Contents

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.

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 $x_{h+1}^k \sim \mathbb{P}_h(\cdot \,|\, x_h^k, a_h^k, b_h^k)$. 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

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

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

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

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

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

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

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]

$$\begin{aligned} Q_h^{\mu,\dagger}(x,a,b) &= (\mathcal{T}_h^\mu Q_{h+1}^{\mu,\dagger})(x,a,b) \\ &:= r_h(x,a,b) + \mathbb{E}_{x'\sim\mathbb{P}_h(\cdot|x,a,b)}\left[\inf_{\nu_{h+1}} \mathbb{D}_{(\mu_{h+1},\nu_{h+1})} Q_{h+1}^{\mu,\dagger}(x')\right], \end{aligned} \tag{C.7}$$

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,

$$\begin{aligned} Q_h^*(x,a,b) &= (\mathcal{T}_h^{\mathrm{NE}} Q_{h+1}^*)(x,a,b) \\ &:= r_h(x,a,b) + \mathbb{E}_{x'\sim\mathbb{P}_h(\cdot|x,a,b)}\left[\sup_{\mu_{h+1}}\inf_{\nu_{h+1}} \mathbb{D}_{(\mu_{h+1},\nu_{h+1})} Q_{h+1}^*(x')\right], \end{aligned} \tag{C.9}$$

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

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

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.

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,

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

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

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

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

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

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

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

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

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

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

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,

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

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

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

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

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,

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

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 $(x, a, b, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$.*

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*

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

and it also holds that

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

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. □

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

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

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

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*

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

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

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

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*

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

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 $\|\phi_h(x,a,b)\|_2 \le 1$ and $\max\{\|\alpha_h\|_2, \|\psi_h^\star(x')\|_2\} \le \sqrt{d}$ for any $(x,a,b,x',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times\mathcal{S}\times[H]$.

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

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

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

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*

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

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

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

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*

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

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*

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

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

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

with probability at least $1-\delta$.

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*

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

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,

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

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*

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

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. □

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. □

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,

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

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 $k \in [K]$,

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

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

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

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

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

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 $\epsilon > 0$,

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

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

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

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

$$\begin{aligned}
\text{Regret}(T) &= \text{Term (i)} + \text{Term (ii)} \\
&\le -c_{\text{(i)}} \cdot \eta \sum_{k=1}^{K} \sum_{s=1}^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f^k;\xi_h)] + c_{\text{(i)}} \cdot \eta BK\big(H \log(HK/\delta) + \log(|\mathcal{H}|)\big), \\
&\quad + c_{\text{(i)}} \cdot \eta \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \pi_{\exp}(f^s)}[\ell_{f^s}(f^k;\xi_h)] + \frac{d_{\text{GEC}}(\epsilon)}{4c_{\text{(i)}}\eta} + \sqrt{d_{\text{GEC}}(\epsilon) HK} + \epsilon HK \\
&= c_{\text{(i)}} \cdot \eta BK\big(H \log(HK/\delta) + \log(|\mathcal{H}|)\big) + \frac{d_{\text{GEC}}(\epsilon)}{4c_{\text{(i)}}\eta} + \sqrt{d_{\text{GEC}}(\epsilon) HK} + \epsilon HK.
\end{aligned} \tag{D.7}$$

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. □

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 $k \in [K]$,

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

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

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

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

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

$$\le -c_{(\mathrm{max.i})} \cdot \eta \sum_{k=1}^{K} \sum_{s=1}^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}[\ell_{f^s}(f; \xi_h)] + c_{(\mathrm{max.i})} \cdot BK\big(H \log(HK/\delta) + \log(|\mathcal{H}|)\big), \tag{D.14}$$

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 $\epsilon > 0$,

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

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

$$\text{Term (Max.ii)} \le c_{(\mathrm{max.i})} \cdot \eta \sum_{h=1}^{H} \sum_{k=1}^{K} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \boldsymbol{\pi}^k}[\ell_{f^s}(f^k; \xi_h)] + \frac{d_{\mathrm{TGEC}}(\epsilon)}{4 c_{(\mathrm{max.i})} \eta} + \sqrt{d_{\mathrm{TGEC}}(\epsilon) HK} + \epsilon HK. \tag{D.15}$$

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 $k \in [K]$,

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

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

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

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

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

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

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

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

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

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. □

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.

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 $\sqrt{\sum_{i=1}^n (\mathbb{E}_{\mu_i}[g])^2} \le \epsilon$ 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 $\epsilon' \ge \epsilon$ 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*

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

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*

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

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

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

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*

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

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

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

Proof of Lemma E.5. See Lemma 3.16 in [89] for a detailed proof. □

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.

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*

$$\begin{aligned}\left|\mathbb{E}_{\pi_f}[Q_{h,f}(x_h, a_h) - R_h(x_h, a_h) - V_{h+1,f}(x_{h+1})]\right| &\leq \left|\langle W_h(f) - W_h(f^*), X_h(f)\rangle_{\mathcal{V}}\right|, \\ \left|\mathbb{E}_{x_h \sim \pi_f, a_h \sim \widetilde{\pi}}[l_f(g; \xi_h)]\right| &= \left|\langle W_h(g) - W_h(f^*), X_h(f)\rangle_{\mathcal{V}}\right|,\end{aligned}$$

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 $\pi_{\mathrm{exp}}(f) = \pi_f$ for Q-type problems and $\pi_{\mathrm{exp}}(f) = \pi_f \circ_h \mathrm{Unif}(\mathcal{A})$ 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 $\pi_{\mathrm{exp}}(f) = \pi_f$, we have that

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

For V-type problems, by choosing $\pi_{\mathrm{exp}}(f) = \pi_f \circ_h \mathrm{Unif}(\mathcal{A})$, 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 $\mathcal{V} = \{\upsilon : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]\}$.

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,*

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

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

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 $\pi_{\mathrm{exp}}(f) = \pi_f$, we have that*

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

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

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

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

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 $1-\delta$,

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

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

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

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

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

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

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

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),

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

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

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

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

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

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

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

and that

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

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],

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

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

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

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

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

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

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

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

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

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

[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}).

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

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

and that

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

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 $(f_h,f_{h+1})\in\mathcal{H}_h\times\mathcal{H}_{h+1}$,

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

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

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

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

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

This finishes the proof of Lemma E.10. □

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

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

This finishes the proof of Proposition 5.1. □

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 $f_h \in \mathcal{H}_h$,

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

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

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{-\frac{1}{2}X_{h,f}^s\right\}\middle|\mathcal{F}_{s-1}\right] \\
&= \mathbb{E}\left[\sqrt{\frac{\mathbb{P}_{h,f}(x_{h+1}^s|x_h^s,a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s|x_h^s,a_h^s)}}\middle|\mathcal{F}_{s-1}\right] \\
&= \mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s),x_{h+1}^s\sim\mathbb{P}_{h,f^*}(\cdot|x_h^s,a_h^s)}\left[\sqrt{\frac{\mathbb{P}_{h,f}(x_{h+1}^s|x_h^s,a_h^s)}{\mathbb{P}_{h,f^*}(x_{h+1}^s|x_h^s,a_h^s)}}\right] \\
&= \mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\left[\int_{\mathcal{S}}\sqrt{\mathbb{P}_{h,f}(x_{h+1}^s|x_h^s,a_h^s)\cdot\mathbb{P}_{h,f^*}(x_{h+1}^s|x_h^s,a_h^s)}\mathrm{d}x_{h+1}^s\right] \\
&= 1 - \frac{1}{2}\mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\left[\int_{\mathcal{S}}\left(\sqrt{\mathbb{P}_{h,f}(x_{h+1}^s|x_h^s,a_h^s)} - \sqrt{\mathbb{P}_{h,f^*}(x_{h+1}^s|x_h^s,a_h^s)}\right)^2\mathrm{d}x_{h+1}^s\right] \\
&= 1 - \mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\Big[D_{\mathrm{H}}(\mathbb{P}_{h,f^*}(\cdot|x_h^s,a_h^s)\|\mathbb{P}_{h,f}(\cdot|x_h^s,a_h^s))\Big],
\end{aligned} \tag{E.28}$$

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

$$\begin{aligned}
-\frac{1}{2}\sum_{s=1}^{k-1} X_{h,f}^s &\le \sum_{s=1}^{k-1}\mathbb{E}\left[\exp\left\{-\frac{1}{2}X_{h,f}^s\right\}\middle|\mathcal{F}_{s-1}\right] - 1 + \log(H|\mathcal{H}_h|/\delta) \\
&= -\sum_{s=1}^{k-1}\mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\left[D_{\mathrm{H}}(\mathbb{P}_{h,f^*}(\cdot|x_h^s,a_h^s)\|\mathbb{P}_{h,f}(\cdot|x_h^s,a_h^s))\right] + \log(H|\mathcal{H}_h|/\delta),
\end{aligned}$$

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

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

$$\begin{aligned}
\sum_{h=1}^{H} L_h^{k-1}(f^*) - L_h^{k-1}(f) &= -\sum_{h=1}^{H}\sum_{s=1}^{k-1} X_{h,f}^s \\
&\le -2\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{(x_h^s,a_h^s)\sim\pi_{\exp}(f^s)}\left[D_{\mathrm{H}}(\mathbb{P}_{h,f^*}(\cdot|x_h^s,a_h^s)\|\mathbb{P}_{h,f}(\cdot|x_h^s,a_h^s))\right] \\
&\quad + 2H\log(H/\delta) + 2\log(|\mathcal{H}|), \\
&= -2\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h\sim\pi_{\exp}(f^s)}[\ell_{f^s}(f;\xi_h)] + 2H\log(H/\delta) + 2\log(|\mathcal{H}|). \qquad \text{(E.29)}
\end{aligned}$$

This finishes the proof of Proposition 5.3. □

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*

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

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*

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

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*

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

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

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

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

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \left|\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^k}\left[\mathcal{E}_h^{\mu}(g_h^k, g_{h+1}^k;\xi_h)\right]\right| \\
&\qquad \le \frac{\widetilde{d}(\epsilon)}{4\eta} + \frac{\eta}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\boldsymbol{\pi}^s}[\ell_{g^s,\mu}(g^k;\xi_h)] + 2\min\{HK, 2\widetilde{d}(\epsilon)\} + HKB_W\epsilon,
\end{aligned} \tag{F.7}$$

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,

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

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,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \left|\mathbb{E}_{\boldsymbol{\pi}^k}\left[\mathcal{E}_h^{\mu}(g_h^k, g_{h+1}^k;\xi_h)\right]\right| \\
&\quad = \sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \langle W_h(g^k,\mu), X_h(g^k,\boldsymbol{\pi}^k,\mu)\rangle\}\mathbf{1}\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}} \le 1\right\} \\
&\qquad + \sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1, \langle W_h(g^k,\mu), X_h(g^k,\boldsymbol{\pi}^k,\mu)\rangle\}\mathbf{1}\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}} > 1\right\} \\
&\quad \le \sum_{k=1}^{K}\sum_{h=1}^{H}\langle W_h(g^k,\mu), X_h(g^k,\boldsymbol{\pi}^k,\mu)\rangle\mathbf{1}\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}} \le 1\right\} + \min\{HK, \widetilde{d}(\epsilon)\} \\
&\quad \le \sum_{k=1}^{K}\sum_{h=1}^{H}\underbrace{\|W_h(g^k,\mu)\|_{\Sigma_{h,k}}\min\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}}, 1\right\}}_{(\mathrm{A})_{h,k}} + \min\{HK, \widetilde{d}(\epsilon)\},
\end{aligned} \tag{F.9}$$

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 $(\mathrm{A})_{h,k}$ in (F.9) as follows.

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

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

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}(\mathrm{A})_{h,k} \qquad\qquad (\mathrm{F.10})\\
&\quad\le \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\sqrt{\epsilon}B_W + \left[\sum_{s=1}^{k-1}|\langle W_h(g^k,\mu), X_h(g^s,\boldsymbol{\pi}^s,\mu)\rangle|^2\right]^{1/2}\right)\cdot \min\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}},1\right\}\\
&\quad\le \left[\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\epsilon}B_W\right]^{1/2}\cdot\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}},1\right\}\right]^{1/2}\\
&\qquad+\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}|\langle W_h(g^k,\mu), X_h(g^s,\boldsymbol{\pi}^s,\mu)\rangle|^2\right]^{1/2}\cdot\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{\|X_h(g^k,\boldsymbol{\pi}^k,\mu)\|_{\Sigma_{h,k}^{-1}},1\right\}\right]^{1/2}\\
&\quad\le \sqrt{HB_WK\epsilon\cdot\min\{2\widetilde{d}(\epsilon),HK\}}+\left[2\widetilde{d}(\epsilon)\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}|\langle W_h(g^k,\mu), X_h(g^s,\boldsymbol{\pi}^s,\mu)\rangle|^2\right]^{1/2}\\
&\quad\le \sqrt{HKB_W\epsilon\cdot\min\{2\widetilde{d}(\epsilon),HK\}}+\left[2\widetilde{d}(\epsilon)\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}[\ell_{g^s,\mu}(g^k;\xi_h)]\right]^{1/2},
\end{aligned}
$$

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

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

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

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

$$
\begin{aligned}
&Q_{h,g}(x,a,b)-r_h(x,a,b)-(\mathbb{P}_hV_{h+1,g}^{\mu,\dagger})(x,a,b)\\
&\qquad=\phi_h(x,a,b)^{\top}\left(\theta_{h,g}-\alpha_h-\int_{\mathcal{S}}\psi_h^{\star}(x')V_{h+1,g}^{\mu,\dagger}(x')\mathrm{d}x'\right),
\end{aligned}
$$

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

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

$$\begin{aligned}\sum_{k=1}^{K} V_1^{\boldsymbol{\pi}^k}(x_1) - V_{1,g^k}^{\mu^k,\dagger}(x_1) &\le \sum_{k=1}^{K}\sum_{h=1}^{H} \left|\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^k}\left[\mathcal{E}_h^{\mu}(g_h^k, g_{h+1}^k;\xi_h)\right]\right| \\ &\le \frac{\widetilde{d}(\epsilon)}{4\eta} + \frac{\eta}{2}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}[\ell_{g^s,\mu}(g^k;\xi_h)] \\ &\quad + 2\min\{HK, 2\widetilde{d}(\epsilon)\} + 4\sqrt{d}H^2K\epsilon,\end{aligned}$$

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

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

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

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

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

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

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

$$\begin{aligned}Q_{h,g}^{\mu,\dagger}(x,a,b) - r_h - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x,a,b) &= (\mathbb{P}_{h,g}V_{h+1,g}^{\mu,\dagger})(x,a,b) - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x,a,b) \\ &= (\theta_{h,g} - \theta_h^\star)^\top\left(\int_{\mathcal{S}}\phi_h(x,a,b,x')V_{h+1,g}^{\mu,\dagger}(x')\mathrm{d}x'\right),\end{aligned} \tag{F.11}$$

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

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

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 $B_X \le 1$ and $B_W \le 4H\sqrt{d}$. 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

$$\begin{aligned}\left(Q_{h,g}^{\mu,\dagger}(x,a,b) - r_h - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x,a,b)\right)^2 &= \left((\mathbb{P}_{h,g}V_{h+1,g}^{\mu,\dagger})(x,a,b) - (\mathbb{P}_h V_{h+1,g}^{\mu,\dagger})(x,a,b)\right)^2 \\ &\le 4\|V_{h+1,g}^{\mu,\dagger}(\cdot)\|_\infty^2 \cdot D_{\mathrm{TV}}(\mathbb{P}_{h,g}(\cdot\,|\,x,a,b)\|\mathbb{P}_h(\cdot\,|\,x,a,b))^2 \\ &\le 2H^2 D_{\mathrm{H}}(\mathbb{P}_{h,g}(\cdot\,|\,x,a,b)\|\mathbb{P}_h(\cdot\,|\,x,a,b))^2 \\ &\le 2H^2 D_{\mathrm{H}}(\mathbb{P}_{h,g}(\cdot\,|\,x,a,b)\|\mathbb{P}_h(\cdot\,|\,x,a,b)),\end{aligned} \tag{F.13}$$

where the second equality comes from Holder inequality and the fact that the TV distance $D_{\mathrm{TV}}(p\|q) = \|p-q\|_1/2$ for any two distributions p and q, the third inequality follows from the fact that $D_{\mathrm{TV}}(p\|q) \le \sqrt{2} D_{\mathrm{H}}(p\|q)$, and the last inequality follows from the fact that $D_{\mathrm{H}}(p\|q) \le 1$. 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,

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

Thus by applying Proposition F.3, we have that

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

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

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

This proves that $d_{\mathrm{TGEC}}(\epsilon) \le \bar{d}(\epsilon)$. 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. □

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 $X^k_{h,f}$ as

$$\begin{aligned}
X^k_{h,f} &= \left(Q_{h,f}(x^k_h,a^k_h,b^k_h) - r^k_h - V_{h+1,f}(x^k_{h+1})\right)^2 \\
&\quad - \left(V_{h+1,f}(x^k_{h+1}) - \mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x^k_h,a^k_h,b^k_h)}[V_{h+1,f}(x_{h+1})]\right)^2. \qquad \text{(F.15)}
\end{aligned}$$

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

$$\mathbb{E}_{x^k_{h+1}\sim\mathbb{P}_h(\cdot|x^k_h,a^k_h,b^k_h)}[X^k_{h,f}] = \left(Q_{h,f}(x^k_h,a^k_h,b^k_h) - r^k_h - \mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x^k_h,a^k_h,b^k_h)}[V_{h+1,f}(x_{h+1})]\right)^2.$$

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

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

where $\mathcal{D}^s_h = \{x^s_h, a^s_h, b^s_h, r^s_h, x^s_{h+1}\}$. With previous arguments, we can derive that

$$\mathbb{E}[X^k_{h,f}|\mathcal{F}_{h,k-1}] = \mathbb{E}\left[\mathbb{E}_{x^k_{h+1}\sim\mathbb{P}_h(\cdot|x^k_h,a^k_h,b^k_h)}[X_{h,f^k}]|\mathcal{F}_{h,k-1}\right] = \mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^k}[\ell_{f^k}(f;\xi_h)], \qquad \text{(F.17)}$$

and that

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

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 $(f_h, f_{h+1})\in\mathcal{H}_h\times\mathcal{H}_{h+1}$,

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

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

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

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

$$\begin{aligned}
&\sum_{s=1}^{k-1}X_{h,f}^s \\
&\quad= \sum_{s=1}^{k-1}\big(Q_{h,f}(x_h^s,a_h^s,b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2 \\
&\qquad- \sum_{s=1}^{k-1}\Big(V_{h+1,f}(x_{h+1}^s) - \mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x_h^s,a_h^s,b_h^s)}[V_{h+1,f}(x_{h+1})]\Big)^2 \\
&\quad= \sum_{s=1}^{k-1}\big(Q_{h,f}(x_h^s,a_h^s,b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2 \\
&\qquad- \sum_{s=1}^{k-1}\big(\mathcal{T}_h f(x_h^s,a_h^s,b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2 \\
&\quad\le \sum_{s=1}^{k-1}\big(Q_{h,f}(x_h^s,a_h^s,b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2 \\
&\qquad- \inf_{f_h'\in\mathcal{H}_h}\sum_{s=1}^{k-1}\big(Q_{h,f'}(x_h^s,a_h^s,b_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)\big)^2 \\
&\quad= L_h^{k-1}(f).
\end{aligned} \tag{F.21}$$

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 $f\in\mathcal{H}$, $k\in[K]$,

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

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

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

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

$$\begin{aligned}\mathbb{E}[W_{h,f}^k|\mathcal{F}_{h,k-1}] &= \mathbb{E}\left[\mathbb{E}_{x_{h+1}^k\sim\mathbb{P}_h(\cdot|x_h^k,a_h^k)}[W_{h,f^k}]|\mathcal{F}_{h,k-1}\right] \\ &= \mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^k}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right],\end{aligned} \tag{F.24}$$

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

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

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

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

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

$$\begin{aligned}-\sum_{s=1}^{k-1} W_{h,f}^s &\lesssim 4B_f^2\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta) - \sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right] \\ &\quad + \sqrt{\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta)\cdot\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}\left[(Q_{h,f}(x_h, a_h, b_h) - Q_{h,f^*}(x_h, a_h, b_h))^2\right]} \\ &\lesssim 8B_f^2\log(HK|\mathcal{H}_h||\mathcal{H}_{h+1}|/\delta),\end{aligned}$$

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

$$\begin{aligned}\sum_{h=1}^H L_h^{k-1}(f^*) &= \sum_{h=1}^H\left(\sum_{s=1}^{k-1}\left(Q_{h,f^*}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f^*}(x_{h+1}^s)\right)^2\right. \\ &\qquad \left. - \inf_{f_h\in\mathcal{H}_h}\sum_{s=1}^{k-1}\left(Q_{h,f}(x_h^s, a_h^s, b_h^s) - r_h^s - V_{h+1,f^*}(x_{h+1}^s)\right)^2\right) \\ &= \sum_{h=1}^H\sup_{f_h\in\mathcal{H}_h}\sum_{s=1}^{k-1} -W_{h,f}^s \lesssim 8HB_f^2\log(HK/\delta) + 16B_f^2\log(|\mathcal{H}|).\end{aligned}$$

This finishes the proof of Lemma F.4. □

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

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

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,

$$X^k_{h,g,\mu_f} = \Big(Q_{h,g}(x^k_h, a^k_h, b^k_h) - r^k_h - V^{\mu_f,\dagger}_{h+1,g}(x^k_{h+1})\Big)^2 - \Big(V^{\mu_f,\dagger}_{h+1,g}(x^k_{h+1}) - \mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x^k_h,a^k_h,b^k_h)}[V^{\mu_f,\dagger}_{h+1,g}(x_{h+1})]\Big)^2. \tag{F.26}$$

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

$$\mathbb{E}_{x^k_{h+1}\sim\mathbb{P}_h(\cdot|x^k_h,a^k_h,b^k_h)}[X^k_{h,g,\mu_f}] = \Big(Q_{h,g}(x^k_h, a^k_h, b^k_h) - r^k_h - \mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x^k_h,a^k_h,b^k_h)}[V^{\mu_f,\dagger}_{h+1,g}(x_{h+1})]\Big)^2.$$

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),

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

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],

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

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

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

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

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

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

$$= \sum_{s=1}^{k-1}\Big(Q_{h,g}(x^s_h, a^s_h, b^s_h) - r^s_h - V^{\mu_f,\dagger}_{h+1,g}(x^s_{h+1})\Big)^2 \tag{F.31}$$

$$- \sum_{s=1}^{k-1}\Big(V^{\mu_f,\dagger}_{h+1,g}(x^s_{h+1}) - \mathbb{E}_{x_{h+1}\sim\mathbb{P}_h(\cdot|x^s_h,a^s_h,b^s_h)}[V^{\mu_f,\dagger}_{h+1,g}(x_{h+1})]\Big)^2$$

$$= \sum_{s=1}^{k-1}\Big(Q_{h,g}(x^s_h, a^s_h, b^s_h) - r^s_h - V^{\mu_f,\dagger}_{h+1,g}(x^s_{h+1})\Big)^2$$

$$- \sum_{s=1}^{k-1}\Big(\mathcal{T}^{\mu_f}_h g(x^s_h, a^s_h, b^s_h) - r^s_h - V^{\mu_f,\dagger}_{h+1,g}(x^s_{h+1})\Big)^2$$

$$\le \sum_{s=1}^{k-1}\Big(Q_{h,g}(x^s_h, a^s_h, b^s_h) - r^s_h - V^{\mu_f,\dagger}_{h+1,g}(x^s_{h+1})\Big)^2 \tag{F.32}$$

$$- \inf_{f'_h\in\mathcal{H}_h}\sum_{s=1}^{k-1}\Big(Q_{h,f'}(x^s_h, a^s_h, b^s_h) - r^s_h - V^{\mu_f,\dagger}_{h+1,g}(x^s_{h+1})\Big)^2$$

$$= L^{k-1}_{h,\mu_f}(f). \tag{F.33}$$

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

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 $f, g \in \mathcal{H}$, $k \in [K]$,

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

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

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

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,

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

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

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

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

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

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

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

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

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

$$\begin{aligned}-\sum_{s=1}^{k-1} W_{h,g,\mu_f}^s &\lesssim 8B_f^2\log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|^2/\delta) - \sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}\left[\Big(Q_{h,g}(x_h, a_h, b_h) - Q_h^{\mu_f,\dagger}(x_h, a_h, b_h)\Big)^2\right] \\ &+ \sqrt{\log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|^2/\delta)\cdot\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\boldsymbol{\pi}^s}\left[\Big(Q_{h,g}(x_h, a_h, b_h) - Q_h^{\mu_f,\dagger}(x_h, a_h, b_h)\Big)^2\right]} \\ &\lesssim 16B_f^2\log(HK|\mathcal{H}_h|^2|\mathcal{H}_{h+1}|^2/\delta),\end{aligned}$$

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. □

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. □

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. □

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. □

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. □

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,

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.

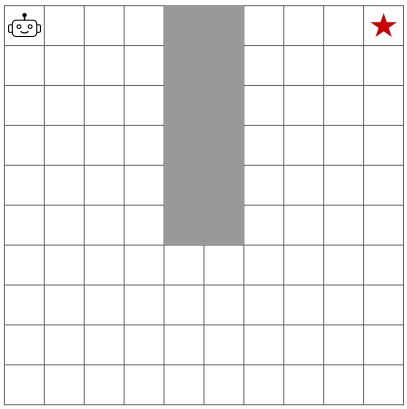

Figure 3: Illustration of the stochastic griworld environment [21].

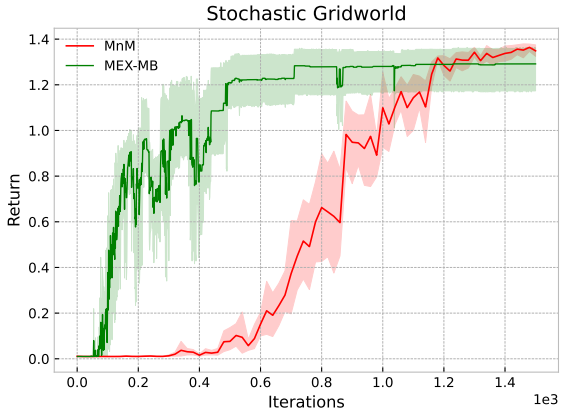

Figure 4: Model-based `MEX-MB` in the stochastic grid-world environment.