | \mathcal{F}_{h,k-1}] = 4B_f^2 \mathbb{E}_{\xi_h \sim \pi^k}[\ell_{f^k}(f; \xi_h)], \tag{F.18}$$

where B is the upper bound of hypothesis in $\mathcal{H}$ by Assumption C.4. By applying Lemma G.2, (F.17), and (F.18), we can obtain that with probability at least $1 - \delta$, for any $(h, k) \in [H] \times [K]$ and $(f_h, f_{h+1}) \in \mathcal{H}_h \times \mathcal{H}_{h+1}$,

$$\left| \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^s}[\ell_{f^s}(f; \xi_h)] - \sum_{s=1}^{k-1} X_{h,f}^s \right| \lesssim \frac{1}{2} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^s}[\ell_{f^s}(f; \xi_h)] + 8B_f^2 \log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta). \tag{F.19}$$

Rearranging terms in (F.19), we can further obtain that

$$- \sum_{s=1}^{k-1} X_{h,f}^s \lesssim -\frac{1}{2} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^s}[\ell_{f^s}(f; \xi_h)] + 8B_f^2 \log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta). \tag{F.20}$$

Meanwhile, by the definition of $X_{h,f}$ in (F.15) and the loss function L in (C.15), we have that

$$\sum_{s=1}^{k-1} X_{h,f}^s$$

$$= \sum_{s=1}^{k-1} \left(Q_{h,f}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s) \right)^2$$

$$\quad - \sum_{s=1}^{k-1} \left(V_{h+1,f}(x_{h+1}^s) - \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot | x_h^s, a_h^s, b_h^s)}[V_{h+1,f}(x_{h+1})] \right)^2$$

$$= \sum_{s=1}^{k-1} \left(Q_{h,f}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s) \right)^2$$

$$\quad - \sum_{s=1}^{k-1} \left(\mathcal{T}_h f(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s) \right)^2$$

$$\leq \sum_{s=1}^{k-1} \left(Q_{h,f}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s) \right)^2$$

$$\quad - \inf_{f_h' \in \mathcal{H}_h} \sum_{s=1}^{k-1} \left(Q_{h,f'}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s) \right)^2$$

$$= L_h^{k-1}(f). \tag{F.21}$$

where the last inequality follows from the completeness assumption (Assumption C.4). Combining (F.20) and (F.21), we can derive that with probability at least $1 - \delta$, for any $f \in \mathcal{H}$, $k \in [K]$,

$$- \sum_{h=1}^{H} L_h^{k-1}(f) \lesssim -\frac{1}{2} \sum_{h=1}^{H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^s}[\ell_{f^s}(f; \xi_h)] + 8HB_f^2 \log(HK/\delta) + 16B_f^2 \log(|\mathcal{H}|). \tag{F.22}$$

Finally, we deal with the term $L_h^k(f^*)$. To this end, we invoke the following lemma.

Lemma F.4. *With probability at least $1 - \delta$, it holds that for each $k \in [K]$,*

$$\sum_{h=1}^{H} L_h^{k-1}(f^*) \lesssim 8HB_f^2 \log(HK/\delta) + 16B_f^2 \log(|\mathcal{H}|).$$

Proof of Lemma F.4. To prove Lemma F.4, we define the random variable $W_{h,f}$ as

$$W_{h,f}^k = \left(Q_{h,f}(x_h^k, a_h^k, b_h^k) - r_h^k - V_{h+1,f^*}(x_{h+1}^k) \right)^2 - \left(Q_{h,f^*}(x_h^k, a_h^k, b_h^k) - r_h^k - V_{h+1,f^*}(x_{h+1}^k) \right)^2.$$

41

Using the Bellman equation for Q_{f^*}, i.e.,

$$Q_{h,f^*}(x_h^k, a_h^k, b_h^k) = r_h^k + \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k, b_h^k)}[V_{h+1,f^*}(x_{h+1})]$$

we can calculate that

$$\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k, b_h^k)}[W_{h,f}^k] = \left(Q_{h,f}(x_h^k, a_h^k, b_h^k) - Q_{h,f^*}(x_h^k, a_h^k, b_h^k)\right)^2. \tag{F.23}$$

Under the filtration $\{\mathcal{F}_{h,k}\}_{k=1}^K$ defined in the proof of Proposition C.8, i.e, (F.16), one can derive that

$$\mathbb{E}[W_{h,f}^k | \mathcal{F}_{h,k-1}] = \mathbb{E}\left[\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)}[W_{h,f^k}]\big| \mathcal{F}_{h,k-1}\right]$$

$$= \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right], \tag{F.24}$$

where $\xi_h = (x_h, a_h, b_h, r_h, x_{h+1})$, and that

$$\mathbb{V}[W_{h,f}^k | \mathcal{F}_{h,k-1}] \leq 4B_f^2 \mathbb{E}[X_{h,f}^k | \mathcal{F}_{h,k-1}] = 4B_f^2 \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right]. \tag{F.25}$$

By applying Lemma G.2, (F.24), and (F.25), we can obtain that with probability at least $1 - \delta$, for any $f \in \mathcal{H}$, $(h, k) \in [H] \times [K]$,

$$\left|\sum_{s=1}^{k-1} W_{h,f}^s - \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right]\right| \lesssim 4B_f^2 \log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta)$$

$$+ \sqrt{\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta) \cdot \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right]}.$$

Rearranging terms, we have that with probability at least $1 - \delta$, for any $(f_h, f_{h+1}) \in \mathcal{H} \times \mathcal{H}_{h+1}$, $(h, k) \in [H] \times [K]$,

$$-\sum_{s=1}^{k-1} W_{h,f}^s \lesssim 4B_f^2 \log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta) - \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right]$$

$$+ \sqrt{\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta) \cdot \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right]}$$

$$\lesssim 8B_f^2 \log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta),$$

where in the second inequality we use the fact that $-x^2 + ax \leq a^2/4$. Thus, with probability at least $1 - \delta$, for any $k \in [K]$, it holds that

$$\sum_{h=1}^H L_h^{k-1}(f^*) = \sum_{h=1}^H \left(\sum_{s=1}^{k-1} \left(Q_{h,f^*}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f^*}(x_{h+1}^s)\right)^2\right.$$

$$\left. - \inf_{f_h \in \mathcal{H}_h} \sum_{s=1}^{k-1} \left(Q_{h,f}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f^*}(x_{h+1}^s)\right)^2\right)$$

$$= \sum_{h=1}^H \sup_{f_h \in \mathcal{H}_h} \sum_{s=1}^{k-1} -W_{h,f}^s \lesssim 8HB_f^2 \log(HK/\delta) + 16B_f^2 \log(|\mathcal{H}|).$$

This finishes the proof of Lemma F.4. $\qquad\square$

Finally, combining (F.22) and Lemma F.4, we have, with probability at least $1 - \delta$, for any $f \in \mathcal{H}$, $k \in [K]$,

$$\sum_{h=1}^H L_h^{k-1}(f^*) - L_h^{k-1}(f) \lesssim -\frac{1}{2} \sum_{h=1}^H \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}[\ell_{f^s}(f; \xi_h)] + 16HB_f^2 \log(HK/\delta) + 32B_f^2 \log(|\mathcal{H}|).$$

42

This finishes the proof of the *first* inequality in Proposition C.8. In the following, we prove the *second* inequality in Proposition C.8. To this end, we define the following random variable, for any $f, g \in \mathcal{H}$ and policy μ_f,

$$X_{h,g,\mu_f}^k = \left(Q_{h,g}(x_h^k, a_h^k, b_h^k) - r_h^k - V_{h+1,g}^{\mu_f,\dagger}(x_{h+1}^k) \right)^2$$

$$- \left(V_{h+1,g}^{\mu_f,\dagger}(x_{h+1}^k) - \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k, b_h^k)}[V_{h+1,g}^{\mu_f,\dagger}(x_{h+1})] \right)^2. \quad \text{(F.26)}$$

After a calculation similar to (E.10) and (E.11), we can derive that

$$\mathbb{E}_{x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k, b_h^k)}[X_{h,g,\mu_f}^k] = \left(Q_{h,g}(x_h^k, a_h^k, b_h^k) - r_h^k - \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k, b_h^k)}[V_{h+1,g}^{\mu_f,\dagger}(x_{h+1})] \right)^2.$$

Following the same argument as in the previous proof of the *first* inequality of Proposition C.8 (see (F.17) and (F.18)), using the definition of ℓ_μ in (C.24), we can derive that, under filtration defined in (F.16),

$$\mathbb{E}[X_{h,f}^k | \mathcal{F}_{h,k-1}] = \mathbb{E}_{\xi_h \sim \pi^k}[\ell_{g^k,\mu_f}(g;\xi_h)], \quad \mathbb{V}[X_{h,f}^k | \mathcal{F}_{h,k-1}] \le 4B_f^2 \mathbb{E}_{\xi_h \sim \pi^k}[\ell_{g^k,\mu_f}(g;\xi_h)]. \quad \text{(F.27)}$$

Using (F.27) and Lemma G.2, we can obtain that with probability at least $1 - \delta$, for any $(h, k) \in [H] \times [K]$ and $(g_h, g_{h+1}, f_{h+1}) \in \mathcal{H}_h \times \mathcal{H}_{h+1} \times \mathcal{H}_{h+1}$[7],

$$\left| \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^s}[\ell_{g^s,\mu_f}(g;\xi_h)] - \sum_{s=1}^{k-1} X_{h,g,\mu_f}^s \right| \quad \text{(F.28)}$$

$$\lesssim \frac{1}{2} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^s}[\ell_{g^s,\mu_f}(g;\xi_h)] + 16B_f^2 \log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|/\delta). \quad \text{(F.29)}$$

Rearranging terms in (F.28), we can further obtain that

$$-\sum_{s=1}^{k-1} X_{h,g,\mu_f}^s \lesssim -\frac{1}{2} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi^s}[\ell_{g^s,\mu_f}(g;\xi_h)] + 16B_f^2 \log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|/\delta). \quad \text{(F.30)}$$

Meanwhile, by the definition of $X_{h,f}$ in (F.26) and the loss function L in (C.15), we have that

$$\sum_{s=1}^{k-1} X_{h,g,\mu_f}^s$$

$$= \sum_{s=1}^{k-1} \left(Q_{h,g}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,g}^{\mu_f,\dagger}(x_{h+1}^s) \right)^2 \quad \text{(F.31)}$$

$$- \sum_{s=1}^{k-1} \left(V_{h+1,g}^{\mu_f,\dagger}(x_{h+1}^s) - \mathbb{E}_{x_{h+1} \sim \mathbb{P}_h(\cdot|x_h^s, a_h^s, b_h^s)}[V_{h+1,g}^{\mu_f,\dagger}(x_{h+1})] \right)^2$$

$$= \sum_{s=1}^{k-1} \left(Q_{h,g}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,g}^{\mu_f,\dagger}(x_{h+1}^s) \right)^2$$

$$- \sum_{s=1}^{k-1} \left(\mathcal{T}_h^{\mu_f} g(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,g}^{\mu_f,\dagger}(x_{h+1}^s) \right)^2$$

$$\le \sum_{s=1}^{k-1} \left(Q_{h,g}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,g}^{\mu_f,\dagger}(x_{h+1}^s) \right)^2 \quad \text{(F.32)}$$

$$- \inf_{f_h' \in \mathcal{H}_h} \sum_{s=1}^{k-1} \left(Q_{h,f'}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,g}^{\mu_f,\dagger}(x_{h+1}^s) \right)^2$$

$$= L_{h,\mu_f}^{k-1}(f). \quad \text{(F.33)}$$

---

[7]Note that $\ell_{g^s,\mu_f}(g;\xi_h)$ and $V_{h+1,g}^{\mu_f,\dagger}$ depend on f only through f_{h+1}.

43

where the last inequality follows from the completeness assumption (Assumption C.4). Combining (F.30) and (F.31), we can derive that with probability at least $1 - \delta$, for any $f, g \in \mathcal{H}$, $k \in [K]$,

$$-\sum_{h=1}^{H} L_{h,\mu_f}^{k-1}(f) \lesssim -\frac{1}{2}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}[\ell_{g^s,\mu_f}(g;\xi_h)] + 16HB_f^2\log(HK/\delta) + 48B_f^2\log(|\mathcal{H}|).$$
(F.34)

Especially, we take $f = f^k$, we can obtain that with probability at least $1 - \delta$, for any $g \in \mathcal{H}$, $k \in [K]$,

$$-\sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(f) \lesssim -\frac{1}{2}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^s}[\ell_{g^s,\mu^k}(g;\xi_h)] + 16HB_f^2\log(HK|\mathcal{H}|/\delta) + 48B_f^2\log(|\mathcal{H}|).$$
(F.35)

Finally, we deal with the term $L_h^k(f^*)$. To this end, we invoke the following lemma.

Lemma F.5. *With probability at least $1 - \delta$, it holds that for each $k \in [K]$,*

$$\sum_{h=1}^{H} L_{h,\mu^k}^{k-1}(Q^{\mu^k,\dagger}) \lesssim 16HB_f^2\log(HK/\delta) + 48B_f^2\log(|\mathcal{H}|).$$

Proof of Lemma F.5. To prove Lemma F.5, we define the following random variable,

$$W_{h,g,\mu_f}^k = \left(Q_{h,g}(x_h^k, a_h^k, b_h^k) - r_h^k - V_{h+1}^{\mu_f,\dagger}(x_{h+1}^k)\right)^2 - \left(Q_h^{\mu_f,\dagger}(x_h^k, a_h^k, b_h^k) - r_h^k - V_{h+1}^{\mu_f,\dagger}(x_{h+1}^k)\right)^2,$$

for any $f, g \in \mathcal{H}$. Using the Bellman equation for $Q^{\mu_f,\dagger}$, i.e.,

$$Q_h^{\mu_f,\dagger}(x_h^k, a_h^k, b_h^k) = r_h^k + \mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x_h^k, a_h^k, b_h^k)}[V_{h+1}^{\mu_f,\dagger}(x_{h+1})]$$

we can calculate that

$$\mathbb{E}_{x_{h+1}^k\sim\mathbb{P}_h(\cdot|x_h^k, a_h^k, b_h^k)}[W_{h,g,\mu_f}^k] = \left(Q_{h,g}(x_h^k, a_h^k, b_h^k) - Q_h^{\mu_f,\dagger}(x_h^k, a_h^k, b_h^k)\right)^2.$$
(F.36)

Under the filtration $\{\mathcal{F}_{h,k}\}_{k=1}^K$ defined in the proof of Proposition C.8, i.e, (F.16), we can derive that

$$\mathbb{E}[W_{h,g,\mu_f}^k|\mathcal{F}_{h,k-1}] = \mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^k}\left[\left(Q_{h,g}(x_h, a_h, b_h) - Q_h^{\mu_f,\dagger}(x_h, a_h, b_h)\right)^2\right],$$
(F.37)

$$\mathbb{V}[W_{h,g,\mu_f}^k|\mathcal{F}_{h,k-1}] \leq 4B_f^2\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^k}\left[\left(Q_{h,g}(x_h, a_h, b_h) - Q_h^{\mu_f,\dagger}(x_h, a_h, b_h)\right)^2\right].$$
(F.38)

Using Lemma G.2, (F.37), (F.38), we have that, with probability at least $1 - \delta$, for any $(h, k) \in [H] \times [K]$, $(g_h, g_{h+1}, f_h, f_{h+1}) \in \mathcal{H}_h \times \mathcal{H}_{h+1} \times \mathcal{H}_h \times \mathcal{H}_{h+1}$,

$$\left|\sum_{s=1}^{k-1} W_{h,g,\mu_f}^s - \sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^k}\left[\left(Q_{h,g}(x_h, a_h, b_h) - Q_h^{\mu_f,\dagger}(x_h, a_h, b_h)\right)^2\right]\right| \lesssim 8B_f^2\log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|^2/\delta)$$

$$+ \sqrt{\log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|^2/\delta) \cdot \sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}\left[\left(Q_{h,g}(x_h, a_h, b_h) - Q_h^{\mu_f,\dagger}(x_h, a_h, b_h)\right)^2\right]}.$$

Rearranging terms, we have that with probability at least $1 - \delta$,

$$-\sum_{s=1}^{k-1} W_{h,g,\mu_f}^s \lesssim 8B_f^2\log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|^2/\delta) - \sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}\left[\left(Q_{h,g}(x_h, a_h, b_h) - Q_h^{\mu_f,\dagger}(x_h, a_h, b_h)\right)^2\right]$$

$$+ \sqrt{\log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|^2/\delta) \cdot \sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}\left[\left(Q_{h,g}(x_h, a_h, b_h) - Q_h^{\mu_f,\dagger}(x_h, a_h, b_h)\right)^2\right]}$$

$$\lesssim 16B_f^2\log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|^2/\delta),$$

44

where in the second inequality we use the fact that $-x^2 + ax \le a^2/4$. Now we take $f = f^k$, which gives that with probability at least $1 - \delta$, for any $k \in [K]$, it holds that

$$
\sum_{h=1}^H L_{h,\mu^k}^{k-1}(Q^{\mu^k,\dagger}) = \sum_{h=1}^H \left(\sum_{s=1}^{k-1} \left(\underbrace{Q_{h,Q^{\mu^k,\dagger}}(x_h^s, a_h^s, b_h^s)}_{=Q_h^{\mu^k,\dagger}(x_h^s, a_h^s, b_h^s)} - r_h^s - \underbrace{V_{h+1,Q^{\mu^k,\dagger}}^{\mu^k,\dagger}(x_{h+1}^s)}_{=V_{h+1}^{\mu^k,\dagger}(x_{h+1}^s)} \right)^2 \right.
$$

$$
\left. - \inf_{g_h \in \mathcal{H}_h} \sum_{s=1}^{k-1} \left(Q_{h,g}(x_h^s, a_h^s, b_h^s) - r_h^s - \underbrace{V_{h+1,Q^{\mu^k,\dagger}}^{\mu^k,\dagger}(x_{h+1}^s)}_{=V_{h+1}^{\mu^k,\dagger}(x_{h+1}^s)} \right)^2 \right)
$$

$$
= \sum_{h=1}^H \sup_{g_h \in \mathcal{H}_h} \sum_{s=1}^{k-1} -W_{h,g,\mu^k}^s \lesssim 16HB_f^2 \log(HK/\delta) + 64B_f^2 \log(|\mathcal{H}|).
$$

This finishes the proof of Lemma F.5. $\qquad\square$

Finally, combining (F.35) and Lemma F.5, we have, with probability at least $1 - \delta$, for any $g \in \mathcal{H}$, $k \in [K]$,

$$
\sum_{h=1}^H L_{h,\mu^k}^{k-1}(Q^{\mu^k,\dagger}) - L_{h,\mu^k}^{k-1}(g)
$$

$$
\lesssim -\frac{1}{2} \sum_{h=1}^H \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}s}[\ell_{g^s,\mu^k}(g; \xi_h)] + 32HB_f^2 \log(HK|\mathcal{H}|/\delta) + 112B_f^2 \log(|\mathcal{H}|).
$$

This finishes the proof of the *second* inequality in Proposition C.8 and completes the proof of Proposition C.8. $\qquad\square$

G   Technical Lemmas

Lemma G.1 (Martingale exponential inequality). *For a sequence of real-valued random variables $\{X_t\}_{t \le T}$ adapted to a filtration $\{\mathcal{F}_t\}_{t \le T}$, the following holds with probability at least $1 - \delta$, for any $t \in [T]$,*

$$
-\sum_{s=1}^t X_s \le \sum_{s=1}^t \log \mathbb{E}[\exp(-X_s)|\mathcal{F}_{s-1}] + \log(1/\delta).
$$

Proof of Lemma G.1. See e.g., Theorem 13.2 of Zhang [88] for a detailed proof. $\qquad\square$

Lemma G.2 (Freedman's inequality). *Let $\{X_t\}_{t \le T}$ be a real-valued martingale difference sequence adapted to filtration $\{\mathcal{F}_t\}_{t \le T}$. If $|X_t| \le R$ almost surely, then for any $\eta \in (0, 1/R)$ it holds that with probability at least $1 - \delta$,*

$$
\sum_{t=1}^T X_t \le \mathcal{O}\left(\eta \sum_{t=1}^T \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] + \frac{\log(1/\delta)}{\eta} \right).
$$

Proof of Lemma G.2. See Freedman [26] for detailed proof. $\qquad\square$

Lemma G.3 (Elliptical potential). *Let $\{x_i\}_{i \in [K]}$ be a sequence of vectors with $x_i \in \mathcal{V}$ for some Hilbert space $\mathcal{V}$. Let Λ_0 be a positive definite matrix and define $\Lambda_k = \Lambda_0 + \sum_{i=1}^k x_i x_i^\top$. Then it holds that*

$$
\sum_{i=1}^K \min\left\{ 1, \|x_i\|_{\Lambda_i^{-1}} \right\} \le 2 \log\left(\frac{\det(\Lambda_{K+1})}{\det(\Lambda_1)} \right).
$$

Proof of Lemma G.3. See Lemma 11 of Abbasi-Yadkori et al. [1] for a detailed proof. $\qquad\square$

45

H Experimental Settings

Our experiments utilize 8 NVIDIA GeForce 1080Ti GPUs and 4 NVIDIA A6000 GPUs. Each result is averaged over five random seeds.

H.1 Environment Setup

In sparse-reward tasks, the agent only receives a reward when it achieves the desired velocity or position. Regarding the model-based sparse-reward experiments, we assign a target value of 1 to the `vel` parameter for the `walker-vel` task and 1.5 for the `hopper-vel`, `cheetah-vel`, `ant-vel` tasks. For the model-free sparse-reward experiments, we set the target `vel` to 3 for the `hopper-vel`, `walker-vel`, `cheetah-vel` tasks, and the target `goal` to $(2, 0)$ for the `ant-goal` task.

H.2 Implementation Details of `MEX-MF`

Below, we describe the detailed implementation of the model-free algorithm `MEX-MF`. We adopt a similar entropy regularization $\mathcal{H}(\mu)$ over μ as in `CQL` [42]. By incorporating such a regularization, we obtain the following soft constrained variant of `MEX-MF`, i.e.

$$\max_{\theta} \max_{\pi} -\mathbb{E}_{\beta}\left[\left(r + \gamma Q_{\theta}(x', a') - Q_{\theta}(x, a)\right)^2\right] + \eta' \cdot \mathbb{E}_{\beta}\left[\mathbb{E}_{a \sim \pi} Q_{\theta}(x, a) - \log \sum_{a \in \mathcal{A}} \exp\left(Q_{\theta}(x, a)\right)\right].$$

We select η' to be $1e-3$ for sparse-reward tasks and $5e-4$ for standard gym tasks since dense reward tasks require less exploration. Other parameters are kept the same with the baseline [27] across all domains and are summarized as in Table 1.

H.3 Implementation Details of `MEX-MB`

When employing the model-based algorithm `MEX-MB`, we configured the parameter η' as $1e-4$ for the `Hopper-v2` and `hopper-vel` tasks, and $1e-3$ for all other tasks. The hyper-parameters are kept the same with the MBPO baseline [34] across all domains and are summarized as in Table 2.

Hyperparameter	Value
Optimizer	Adam
Critic learning rate	3e-4
Actor learning rate	3e-4
Mini-batch size	256
Discount factor	0.99
Target update rate	5e-3
Policy noise	0.2
Policy noise clipping	(-0.5, 0.5)
TD3+BC parameter α	2.5

Architecture	Value
Critic hidden dim	256
Critic hidden layers	2
Critic activation function	ReLU
Actor hidden dim	256
Actor hidden layers	2
Actor activation function	ReLU

Table 1: Hyper-parameters sheet of `MEX-MF`.

Hyperparameter	Value
Optimizer	Adam
Critic learning rate	3e-4
Actor learning rate	3e-4
Model learning rate	1e-3
Mini-batch size	256
Discount factor	0.99
Target update rate	5e-3
SAC updates per step	40

Architecture	Value
Critic hidden layers	3
Critic activation function	ReLU
Actor hidden layers	2
Actor activation function	ReLU
Model hidden dim	200
Model hidden layers	4
Model activation function	SiLU

Table 2: Hyper-parameters sheet of `MEX-MB`.

H.4 Tabular Experiments

We also conduct experiments in tabular MDPs. Specifically, we evaluate `MEX-MB` and `MnM` [21] in a 10x10 gridworld with stochastic dynamics and sparse reward functions. As illustrated in Figure 3,

46

the stochastic gridworld environment is associated with a navigation task to reach the red star from the initial upper left cell position. The action space contains four discrete actions, corresponding to moving to the four adjacent cells. The transition noise moves the agent to neighbor states with equal probability. The black region represents the obstacle that the agent cannot enter. The agent receives a +0.001 reward at every timestep and a +10 when reaching the goal state. Each episode has 200 timesteps. The performance results are shown in Figure 4.
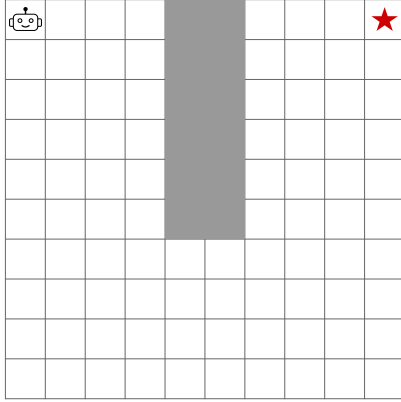
Figure 3: Illustration of the stochastic griworld environment [21].
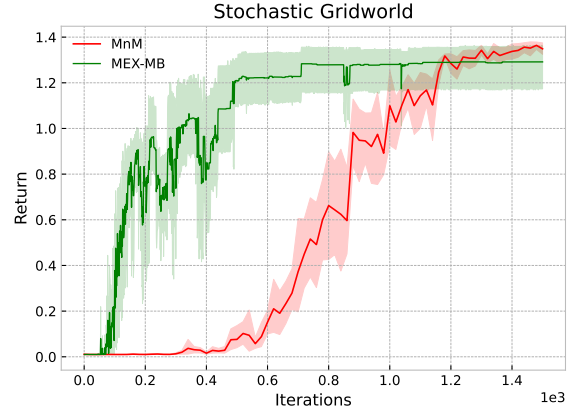
Figure 4: Model-based MEX-MB in the stochastic gridworld environment.

47