A BACKGROUNDS AND TECHNICAL NOVELTIES

Notations. For any integer $n \in \mathbb{N}^+$, we take the convention to use $[n] = \{1, \ldots, n\}$. Consider two non-negative sequences $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$, if $\limsup a_n/b_n < \infty$, then we write it as $a_n = \mathcal{O}(b_n)$. Else if $\limsup a_n/b_n = 0$, then we write it as $a_n = o(b_n)$. And we use $\tilde{\mathcal{O}}$ to omit the logarithmic terms. Denote $\Delta(\mathcal{X})$ be the probability simplex over the set $\mathcal{X}$. Denote by $\sup_x |v(x)|$ the supremum norm of a given function. $x \wedge y$ stands for $\min\{x, y\}$ and $x \vee y$ stands for $\max\{x, y\}$. Given any continuum $\mathcal{S}$, let $|\mathcal{S}|$ be the cardinality. Given two distributions $P, Q \in \Delta(\mathcal{X})$, the TV distance of the two distributions is defined as $\mathrm{TV}(P, Q) = \frac{1}{2}\mathbb{E}_{x\sim P}[|\mathrm{d}Q(x)/\mathrm{d}P(x) - 1|]$.

Infinite-horizon Average-reward MDPs. Pioneering works by Auer et al. (2008) and Bartlett & Tewari (2012) laid foundation for model-based algorithms operating within the online framework with sub-linear regret. In recent years, the pursuit of improved regret guarantees has led to the emergence of a multitude of new algorithms. In tabular case, these advancements include numerous model-based approaches (Ouyang et al., 2017; Fruit et al., 2018; Zhang & Ji, 2019; Ortner, 2020) and model-free algorithms (Abbasi-Yadkori et al., 2019; Wei et al., 2020; Hao et al., 2021; Lazic et al., 2021; Zhang & Xie, 2023). In the context of function approximation, POLITEX (Abbasi-Yadkori et al., 2019), a variant of the regularized policy iteration, is the first model-free algorithm with linear value-function approximation, and achieves $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$ regret for the ergodic MDP. The work by Hao et al. (2021) followed the same setting and improved the results to $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ with an adaptive approximate policy iteration (AAPI) algorithm. Wei et al. (2021) proposed an optimistic Q-learning algorithm FOPO for the linear function approximation, and achieve a near-optimal $\tilde{\mathcal{O}}(\sqrt{T})$ regret. On another line of research, Wu et al. (2022) delved into the linear function approximation under the framework of linear mixture model, which is mutually uncoverable concerning linear MDPs (Wei et al., 2021), and proposed UCRL2-VTR based on the value-targeted regression (Ayoub et al., 2020). Recent work of Chen et al. (2022a) expanded the scope of research by addressing the general function approximation problem in average-reward RL and proposed the SIM-TO-REAL algorithm, which can be regarded as an extension to UCRL2-VTR. In comparison to the works mentioned, our algorithm, LOOP, not only addresses all the problems examined in those studies but also extends its applicability to newly identified models. See Table 1 for a summary.

Function Approximation in Finite-horizon MDPs. In the pursuit of developing sample-efficient algorithms capable of handling large state spaces, extensive research efforts have converged on the linear function approximation problems within the finite-horizon setting. See Yang & Wang (2019); Wang et al. (2019); Jin et al. (2020); Ayoub et al. (2020); Cai et al. (2020); Zhou et al. (2021a;b); Zhou & Gu (2022); Agarwal et al. (2022); He et al. (2022); Zhong & Zhang (2023); Zhao et al. (2023); Huang et al. (2023); Li & Sun (2023) and references therein. Furthermore, Wang et al. (2020) studied RL with general function approximation and adopted the eluder dimension (Russo & Van Roy, 2013) as a complexity measure. Before this, Jiang et al. (2017) considered a substantial subset of problems with low Bellman ranks. Building upon these foundations, Jin et al. (2021) combined both the Eluder dimension and Bellman error, thereby broadening the scope of solvable problems under the concept of the Bellman Eluder (BE) dimension. In a parallel line of research, Sun et al. (2019) proposed the witness ranking focusing on the low-rank structures, and Du et al. (2021) extended it to encompass more scenarios with the bilinear class. Besides, Foster et al. (2021; 2023) provided a unified framework, decision estimation coefficient, for interactive decision making. The work of Chen et al. (2022b) extended the value-based GOLF (Jin et al., 2021) with the introduction of the discrepancy loss function to handle the broader admissible Bellman characterization (ABC) class. More recently, Zhong et al. (2022); Liu et al. (2023b) proposed a unified framework measured by generalized eluder coefficient (GEC), an extension to Dann et al. (2021) that captures almost all known tractable problems. All these works are restricted to the finite-horizon regime, and their complexity measure and algorithms are not applicable in the infinite-horizon average-reward setting.

Low-Switching Cost Algorithms. Addressing low-switching cost problems in bandit and reinforcement learning has seen notable progress. Abbasi-Yadkori et al. (2011) first proposed an algorithm for linear bandits with $\mathcal{O}(\log T)$ switching cost. Subsequent research extended this to tabular MDPs, including works of Bai et al. (2019); Zhang et al. (2020). A significant stride was made by Kong et al. (2021), who introduced importance scores to handle low-switching cost scenarios in general function approximation with complexity measured by eluder dimension (Russo & Van Roy, 2013). Recently, Xiong et al. (2023) introduced the eluder condition (EC) class, offering a comprehensive framework to address all tractable low-switching cost problems above. In the context of average-reward RL, Wei et al. (2021); Wu et al. (2022); Chen et al. (2022a); Hu et al. (2022) developed low-switching algorithms to control the regret under linear structure or model-based class, leaving a unifying framework for both value-based and model-based problems an open problem.

Further Elaboration on Our Contributions and Technical Novelties. Compared to episodic MDPs or discounted MDPs, AMDPs present unique challenges that prevent a straightforward extension of existing algorithms and analyses from these well-studied domains. One notable distinction is a different regret notion in average-reward RL due to a different form of the Bellman optimality equation. Furthermore, such a difference is coupled with the challenge of exploration in the context of general function approximation. To effectively bound this regret, we introduce a new regret decomposition approach within the context of general function approximation (refer to (5.1) and (5.3)). This regret decomposition suggests that the total regret can be controlled by the cumulative Bellman error and the switching cost. Inspired by this, we propose an optimistic algorithm with lazy updates in the general function approximation setting, which uses the residue of the loss function as the indicator for deciding when to conduct policy updates. Such a lazy policy update scheme adaptively divides the total of T steps into $\mathcal{O}(\log T)$ epochs, which is significantly different from (OLSVI.FH; Wei et al., 2021) that reduces the infinite-horizon setting to the finite-horizon setting by splitting the whole learning procedure into several H-length epoch, where H typically chosen as $\Theta(\sqrt{T})$ (Wei et al., 2021). We remark that such an adaptive lazy updating design and corresponding analysis are pivotal in achieving the optimal $\tilde{\mathcal{O}}(\sqrt{T})$ rate, as opposed to the $\tilde{\mathcal{O}}(T^{3/4})$ regret in (OLSVI.FH; Wei et al., 2021). Moreover, our approach is an extension to the existing lazy update approaches for average-reward setting (Wei et al., 2021; Wu et al., 2022) that leverages the postulated linear structure and is not applicable to problems with general function approximation. Furthermore, to accommodate the average-reward term, we introduce a new complexity measure AGEC, which characterizes the exploration challenge in general function approximation. Compared with Zhong et al. (2022), our additional transferability restriction is tailored for the infinite-horizon setting and plays a crucial role in analyzing the low-switching error. Despite this additional transferability restriction, AGEC can still serve as a unifying complexity measure in the infinite-horizon average-reward setting, like the role of GEC in the finite-horizon setting. Specifically, AGEC captures a rich class of tractable AMDP models, including all previously recognized AMDPs, including all known tractable AMDPs, and some newly identified AMDPs. See Table 1 for a summary.

Discussion about distribution families Beyond the singleton distribution family $\mathcal{D}_\Delta$ taken in this paper, there exists a notable distribution family $\mathcal{D}_\mathcal{H} = \{\mathcal{D}_{\mathcal{H},t}\}_{t\in[T]}$, proposed in Jin et al. (2021), where $\mathcal{D}_{\mathcal{H},t}$ characterizes probability measure over $\mathcal{S}\times\mathcal{A}$ obtained by executing different policies induced by $f_1,\dots,f_{t-1}\in\mathcal{H}$, measures the detailed distribution under sequential policies. However, in this paper, we exclude the consideration of $\mathcal{D}_\mathcal{H}$ for two principal reasons. First, evaluations of average-reward RL focus on the difference between *observed* rewards $r(s_t,a_t)$ and optimal average reward J^* — as opposed to the expected value V_h^π (i.e *expected* sum of reward) under specific policy and optimal value at step $h\in[H]$ in episodic setting — rendering the introduction of $\mathcal{D}_\mathcal{H}$ unnecessary. Second, in infinite settings, the measure of such distribution becomes highly intricate and impractical given *different* policy induced by $f_1,\dots,f_T$ over a potentially infinite T-steps. As a comparison, in the episode setting a *fixed* policy induced by f_t is executed over a finite H-step.

B Alternative Choices of Discrepancy Function

Note that there is another line of research that addresses model-based problems using Maximum Likelihood Estimator (MLE)-based approaches (Liu et al., 2023a; Zhong et al., 2022), as opposed to the value-targeted regression with reward function known. We remark that these MLE-based approaches can be also incorporated within our framework through the use of the discrepancy function:

$$l_{f'}(f,g,\zeta_t)=\frac{1}{2}|\mathbb{P}_g(s_{t+1}|s_t,a_t)/\mathbb{P}_{f^*}(s_{t+1}|s_t,a_t)-1|, \tag{B.2}$$

where the trajectory is $\zeta_t=(s_t,a_t,s_{t+1})$ with expectation taken over the next transition state s_{t+1} from $\mathbb{P}(\cdot|s_t,a_t)$ such that $\mathbb{E}_{\zeta_t}[l_{f'}(f,g,\zeta_t)]=\mathrm{TV}\big(\mathbb{P}_f(\cdot|s_t,a_t),\mathbb{P}_{f^*}(\cdot|s_t,a_t)\big)$. To accommodate the discrepancy function in (B.2), we introduce a natural variant of AGEC defined below.

Algorithm 2 MLE-based Local-fitted Optimization with Optimism - MLE-LOOP$(\mathcal{H}, T, \delta)$

Parameter: Initial s_1, span $\mathrm{sp}(V^*)$, optimistic parameter $\beta = c\log\big(T\mathcal{N}_{\mathcal{H}}(1/T)/\delta\big)$
Initialize: Draw $a_1 \sim \mathrm{Unif}(\mathcal{A})$ and set $\tau_0 \leftarrow 0$, $\Upsilon_0 \leftarrow 0$, $\mathcal{B}_0 \leftarrow \emptyset$, $\mathcal{D}_0 \leftarrow \emptyset$.

for $t = 1, \ldots, T$ **do**
  if $t = 1$ or $\Upsilon_{t-1} \geq 3\sqrt{\beta t}$ **then**
    Update $\tau_t = t$.
    Solve optimization problem $f_t = \mathrm{argmax}_{f_t \in \mathcal{B}_t} J_{f_t}$, where

$$\mathcal{B}_t = \left\{ f \in \mathcal{H} : \mathcal{L}_{\mathcal{D}_{t-1}}(f, f) - \inf_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{D}_{t-1}}(f, g) \leq \beta \right\}, \tag{B.1}$$

    Update $Q_t = Q_{f_t}$, $V_t = V_{f_t}$, $J_t = J_{f_t}$ and $g_t = \inf_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{D}_{t-1}}(f_t, g)$.
  else
    Retain $(f_t, g_t, J_t, V_t, Q_t, \tau_t) = (f_{t-1}, g_{t-1}, J_{t-1}, V_{t-1}, Q_{t-1}, \tau_{t-1})$.
  Execute $a_t = \mathrm{argmax}_{a \in \mathcal{A}} Q_t(s_t, a)$.
  Collect reward $r_t = r(s_t, a_t)$ and transition state s_{t+1}.
  Update $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(s_t, a_t, r_t, s_{t+1})\}$, $\Upsilon_t = \sum_{s,a \in \mathcal{D}_t} \mathrm{TV}\big(\mathbb{P}_{f_t}(\cdot|s,a), \mathbb{P}_{g_t}(\cdot|s,a)\big)$.

Definition 9 (MLE-AGEC). Given hypothesis class $\mathcal{H}$, the MLE-discrepancy function $\{l_f\}_{f \in \mathcal{H}}$ in (B.2) and constant $\epsilon > 0$, the MLE-based average-reward generalized eluder coefficients MLE-AGEC$(\mathcal{H}, \{l_f\}, \epsilon)$ is defined as the smallest coefficients κ_{G} and d_{G} such that following two conditions hold with absolute constants $C_1 > 0$:

(i) (MLE-Bellman dominance) There exists constant $d_{\mathrm{G}} > 0$ such that

$$\sum_{t=1}^{T} \mathcal{E}(f_t)(s_t, a_t) \leq d_{\mathrm{G}} \cdot \mathrm{sp}(V^*) \sum_{t=1}^{T} \|\mathbb{E}_{\zeta_t}[l_{f_t}(f_t, f_t, \zeta_i)]\|_1.$$

(ii) (MLE-Transferability) There exists constant $\kappa_{\mathrm{G}} > 0$ such that for hypotheses $f_1, \ldots, f_T \in \mathcal{H}$, if it holds that $\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\|_1 \leq \sqrt{\beta t}$ for all $t \in [T]$, then we have

$$\sum_{t=1}^{T} \|\mathbb{E}_{\zeta_t}[l_{f_t}(f_t, f_t, \zeta_t)]\|_1 \leq \mathrm{poly}(\log T)\sqrt{\kappa_{\mathrm{G}} \cdot \beta T} + C_1 \cdot \mathrm{sp}(V^*)^2 \min\{\kappa_{\mathrm{G}}, T\} + 2T\epsilon^2.$$

The main difference between the MLE-based variant (see Definition 9) and the original AGEC (see Definition 3) is that the coefficients are defined over the ℓ_1-norm rather than the ℓ_2-norm, and a similar condition is considered in Liu et al. (2023a); Xiong et al. (2023). Now, we are ready to introduce the algorithm for the alternative discrepancy function in (B.2) and please see Algorithm 2 for complete pseudocode. The main modification lies in the construction of confidence set and update condition Υ_t. Here, the confidence set now follows

$$\mathcal{B}_t = \{f_t \in \mathcal{H} : \mathcal{L}_{\mathcal{D}_{t-1}}(f_t, f_t) - \inf_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{D}_{t-1}}(f_t, g) \leq \beta\}, \quad \mathcal{L}_{\mathcal{D}}(f, g) = -\sum_{(s,a,s') \in \mathcal{D}} \log \mathbb{P}_g(s'|s,a).$$

In comparison, the update condition follows that $\Upsilon_t = \sum_{(s,a) \in \mathcal{D}_t} \mathrm{TV}\big(\mathbb{P}_{f_t}(\cdot|s,a), \mathbb{P}_{g_t}(\cdot|s,a)\big)$. Unlike the standard LOOP algorithm, the the confidence set and update condition in the MLE-based varint no longer shares the same construction. Following the literature of MLE-based algorithms, we adopt the bracket number to approximation the cardinality of the function class.

Definition 10 (ρ-bracket). Let $\rho > 0$ and $\mathcal{F}$ is a set of functions defined over $\mathcal{X}$. Under ℓ_1-norm, a set of functions $\mathcal{V}_\rho(\mathcal{F})$ is an ρ-bracket of $\mathcal{F}$ if for any $f \in \mathcal{F}$, there exists a function $f' \in \mathcal{F}$ such that the following two properties hold: (i) $f'(x) \geq f(x)$ for all $x \in \mathcal{X}$, and (ii) $\|f - f'\|_1 \leq \rho$. The bracketing number $\mathcal{B}_{\mathcal{F}}(\rho)$ is the cardinality of the smallest ρ-bracket needed to cover $\mathcal{F}$.

The theoretical guarantee is provided below.

Theorem 4 (Cumulative regret)**.** Under Assumptions 1-2 and the discepancy function in (B.2) with self-completeness such that $\mathcal{G} = \mathcal{H}$, there exists constant c such that for any $\delta \in (0, 1)$ and time horizon T, with probability at least $1 - 4\delta$, the regret of MLE-LOOP satisfies that

$$\mathrm{Reg}(T) \leq \mathcal{O}\Big(\mathrm{sp}(V^*) \cdot d\sqrt{T\beta}\Big),$$

where $\beta = c \log\big(T\mathcal{B}_{\mathcal{H}}(1/T)/\delta\big) \cdot \mathrm{sp}(V^*)$, $d = d_{\mathrm{G}}\sqrt{\kappa_{\mathrm{G}}}$. Here, $(d_{\mathrm{G}}, \kappa_{\mathrm{G}}) = $ MLE-AGEC$(\mathcal{H}, \{l_f\}, 1/\sqrt{T})$ denote MLE-AGEC defined in Definition 9 and $\mathcal{B}_{\mathcal{H}}(\cdot)$ denotes the bracketing number.

The proof of Theorem 4 is similar to that of Theorem 3, and can be found in Appendix H.1.

C Concrete Examples

In this section, we present concrete examples of problems for AMDP. We remark that the understanding of function approximation problems under the average-reward setting is quite limited, and to our best knowledge, existing works have primarily focused on linear approximation (Wei et al., 2021; Wu et al., 2022) and model-based general function approximation (Chen et al., 2022a). Here, we introduce a variety of function classes with low AGEC. Beyond the examples considered in existing work, these newly proposed function classes are mostly natural extensions from their counterpart the finite-horizon episode setting (Jin et al., 2020; Zanette et al., 2020; Du et al., 2021; Domingues et al., 2021), which can be extended to the average-reward problems with moderate justifications.

C.1 Linear Function Approximation and Variants

Linear function approximation Consider the linear FA, which encompasses a wide range of concrete problems with state-action bias function linear in a d-dimensional feature mapping. Specifically, a linear function class $\mathcal{H}$ is defined as $\mathcal{H} = \{Q(\cdot,\cdot) = \langle \omega, \phi(\cdot,\cdot)\rangle, J \in \mathcal{J}_{\mathcal{H}} | \|\omega\|_2 \leq \frac{1}{2}\mathrm{sp}(V^*)\sqrt{d}\}$, where the feature satisfies that $\|\phi\|_{2,\infty} \leq \sqrt{2}$ with first coordinate fixed to 1. We remark that such scaling is without loss of generality as justified in Lemma 19. To begin with, we first introduce two specific problems: linear AMDP and AMDP with linear Bellman completion.

Definition 11 (Linear AMDP, Wei et al. (2021))**.** There exists a known feature mapping $\phi : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^d$, an unknown d-dimensional signed measures $\mu = \big(\mu_1, \dots, \mu_d\big)$ over $\mathcal{S}$, and an unknown reward parameter $\theta \in \mathbb{R}^d$, such that the transition kernel the reward function can be written as

$$\mathbb{P}(\cdot|s,a) = \langle \phi(s,a), \mu(\cdot)\rangle, \quad r(s,a) = \langle \phi(s,a), \theta\rangle. \tag{C.1}$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Without loss of generality, we assume that the feature mapping ϕ satisfies that $\|\phi\|_{2,\infty} \leq \sqrt{2}$ with first coordinate fixed to 1, $\|\theta\|_2 \leq \sqrt{d}$ and $\|\mu(\mathcal{S})\|_2 \leq \sqrt{d}$, where we denote $\mu(\mathcal{S}) = \big(\mu_1(\mathcal{S}), \dots, \mu_d(\mathcal{S})\big)$ and $\mu_i(\mathcal{S}) = \int_{\mathcal{S}} \mathrm{d}\mu_i(s)$ be the total measure of $\mathcal{S}$.

We remark that the scaling on the feature mapping can help in overcoming the gap between the episodic setting and the average-reward one by ensuring the linear structure of Q- and V-value function under optimality (Wei et al., 2021). To illustrate the necessity, note that

$$Q^*(s,a) = r(s,a) + \mathbb{E}_{s'\sim\mathbb{P}(s,a)}[V^*(s')] - J^* = \phi(s,a)^\top \left(\theta - J^*\mathbf{e}_1 + \int_{\mathcal{S}} V^*(s')\mathrm{d}\mu(s')\right),$$

where denote $\mathbf{e}_1 = (1, 0, \dots, 0) \in \mathbb{R}^d$. Next, we provide the AMDPs with linear Bellman completion, modified from Zanette et al. (2020), which is a more general setting than linear AMDPs.

Definition 12 (Linear Bellman completion)**.** There exists a known feature mapping $\phi : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^d$ such that for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\omega \in \mathcal{W}_{\mathcal{H}}$ and $J \in \mathcal{J}_{\mathcal{H}}$, we have

$$\langle \mathcal{T}(\omega, J), \phi(s,a)\rangle := r(s,a) + \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}}\Big\{\omega^\top\phi(s',a')\Big\}\right] - J, \tag{C.2}$$

Generalized linear function approximation To introduce the nonlinearity beyond linear FA, we extend by incorporating a link function. In generalized linear FA, the hypotheses class is defined as

$\mathcal{H} = \{Q(\cdot,\cdot) = \sigma\left(\omega^\top \phi(\cdot,\cdot)\right), J \in \mathcal{J}_\mathcal{H} | \|\omega\|_2 \leq \sqrt{d}\}$, where $\|\phi(s,a)\|_{2,\infty} \leq 1$ and $\sigma: \mathbb{R} \mapsto \mathbb{R}$ is an α-bi-Lipschitz function with $\|\sigma\|_\infty \leq \frac{1}{2}\mathrm{sp}(V^*)$. We say σ is α-bi-Lipschitz continuous if

$$\frac{1}{\alpha} \cdot |x - x'| \leq |\sigma(x) - \sigma(x')| \leq \alpha \cdot |x - x'|, \quad \forall x, x' \in \mathbb{R}. \tag{C.3}$$

We remark that the generalized linear function class $\mathcal{H}$ degenerates to the standard linear function class $\mathcal{H}$ if we choose $\sigma(x) = x$. Modified from Wang et al. (2019) for the episodic setting, we define AMDPs with generalized linear Bellman completion as follows.

Definition 13 (Generalized linear Bellman completion)**.** There exists a known feature mapping $\phi: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\omega \in \mathcal{W}_\mathcal{H}$ and $J \in \mathcal{J}_\mathcal{H}$, we have

$$\sigma\left(\mathcal{T}(\omega, J)^\top \phi(\cdot,\cdot)\right) := r(s,a) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}}\left\{\sigma\left(\omega^\top \phi(s',a')\right)\right\}\right] - J. \tag{C.4}$$

The proposition below states that (generalized) linear function classes have low AGEC.

Proposition 5 (Linear FA $\subset$ Low AGEC)**.** Consider linear function class $\mathcal{H}_{\mathrm{Lin}}$ and generalized linear function class $\mathcal{H}_{\mathrm{Glin}}$ with a d-dimensional feature mapping $\phi: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, if the problem follows one of Definitions 11-13, then it have low AGEC under Bellman discrepancy in (3.1):

$$d_{\mathrm{G}} \leq \mathcal{O}\big(d \log\left(\mathrm{sp}(V^*)\sqrt{d}\epsilon^{-1}\right) \log T\big), \quad \kappa_{\mathrm{G}} \leq \mathcal{O}\big(d \log\left(\mathrm{sp}(V^*)\sqrt{d}\epsilon^{-1}\right)\big).$$

Linear Q^*/V^* AMDP Moreover, we consider the linear Q^*/V^* AMDPs, which is modified from the one in Du et al. (2021) under the episodic setting.

Definition 14 (Linear Q^*/V^* AMDP)**.** There exists known feature mappings $\phi: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_1}$, $\psi: \mathcal{S} \mapsto \mathbb{R}^{d_2}$, and unknown vectors $\omega^* \in \mathbb{R}^{d_1}$, $\theta^* \in \mathbb{R}^{d_2}$ such that optimal value functions follow

$$Q^*(s,a) = \langle \phi(s,a), \omega^* \rangle, \quad V^*(s') = \langle \psi(s'), \theta^* \rangle,$$

for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Without loss of generality, we assume that features $\|\phi\|_{2,\infty} \leq \sqrt{2}$ and $\|\psi\|_{2,\infty} \leq \sqrt{2}$ with first coordinate fixed to 1, and $\|\omega^*\|_2 \leq \frac{1}{2}\mathrm{sp}(V^*)\sqrt{d_1}$, $\|\theta^*\|_2 \leq \frac{1}{2}\mathrm{sp}(V^*)\sqrt{d_2}$.

The proposition below states that linear Q^*/V^* also has low AGEC.

Proposition 6 (Linear $Q^*/V^* \subset$ Low AGEC)**.** Linear Q^*/V^* AMDPs with coupled (d_1, d_2)-dimensional feature mappings $\phi: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_1}$ and $\psi: \mathcal{S} \mapsto \mathbb{R}^{d_2}$ have low AGEC such that

$$d_{\mathrm{G}} \leq \mathcal{O}\big(d^+ \log\left(\mathrm{sp}(V^*)\sqrt{d^+}\epsilon^{-1}\right) \log T\big), \quad \kappa_{\mathrm{G}} \leq \mathcal{O}\big(d^+ \log\left(\mathrm{sp}(V^*)\sqrt{d^+}\epsilon^{-1}\right)\big),$$

where denote $d^+ = d_1 + d_2$ as the sum of dimensions of features.

The proposition above asserts that in linear Q^*/V^* AMDPs, with additional structural information in state bias function, add $\tilde{\mathcal{O}}(d_2)$ in complexity from the AGEC perspective. We remark that linear FA, generalized linear FA, and linear Q^*/V^* AMDPs are typical value-based problems. The proof of this proposition relies on ABE dimension as an intermediate, and then uses Lemma 2.

C.2 Kernel Function Approximation

In this subsection, we first introduce the notion of effective dimension. With this notion, we prove a useful proposition that any kernel function class with a low effective dimension has low AGEC. Consider kernel FA, a natural extension to linear FA from d-dimensional Euclidean space $\mathbb{R}^d$ to a decomposable kernel Hilbert space $\mathcal{K}$. Formally, a kernel function class is defined as $\mathcal{H} = \left\{Q(\cdot,\cdot) = \langle \phi(\cdot,\cdot), \omega \rangle_\mathcal{K}, J \in \mathcal{J}_\mathcal{H} | \ \|\omega\|_\mathcal{K} \leq \mathrm{sp}(V^*)R\right\}$, where the feature mapping $\phi: \mathcal{S} \times \mathcal{A} \mapsto \mathcal{K}$ satisfies that $\|\phi\|_{\mathcal{K},\infty} \leq 1$. To measure the complexity of problems in a Hilbert space $\mathcal{K}$ with a potentially infinite dimension, we introduce the ϵ-effective dimension below.

Definition 15 (ϵ-effective dimension)**.** Consider a set $\mathcal{Z}$ with the possibly infinite elements in a given separable Hilbert space $\mathcal{K}$, the ϵ-effective dimension, denoted by $\dim_{\mathrm{eff}}(\mathcal{Z}, \epsilon)$, is defined as the length n of the longest sequence satisfying the condition below:

$$\sup_{\mathbf{z}_1,\ldots,\mathbf{z}_n \in \mathcal{Z}} \left\{\frac{1}{n} \log\det\left(\mathbf{I} + \frac{1}{\epsilon^2}\sum_{t=1}^n \mathbf{z}_i \mathbf{z}_i^\top\right) \leq \frac{1}{e}\right\}.$$

Here, the concept of ϵ-effective dimension is inspired by the measurement of maximum information gain (Srinivas et al., 2009) and is later introduced as a complexity measure of Hilbert space in Du et al. (2021); Zhong et al. (2022). Similar to Jin et al. (2021), we augment the assumption by requiring the $\mathcal{H}$ to be self-complete under average-reward Bellman operator, i.e., $\mathcal{G} = \mathcal{H}$. Next, the proposition below demonstrates that kernel FA has low AGEC.

Proposition 7 (Kernel FA $\subset$ Low AGEC)**.** Under the self-completeness, kernel FA with function class $\mathcal{H}_{\text{Ker}}$ concerning a known feature mapping $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{K}$ have low AGEC such that

$$d_{\text{G}} \le \dim_{\text{eff}}\big(\mathcal{X}, \epsilon/2\text{sp}(V^*)R\big) \log T, \quad \kappa_{\text{G}} \le \dim_{\text{eff}}\big(\mathcal{X}, \epsilon/2\text{sp}(V^*)R\big),$$

where denote $\mathcal{X} = \{\phi(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$ as the collection of feature mappings.

The proposition above shows that the kernel FA with a low ϵ-effective dimension over the Hilbert space also has low AGEC. As a special case of kernel FA, if we choose $\mathcal{K} = \mathbb{R}^d$, then we can prove that the RHS in the proposition above is upper bounded by $\tilde{\mathcal{O}}(d)$.

C.3 Linear Mixture AMDP

In this subsection, we focus on the average-reward linear mixture problem considered in Wu et al. (2022). In this context, the hypotheses function class is defined as $\mathcal{H} = \{\mathbb{P}(s'|s,a) = \langle \theta, \phi(s,a,s')\rangle$,
$r(s,a) = \langle \theta, \psi(s,a)\rangle | \|\theta\|_2 \le 1\}$ with known feature mappings $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}^d$, $\psi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, and an unknown parameter $\theta \in \mathbb{R}^d$. The problem is defined as below.

Definition 16 (Linear mixture AMDPs, Wu et al. (2022))**.** There exists a known feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}^d$, $\psi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, and an unknown vector $\theta \in \mathbb{R}^d$, it holds that

$$\mathbb{P}(s'|s,a) = \langle \theta, \phi(s,a,s')\rangle, \quad r(s,a) = \langle \theta, \psi(s,a)\rangle,$$

for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Without loss of generality, we assume $\|\phi\|_{2,\infty} \le \sqrt{d}$, $\|\psi\|_{2,\infty} \le \sqrt{d}$.

Now we show that the linear mixture problem is tractable under the framework of AGEC.

Proposition 8 (Linear mixture $\subset$ Low AGEC)**.** Consider linear mixture problem with hypotheses class $\mathcal{H}$ and d-dimensional feature mappings (ϕ, ψ). If we choose discrepancy function as

$$l_{f'}(f, g, \zeta_t) = \theta_g^\top \left(\psi(s_t, a_t) + \int_{\mathcal{S}} \phi(s_t, a_t, s') V_{f'}(s') \mathrm{d}s' \right) - r(s_t, a_t) - V_{f'}(s_{t+1}), \tag{C.5}$$

and takes $\mathcal{H} = \mathcal{G}$ with operator following $\mathcal{P}(f) = f^*$ for all $f \in \mathcal{H}$, it has low AGEC such that

$$d_{\text{G}} \le \mathcal{O}\left(d \log\left(\text{sp}(V^*)T/\sqrt{d}\epsilon\right)\right), \quad \kappa_{\text{G}} \le \mathcal{O}\left(d \log\left(\text{sp}(V^*)T/\sqrt{d}\epsilon\right)\right).$$

The proposition posits that AGEC can capture the linear mixture AMDP, based on a modified version of the Bellman discrepancy function in (2.2). In contrast to the linear FA discussed in Appendix C.1, the presence of the average-reward term in this model-based problem does not impose any additional computational or statistical burden, and there is no need for structural assumptions on feature mappings, such as a fixed first coordinate, considering discrepancy in (C.5).

D Proof of Main Results for Loop

D.1 Proof of Theorem 3

Proof of Theorem 3. Note that the regret can be decomposed as

$$\begin{aligned}
\text{Reg}(T) &= \sum_{t=1}^{T} \Big(J^* - r(s_t, a_t) \Big) \le \sum_{t=1}^{T} \Big(J_t - r(s_t, a_t) \Big) && \text{(optimism)} \\
&\overset{(a)}{=} \sum_{t=1}^{T} \mathcal{E}(f_t)(s_t, a_t) - \sum_{t=1}^{T} \mathbb{E}_{s_{t+1} \sim \mathbb{P}(\cdot|s_t,a_t)} \Big[Q_t(s_t, a_t) - \max_{a \in \mathcal{A}} Q_t(s_{t+1}, a) \Big] \\
&\overset{(b)}{=} \underbrace{\sum_{i=1}^{T} \mathcal{E}(f_t)(s_t, a_t)}_{\text{Bellman error}} + \underbrace{\sum_{t=1}^{T} \Big[\mathbb{E}_{s_{t+1} \sim \mathbb{P}(\cdot|s_t,a_t)} [V_t(s_{t+1})] - V_t(s_t) \Big]}_{\text{Realization error}}, && \text{(D.1)}
\end{aligned}$$

where step (a) and step (b) follow the definition of the Bellman optimality operator and the greedy policy. Below, we will present the bound of Bellman error and Realization error respectively.

Step 1: Bound over Bellman error Recall that the of confidence set ensures that $\mathcal{L}_{\mathcal{D}_{t-1}}(f_t, f_t) - \inf_{g\in\mathcal{G}} \mathcal{L}_{\mathcal{D}_{t-1}}(f_t, g) \le \mathcal{O}(\beta)$ across all steps. Using the concentration arguments, we can infer

$$\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\|_2^2 \le \mathcal{O}(\beta), \tag{D.2}$$

with high probability and the formal statements are deferred to Lemma 10 in Appendix D.3. In the following arguments, we assume the above event holds. Take $\epsilon = 1/\sqrt{T}$, recall the definition of dominance coefficient d_{G} in AGEC$(\mathcal{H}, \mathcal{J}, l, \epsilon)$ and it directly indicates that

$$\text{Bellman error} = \sum_{t=1}^{T} \mathcal{E}(f_t)(s_t, a_t) \le \left[d_{\mathrm{G}} \sum_{t=1}^{T}\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta)]\|_2^2 \right]^{1/2} + \mathcal{O}\left(\mathrm{sp}(V^*)\sqrt{d_{\mathrm{G}} T}\right),$$

and thus the Bellman error can be upper bounded by $\mathcal{O}\left(\mathrm{sp}(V^*)\sqrt{d_{\mathrm{G}}\beta T}\right)$.

Step 2: Bound over Realization error To bound Realization error, we use the concentration argument and the upper-boundded switching cost. Note that

$$\begin{aligned}
\text{Realization error} &\overset{(c)}{=} \sum_{t=1}^{T} [V_t(s_{t+1}) - V_t(s_t)] + \mathcal{O}\big(\mathrm{sp}(V^*)\sqrt{T\log(1/\delta)}\big),\\
&= \sum_{t=1}^{T} \left[V_{\tau_t}(s_{t+1}) - V_{\tau_{t+1}}(s_{t+1})\right] + \mathcal{O}\big(\mathrm{sp}(V^*)\sqrt{T\log(1/\delta)}\big), \qquad \text{(Shift)}\\
&= \sum_{t=1}^{T} \left[V_{\tau_t}(s_{t+1}) - V_{\tau_{t+1}}(s_{t+1})\right] \mathbb{1}(\tau_t \ne \tau_{t+1}) + \mathcal{O}\big(\mathrm{sp}(V^*)\sqrt{T\log(1/\delta)}\big)\\
&\le \mathrm{sp}(V^*)\cdot\mathcal{N}(T) + \mathcal{O}\big(\mathrm{sp}(V^*)\sqrt{T\log(1/\delta)}\big) \overset{(d)}{\le} \mathcal{O}\big(\mathrm{sp}(V^*)\cdot\kappa_{\mathrm{G}}\sqrt{T\log(1/\delta)}\big),
\end{aligned} \tag{D.3}$$

where step (c) directly follows the Azuma-Hoeffding inequality and step (d) is based the fact that $\|V_{\tau_t} - V_{\tau_{t+1}}\|_\infty \le \mathrm{sp}(V^*)$ and the bounded switching cost such that $\mathcal{N}(T) \le \mathcal{O}(\kappa_{\mathrm{G}}\log T)$, where κ_{G} is the transferability coefficient in AGEC with $\epsilon = 1/\sqrt{T}$. Please refer to Lemma 11 in Appendix D.4 for the detailed statement and proof of the bounded switching cost.

Step 3: Combine the bounded erroes Plugging (D.2) and (D.3) back into (D.1), we have

$$\begin{aligned}
\mathrm{Reg}(T) &\le \text{Bellman error} + \text{Realization error}\\
&\le \mathcal{O}\left(\mathrm{sp}(V^*)\sqrt{d_{\mathrm{G}}\beta T}\right) + \mathcal{O}\big(\mathrm{sp}(V^*)\kappa_{\mathrm{G}}\sqrt{T\log(1/\delta)}\big) = \mathcal{O}\Big(\mathrm{sp}(V^*)\cdot d\sqrt{T\beta}\Big),
\end{aligned}$$

where $d = \max\{\sqrt{d_{\mathrm{G}}}, \kappa_{\mathrm{G}}\}$ is a function of $(d_{\mathrm{G}}, \kappa_{\mathrm{G}}) = \mathrm{AGEC}(\mathcal{H}, \{l_f\}_{f\in\mathcal{H}}, 1/\sqrt{T})$. In the arguments above, the optimistic parameter is chosen as $\beta = c\log\left(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(1/T)/\delta\right)\cdot\mathrm{sp}(V^*)$, which takes the upper bound of the optimistic parameters, aligning with the choice in Lemma 9, Lemma 10, and Lemma 11. Then finish the proof of cumulative regret for LOOP in Algorithm 1. □

D.2 Proof of Lemma 9

Lemma 9 (Optimism)**.** Under Assumptions 1-4, LOOP is an optimistic algorithm such that it ensures $J_t \ge J^*$ for all $t \in [T]$ with probability greater than $1-\delta$.

Proof of Lemma 9. Denote $\mathcal{V}_\rho(\mathcal{G})$ be the ρ-cover of $\mathcal{G}$ and $\mathcal{N}_{\mathcal{G}}(\rho)$ be the size of ρ-cover $\mathcal{V}_\rho(\mathcal{G})$. Consider fixed $(i, g) \in [T]\times\mathcal{G}$ and define the auxiliary function

$$X_{i,f_i}(g) := \|l_{f_i}(f^*, g, \zeta_i)\|_2^2 - \|l_{f_i}(f^*, f^*, \zeta_i)\|_2^2, \tag{D.4}$$

where f^* is the optimal hypothesis in value-based problems and the true hypothesis in model-based ones. Let $\mathscr{F}_t$ be the filtration induced by $\{s_1, a_1, \ldots, s_t, a_t\}$ and note that $f_1, \ldots, f_t$ is fixed under the filtration, then we have

$$\begin{aligned}
\mathbb{E}[X_{i,f_i}(g)|\mathscr{F}_i] &= \mathbb{E}_{\zeta_i}[\|l_{f_i}(f^*, g, \zeta_i)\|_2^2 - \|l_{f_i}(f^*, \mathcal{P}(f^*), \zeta_i)\|_2^2 | \mathscr{F}_i] \\
&= \mathbb{E}_{\zeta_i}\left[\left[l_{f_i}(f^*, g, \zeta_i) - l_{f_i}(f^*, \mathcal{P}(f^*), \zeta_i)\right] \cdot \left[l_{f_i}(f^*, g, \zeta_i) + l_{f_i}(f^*, \mathcal{P}(f^*), \zeta_i)\right] \Big| \mathscr{F}_i\right] \\
&= \mathbb{E}_{\zeta_i}\left[\mathbb{E}_{\zeta_i}\left[l_{f_i}(f^*, g, \zeta_i)\right] \cdot \left[l_{f_i}(f^*, g, \zeta_i) + l_{f_i}(f^*, \mathcal{P}(f^*), \zeta_i)\right] \Big| \mathscr{F}_i\right] \\
&= \|\mathbb{E}_{\zeta_i}\left[l_{f_i}(f^*, g, \zeta_i)\right]\|_2^2,
\end{aligned}$$

where the equation follows the definition of generalized completeness (see Assumption 4):

$$\begin{cases}
\mathbb{E}_{\zeta_i}[l_{f'}(f, g, \zeta)] = l_{f'}(f, g, \zeta) - l_{f'}(f, \mathcal{P}(f), \zeta), \\
\mathbb{E}_{\zeta_i}[l_{f'}(f, g, \zeta)] = \mathbb{E}_{\zeta_i}\left[l_{f'}(f, g, \zeta) + l_{f'}(f, \mathcal{P}(f), \zeta)\right].
\end{cases}$$

Similarly, we can obtain that the second moment of the auxiliary function is bounded by

$$\mathbb{E}[X_{i,f_i}(g)^2|\mathscr{F}_i] \leq \mathcal{O}\Big(\mathrm{sp}(V^*)^2 \, \|\mathbb{E}_{\zeta_i}[l_{f_i}(f^*, g, \zeta_i)]\|_2^2\Big),$$

By Freedman's inequality (see Lemma 18), with probability greater than $1 - \delta$ it holds that

$$\begin{aligned}
&\Big|\sum_{i=1}^t X_{i,f_i}(g) - \sum_{i=1}^t \|\mathbb{E}_{\zeta_i}[l_{f_i}(f^*, g, \zeta_i)]\|_2^2\Big| \\
&\quad \leq \mathcal{O}\left(\sqrt{\log(1/\delta) \cdot \mathrm{sp}(V^*)^2 \sum_{i=1}^t \|\mathbb{E}_{\zeta_i}[l_{f_i}(f^*, g, \zeta_i)]\|_2^2} + \log(1/\delta)\right).
\end{aligned}$$

By taking union bound over $[T] \times \mathcal{V}_\rho(\mathcal{G})$, for any $(t, \phi) \in [T] \times \mathcal{V}_\rho(\mathcal{G})$ we have $-\sum_{i=1}^t X_{i,f_i}(\phi) \leq \mathcal{O}(\zeta)$, where $\zeta = \mathrm{sp}(V^*) \log(T\mathcal{N}_{\mathcal{G}}(\rho)/\delta)$ and we use the fact that $\|\mathbb{E}_{\zeta_i}[l_{f_i}(f^*, g, \zeta_i)]\|_2^2$ is non-negative. Recall the definition of ρ-cover, it ensures that for any $g \in \mathcal{G}$, there exists $\phi \in \mathcal{V}_\epsilon(\mathcal{G})$ such that $\|g(s, a) - \phi(s, a)\|_1 \leq \rho$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Therefore, for any $g \in \mathcal{G}$ we have

$$-\sum_{i=1}^t X_{i,f_i}(g) \leq \mathcal{O}\Big(\zeta + t\rho\Big). \tag{D.5}$$

Combine the (D.5) above and the designed confidence set, then for all $t \in [T]$ it holds that

$$\mathcal{L}_{\mathcal{D}_{t-1}}(f^*, f^*) - \inf_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{D}_{t-1}}(f^*, g) = -\sum_{i=1}^{t-1} X_{i,f_i}(\tilde{g}) \leq \mathcal{O}\Big(\zeta + t\rho\Big) < \beta, \tag{D.6}$$

where $\tilde{g}$ is the local minimizer to $\mathcal{L}_{\mathcal{D}_{t-1}}(f^*, g)$, and we take the covering coefficient as $\rho = 1/T$ and optimistic parameter as $\beta = c \log\left(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(1/T)/\delta\right) \cdot \mathrm{sp}(V^*)$. Based on (D.6), with probability greater than $1 - \delta$, f^* is a candidate of the confidence set such that $J_t \geq J^*$ for all $t \in [T]$. □

D.3 Proof of Lemma 10

Lemma 10. For fixed $\rho > 0$ and the optimistic parameter $\beta = c\big(\mathrm{sp}(V^*) \cdot \log\left(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(\rho)/\delta\right) + T\rho\big)$ where $c > 0$ is constant large enough, then it holds that

$$\sum_{i=1}^{t-1} \mathbb{E}_{\zeta_i} \|l_{f_i}(f_t, f_t, \zeta_i)\|^2 \leq \mathcal{O}(\beta), \tag{D.7}$$

for all $t \in [T]$ with probability greater than $1 - \delta$.

Proof of Lemma 10. Denote $\mathcal{V}_\rho(\mathcal{H})$ be the ρ-cover of $\mathcal{H}$ and $\mathcal{N}_{\mathcal{H}}(\rho)$ be the size of ρ-cover $\mathcal{V}_\rho(\mathcal{H})$. Consider fixed $(i, f) \in [T] \times \mathcal{H}$ and define the auxiliary function

$$X_{i,f_i}(f) := \big\|l_{f_i}(f, f, \zeta_i)\big\|_2^2 - \big\|l_{f_i}(f, \mathcal{P}(f), \zeta_i)\big\|_2^2,$$

Let $\mathscr{F}_t$ be the filtration induced by $\{s_1, a_1, \ldots, s_t, a_t\}$ and note that $f_1, \ldots, f_t$ is fixed under the filtration, then we have

$$\begin{aligned}\mathbb{E}[X_{i,f_i}(f)|\mathscr{F}_i] &= \mathbb{E}_{\zeta_i}[\|l_{f_i}(f,f,\zeta_i)\|_2^2 - \|l_{f_i}(f,\mathcal{P}(f),\zeta_i)\|_2^2 \,|\mathscr{F}_i] \\ &= \mathbb{E}_{\zeta_i}\Big[\big[l_{f_i}(f,f,\zeta_i) - l_{f_i}(f,\mathcal{P}(f),\zeta_i)\big] \cdot \big[l_{f_i}(f,f,\zeta_i) + l_{f_i}(f,\mathcal{P}(f),\zeta_i)\big]\Big|\mathscr{F}_i\Big] \\ &= \mathbb{E}_{\zeta_i}\big[l_{f_i}(f,f,\zeta_i)\big] \cdot \mathbb{E}_{\zeta_i}\big[l_{f_i}(f,f,\zeta_i) + l_{f_i}(f,\mathcal{P}(f),\zeta_i)\big|\mathscr{F}_i\big] \\ &= \big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,f,\zeta_i)\big]\big\|_2^2,\end{aligned}$$

where the equation generalized completeness (see Lemma 9). Similarly, we can obtain that the second moment of the auxiliary function is bounded by

$$\mathbb{E}[X_{i,f_i}(f)^2|\mathscr{F}_i] \le \mathcal{O}\Big(\mathrm{sp}(V^*)^2 \,\big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,f,\zeta_i)\big]\big\|_2^2\Big),$$

By Freedman's inequality in Lemma 18, with probability greater than $1-\delta$ we have

$$\begin{aligned}&\Big|\sum_{i=1}^t X_{i,f_i}(f) - \sum_{i=1}^t \big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,f,\zeta_i)\big]\big\|_2^2\Big| \\ &\quad\le \mathcal{O}\left(\sqrt{\log(1/\delta)\cdot \mathrm{sp}(V^*)^2 \sum_{i=1}^t \big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,f,\zeta_i)\big]\big\|_2^2 + \log(1/\delta)}\right).\end{aligned}$$

Define $\zeta = \mathrm{sp}(V^*)\log(T\mathcal{N}_{\mathcal{H}}(\rho)/\delta)$, by taking a union bound over ρ-covering of hypothesis set $\mathcal{H}$, we can obtain that with probability greater than $1-\delta$, for all $(t,\phi) \in [T]\times\mathcal{V}_\rho(\mathcal{H})$ we have

$$\begin{aligned}&\Big|\sum_{i=1}^t X_{i,f_i}(\phi) - \sum_{i=1}^t \|\mathbb{E}_{\zeta_i}\big[l_{f_i}(\phi,\phi,\zeta_i)\big]\|_2^2\Big| \\ &\quad\le \mathcal{O}\left(\sqrt{\zeta\cdot \mathrm{sp}(V^*)^2 \sum_{i=1}^t \|\mathbb{E}_{\zeta_i}\big[l_{f_i}(\phi,\phi,\zeta_i)\big]\|_2^2 + \zeta}\right).\end{aligned} \tag{D.8}$$

The following analysis assumes that the event above is true. Recall that the Loop ensures that

$$\begin{aligned}\sum_{i=1}^{t-1} X_{i,f_i}(f_t) &= \sum_{i=1}^{t-1}\|l_{f_i}(f_t,f_t,\zeta_i)\|_2^2 - \sum_{i=1}^{t-1}\|l_{f_i}\big(f_t,\mathcal{P}(f_t),\zeta_i\big)\|_2^2, \\ &\le \sum_{i=1}^{t-1}\big\|l_{f_i}\big(f_t,f_t,\zeta_i\big)\big\|_2^2 - \inf_{g\in\mathcal{G}}\sum_{i=1}^{t-1}\big\|l_{f_i}\big(f_t,g,\zeta_i\big)\big\|_2^2, \\ &= \mathcal{L}_{\mathcal{D}_{t-1}}(f_t,f_t) - \inf_{g\in\mathcal{G}}\mathcal{L}_{\mathcal{D}_{t-1}}(f_t,g) \le \mathcal{O}(\beta),\end{aligned} \tag{D.9}$$

where the last inequality is based on the confidence set and the update condition combined. Note that if the update is executed at time t, the confidence set ensures that

$$\mathcal{L}_{\mathcal{D}_{t-1}}(f_t,f_t) - \inf_{g\in\mathcal{G}}\mathcal{L}_{\mathcal{D}_{t-1}}(f_t,g) \le \beta,$$

within the update step t. Otherwise, if the update condition is not triggered, we have $f_{\tau_t} = f_t$ and

$$\Upsilon_{t-1} = \mathcal{L}_{\mathcal{D}_{t-1}}(f_t,f_t) - \inf_{g\in\mathcal{G}}\mathcal{L}_{\mathcal{D}_{t-1}}(f_t,g) \le 4\beta.$$

Recall that based on the definition of ρ-cover for any $f\in\mathcal{H}$, there exists $\phi\in\mathcal{V}_\rho(\mathcal{H})$ such that $\|g(s,a)-\phi(s,a)\|_1 \le \rho$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have the in-sample training error is bounded by

$$\begin{aligned}\sum_{i=1}^{t-1}\big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f_t,f_t,\zeta_i)\big]\big\|_2^2 &\le \sum_{i=1}^{t-1}\big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(\phi_t,\phi_t,\zeta_i)\big]\big\|_2^2 + \mathcal{O}(t\rho), && (\rho\text{-approximation}) \\ &= \sum_{i=1}^{t-1} X_{i,f_i}(\phi_t) + \mathcal{O}(t\rho+\zeta) && (\,(\text{D.8})) \\ &= \sum_{i=1}^{t-1} X_{i,f_i}(f_t) + \mathcal{O}(t\rho+\zeta) \le \mathcal{O}(T\rho+\zeta+\beta) = \mathcal{O}(\beta), && (\text{D.10})\end{aligned}$$

where the last inequality follows (D.9), and takes $\beta = c\big((\mathrm{sp}(V^*)\log\big(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(\rho)/\delta\big) + T\rho\big)$. □

D.4 PROOF OF LEMMA 11

Lemma 11. Let $\mathcal{N}(T)$ be the switching cost with time horizon T, defined as

$$\mathcal{N}(T) = \#\{t \in [T] : \tau_t \neq \tau_{t-1}\}.$$

Given fixed $\rho > 0$ and the optimistic parameter $\beta = c\big(\mathrm{sp}(V^*)\log\big(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(\rho)/\delta\big) + T\rho\big)$, where $c > 0$ is large enough constant, then with probability greater than $1-2\delta$ we have

$$\mathcal{N}(T) \leq \mathcal{O}\big(\kappa_{\mathrm{G}}\log T + \beta^{-1}T\epsilon^2\big),$$

where κ_{G} is the transferability coefficient with respect to $\mathrm{AGEC}(\mathcal{H}, \{l_{f'}\}, \epsilon)$.

Proof of Lemma 11. Denote $\mathcal{V}_\rho(\mathcal{H})$ be the ρ-cover of $\mathcal{H}$ and $\mathcal{N}_{\mathcal{H}}(\rho)$ be the size of ρ-cover $\mathcal{V}_\rho(\mathcal{H})$.

Step 1: Bound the difference of discrepancy between the minimizer and $\mathcal{P}(f)$.

Consider fixed tuple $(i, f, g) \in [T] \times \mathcal{H} \times \mathcal{G}$ and define auxiliary function as

$$X_{i,f_i}(f,g) := \big\|l_{f_i}(f,g,\zeta_i)\big\|_2^2 - \big\|l_{f_i}(f,\mathcal{P}(f),\zeta_i)\big\|_2^2$$

Let $\mathscr{F}_t$ be the filtration induced by $\{s_1, a_1, \ldots, s_t, a_t\}$ and note that $f_1, \ldots, f_t$ is fixed under the filtration, then we have

$$\begin{aligned}
\mathbb{E}[X_{i,f_i}(f,g)|\mathscr{F}_i] &= \mathbb{E}_{\zeta_i}[\big\|l_{f_i}(f,g,\zeta_i)\big\|_2^2 - \big\|l_{f_i}(f,\mathcal{P}(f),\zeta_i)\big\|_2^2|\mathscr{F}_i] \\
&= \mathbb{E}_{\zeta_i}\Big[\big[l_{f_i}(f,g,\zeta_i) - l_{f_i}(f,\mathcal{P}(f),\zeta_i)\big]\cdot\big[l_{f_i}(f,g,\zeta_i) + l_{f_i}(f,\mathcal{P}(f),\zeta_i)\big]\Big|\mathscr{F}_i\Big] \\
&= \mathbb{E}_{\zeta_i}\big[l_{f_i}(f,g,\zeta_i)\big]\cdot\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,g,\zeta_i) + l_{f_i}(f,\mathcal{P}(f),\zeta_i)\big|\mathscr{F}_i\big] \\
&= \big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,g,\zeta_i)\big]\big\|_2^2,
\end{aligned}$$

where the equation generalized completeness (see Lemma 9). Similarly, we can obtain that the second moment of the auxiliary function is bounded by

$$\mathbb{E}[X_{i,f_i}(f,g)^2|\mathscr{F}_i] \leq \mathcal{O}\Big(\mathrm{sp}(V^*)^2\,\big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,g,\zeta_i)\big]\big\|_2^2\Big),$$

By Freedman's inequality in Lemma 18, with probability greater than $1-\delta$

$$\begin{aligned}
&\Big|\sum_{i=1}^t X_{i,f_i}(f,g) - \sum_{i=1}^t \big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,g,\zeta_i)\big]\big\|_2^2\Big| \\
&\leq \mathcal{O}\left(\sqrt{\log(1/\delta)\cdot\mathrm{sp}(V^*)^2\sum_{i=1}^t\big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(f,g,\zeta_i)\big]\big\|_2^2} + \log(1/\delta)\right)
\end{aligned}$$

Define $\zeta = \mathrm{sp}(V^*)\log(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(\rho)/\delta)$, by taking a union bound over ρ-covering of hypothesis set $\mathcal{H}\times\mathcal{G}$, with probability greater than $1-\delta$, for all $(t,\phi,\varphi) \in [T]\times\mathcal{V}_\rho(\mathcal{H})\times\mathcal{V}_\rho(\mathcal{G})$ it holds

$$\begin{aligned}
&\Big|\sum_{i=1}^t X_{i,f_i}(\phi,\psi) - \sum_{i=1}^t \big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(\phi,\psi,\zeta_i)\big]\big\|_2^2\Big| \\
&\leq \mathcal{O}\left(\sqrt{\zeta\cdot\mathrm{sp}(V^*)^2\sum_{i=1}^t\big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(\phi,\psi,\zeta_i)\big]\big\|_2^2} + \zeta\right),
\end{aligned} \tag{D.11}$$

where $\zeta = \mathrm{sp}(V^*)\log(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(\rho)/\delta)$. Note that $\big\|\mathbb{E}_{\zeta_i}\big[l_{f_i}(\phi,\psi,\zeta_i)\big]\big\|_2^2$ is non-negative, then it holds that $-\sum_{i=1}^t X_{i,f_i}(\phi,\varphi) \leq \mathcal{O}(\zeta)$ for all $t \in [T]$. Based on (D.11) and the ρ-approximation, we have

$$-\sum_{i=1}^t X_{i,f_i}(f,g) \leq \mathcal{O}\Big(\zeta + t\rho\Big), \qquad \forall t \in [T],$$

for any $(f, g) \in \mathcal{H} \times \mathcal{G}$. Recall that $\beta = c \log(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(\rho)/\delta)\mathrm{sp}(V^*)$, for all $t \in [T]$ we have

$$\mathcal{L}_{\mathcal{D}_t}(f_t, \mathcal{P}(f_t)) - \inf_{g\in\mathcal{G}} \mathcal{L}_{\mathcal{D}_t}(f_t, g) = \sum_{i=1}^{t} \|l_{f_i}(f_t, \mathcal{P}(f_t), \zeta_i)\|_2^2 - \inf_{g\in\mathcal{G}} \sum_{i=1}^{t} \|l_{f_i}(f_t, g, \zeta_i)\|_2^2$$
$$= -\sum_{i=1}^{t} X_{i,f_i}(f_t, \tilde{g}) \le \mathcal{O}\Big(\zeta + t\rho\Big) \le \beta. \tag{D.12}$$

Combine (D.12), and the fact that g is defined as the local minimizer among auxiliary class $\mathcal{G}$ and $\mathcal{P}(f_t) \in \mathcal{G}$, then for all $t \in [T]$ we have the difference of discrepancy bounded by

$$0 \le \mathcal{L}_{\mathcal{D}_t}(f_t, \mathcal{P}_{J_t}(f_t)) - \inf_{g\in\mathcal{G}} \mathcal{L}_{\mathcal{D}_t}(f_t, g) \le \beta. \tag{D.13}$$

Step 2: Bound the out-sample training error between updates.

Consider an update is executed at step $t+1$, it directly implies that $\mathcal{L}_{\mathcal{D}_t}(f_t, f_t) - \inf_{g\in\mathcal{G}} \mathcal{L}_{\mathcal{D}_t}(f_t, g) > 4\beta$, while the latest update at step τ_t ensures that $\mathcal{L}_{\mathcal{D}_{\tau_t-1}}(f_{\tau_t}, f_{\tau_t}) - \inf_{g\in\mathcal{G}} \mathcal{L}_{\mathcal{D}_{\tau_t-1}}(f_{\tau_t}, g) \le \beta$, where τ_t is the pointer of the lastest update. Combined the results above with (D.13), we have

$$\mathcal{L}_{\mathcal{D}_t}(f_t, f_t) - \mathcal{L}_{\mathcal{D}_t}\big(f_t, \mathcal{P}(f_t)\big) > 3\beta, \quad \mathcal{L}_{\mathcal{D}_{\tau_t-1}}(f_{\tau_t}, f_{\tau_t}) - \mathcal{L}_{\mathcal{D}_{\tau_t-1}}\big(f_{\tau_t}, \mathcal{P}(f_{\tau_t})\big) \le \beta. \tag{D.14}$$

It indicates that the sum of squared empirical discrepancy between two adjacent updates follows

$$\sum_{i=\tau_t}^{t} \|l_{f_t}(f_t, f_t, \zeta_t)\|_2^2 = \mathcal{L}_{\mathcal{D}_{\tau_t:t}}(f_t, f_t) - \mathcal{L}_{\mathcal{D}_{\tau_t:t}}(f_t, \mathcal{P}_{J_t}(f_t)) > 2\beta, \tag{D.15}$$

where denote $\mathcal{D}_{\tau_t:t} = \mathcal{D}_t/\mathcal{D}_{\tau_t}$. Based on the similar concentration arguments as Lemma 10, we have the out-sample training error between updates is bounded by $\sum_{i=\tau_t}^{t} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_i, f_i, \zeta_i)]\|_2^2 > \beta$.

Step 3: Bound the switching cost under the transferability constraint.

Denote $b_1, \dots, b_{\mathcal{N}(T)}, b_{\mathcal{N}(T)+1}$ be the sequence of updated steps such that $\tau_t \in \{b_t\}$ for all $t \in [T]$, and we fix the recorder $b_1 = 1$ and $b_{\mathcal{N}(T)+1} = T + 1$. Note that based on (D.15), the sum of out-sample training error shall have a lower bound such that

$$\sum_{t=1}^{T} \|\mathbb{E}_{\zeta_t}[l_{f_t}(f_t, f_t, \zeta_t)]\|_2^2 = \sum_{u=1}^{\mathcal{N}(T)} \sum_{t=b_u}^{b_{u+1}-1} \|\mathbb{E}_{\zeta_t}[l_{f_t}(f_t, f_t, \zeta_t)]\|_2^2 \ge \mathcal{N}(T) \cdot \beta. \tag{D.16}$$

Besides, note that the in-sample training error $\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_t}\left[l_{f_i}(f_t, f_t, \zeta_t)\right]\|_2^2 \le \mathcal{O}\big(\beta\big)$ for all $t \in [T]$ and based on the definition of transferability coefficient κ_{G} (see Definition 3), we have

$$\sum_{t=1}^{T} \|\mathbb{E}_{\zeta_t}[l_{f_t}(f_t, f_t, \zeta_t)]\|_2^2 \le \mathcal{O}\left(\kappa_{\mathrm{G}} \cdot \beta \log T + \mathrm{sp}(V^*)^2 \min\{\kappa_{\mathrm{G}}, T\} + T\epsilon^2\right) \tag{D.17}$$

Combine (D.16) and (D.17), it holds $\mathcal{N}(T) \le \mathcal{O}\big(\kappa_{\mathrm{G}} \log T + \beta^{-1} T \log T \epsilon^2\big)$ and finish the proof. □

E Proof of Results about Complexity Measures

In this section, we provide the proof of results about the complexity metrics in Section 3. We remark that the proof highly relies on Lemma 13 and Lemma 14, which are natural extentions to original results in Jin et al. (2020); Zhong et al. (2022) and proofs are provided in Section G.1.

E.1 Proof of Lemma 1

Proof of Lemma 1. Recall that the eluder dimension is defined over the function class following

$$\mathcal{X}_{\mathcal{H}} := \big\{X_{f,f'}(s,a) = \big(r_f + \mathbb{P}_{f'} V_f\big)(s,a) : f, f' \in \mathcal{H}\big\},$$

and for model-based problems, the discrepancy function is chosen as

$$l_{f'}(f, g, \zeta_t) = \big(r_g + \mathbb{P}_g V_{f'}\big)(s_t, a_t) - r(s_t, a_t) - V_{f'}(s_{t+1}).$$

Step 1: Bound over transferability coefficient.

Start with the transferability coefficient, the condition can be equivalently written as

$$\begin{aligned}\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\|_2^2 &= \sum_{i=1}^{t-1} \|\big(r_{f_t} + \mathbb{P}_{f_t} V_{f_t} - r_{f^*} + \mathbb{P}_{f^*} V_{f_t}\big)(s_i, a_i)\|_2^2 \\ &= \sum_{i=1}^{t-1} \big(X_{f_t,f_t} - X_{f_t,f^*}\big)(s_i, a_i)^2 \leq \beta, \quad \forall t \in [T]. \qquad \text{(E.1)}\end{aligned}$$

Let $\breve{\mathcal{X}}_{\mathcal{H}} = \{f - f' : f, f' \in \mathcal{X}_{\mathcal{H}}\}$, the generalized pigeon-hole principle (see Lemma 13) indicates that if we take $\Gamma = \mathcal{D}_\Delta$, $\phi_t = X_{f_t,f_t} - X_{f_t,f^*}$, $\Phi = \breve{\mathcal{X}}_{\mathcal{H}}$, $\|\phi_t\|_\infty \leq \mathrm{sp}(V^*) + 2$, then it holds that

$$\begin{aligned}\sum_{i=1}^{t} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_i, f_i, \zeta_i)]\|^2 &= \sum_{i=1}^{t} \big(X_{f_i,f_i} - X_{f_i,f^*}\big)(s_i, a_i)^2 \\ &\leq \dim_{\mathrm{DE}}(\breve{\mathcal{X}}_{\mathcal{H}}, \mathcal{D}_\Delta, \epsilon) \cdot \beta \log t + (\mathrm{sp}(V^*) + 2)^2 \min\{\dim_{\mathrm{DE}}(\breve{\mathcal{X}}_{\mathcal{H}}, \mathcal{D}_\Delta, \epsilon), t\} + t\epsilon^2 \\ &= \dim_{\mathrm{E}}(\mathcal{X}_{\mathcal{H}}, \epsilon) \cdot \beta \log t + (\mathrm{sp}(V^*) + 2)^2 \min\{\dim_{\mathrm{E}}(\mathcal{X}_{\mathcal{H}}, \epsilon), t\} + t\epsilon^2, \qquad \text{(E.2)}\end{aligned}$$

given condition that (E.1) holds for all $t \in [T]$, where the last equation uses $\dim_{\mathrm{DE}}(\breve{\mathcal{X}}_{\mathcal{H}}, \mathcal{D}_\Delta, \epsilon) = \dim_{\mathrm{E}}(\mathcal{X}_{\mathcal{H}}, \epsilon)$. Denote $d_{\mathrm{E}} = \dim_{\mathrm{E}}(\mathcal{X}_{\mathcal{H}}, \epsilon)$, then we have $\kappa_{\mathrm{G}} \leq d_{\mathrm{E}}$ based on (E.2).

Step 2: Bound over dominance coefficient.

Based on Lemma 14 and $\mathcal{E}(f_t)(s_t, a_t) = \mathbb{E}_{\zeta_t}\big[l_{f_t}(f_t, f_t, \zeta_t)\big]$ based on definition, it holds that

$$\begin{aligned}\sum_{t=1}^{T} \|\mathbb{E}_{\zeta_t}\big[l_{f_t}(f_t, f_t, \zeta_t)\big]\|_2^2 &= \sum_{t=1}^{T} \big[\big(X_{f_t,f_t} - X_{f_t,f^*}\big)(s_t, a_t)\big]^2 \\ &\leq \left[2 d_{\mathrm{DE}} \log T \sum_{t=1}^{T} \sum_{i=1}^{t-1} \big[\big(X_{f_t,f_t} - X_{f_t,f^*}\big)(s_i, a_i)\big]^2\right]^{1/2} + (\mathrm{sp}(V^*) + 2) \min\{d_{\mathrm{DE}}, T\} + T\epsilon \\ &= \left[2 d_{\mathrm{E}} \log T \sum_{t=1}^{T} \sum_{i=1}^{t-1} \big\|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\big\|_2^2\right]^{1/2} + (\mathrm{sp}(V^*) + 2) \min\{d_{\mathrm{E}}, T\} + T\epsilon, \qquad \text{(E.3)}\end{aligned}$$

by taking $\Gamma = \mathcal{D}_\Delta$, $\phi_t = X_{f_t,f_t} - X_{f_t,f^*}$, $\Phi = \breve{\mathcal{X}}_{\mathcal{H}}$, $\|\phi_t\|_\infty \leq \mathrm{sp}(V^*) + 2$, and $1 + \log T \leq 2 \log T$. □

E.2 Proof of Lemma 2

Proof of Lemma 2. Consider the Bellman discrepancy function, defined as

$$l_{f'}(f, g, \zeta_t) = Q_g(s_t, a_t) - r(s_t, a_t) - V_f(s_{t+1}) + J_g,$$

and the expectation is taken over s_{i+1} from $\mathbb{P}(\cdot|s_i, a_i)$ such that $\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)] = \mathcal{E}(f_t)(s_i, a_i)$.

Step 1: Bound over transferability coefficient.

First, we're going to demonstrate the transferability. Note that the generalized pigeon-hole principle (see Lemma 13) directly indicates that, given

$$\sum_{i=1}^{t-1} \|\mathcal{E}(f_t)(s_i, a_i)\|_2^2 = \sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\|_2^2 \leq \beta, \quad \forall t \in [T],$$

if we take $\phi_t = \mathcal{E}(f_t)$, $\Phi = \mathcal{E}_{\mathcal{H}}$ and $\Gamma = \mathcal{D}_\Delta$, then for all $t \in [T]$ we have

$$\sum_{i=1}^{t} \|\mathcal{E}(f_i)(s_i, a_i)\|_2^2 \leq d_{\mathrm{ABE}} \cdot \beta \log t + (\mathrm{sp}(V^*) + 2)^2 \min\{d_{\mathrm{ABE}}, t\} + t\epsilon^2, \qquad \text{(E.4)}$$

and thus we upper bound $\kappa_{\mathrm{G}} \leq d_{\mathrm{ABE}} := \dim_{\mathrm{ABE}}(\mathcal{H}, \epsilon)$.

Step 2: Bound over dominance coefficient.

Based on Lemma 14 and $\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)] = \mathcal{E}(f_t)(s_i, a_i)$, it holds that

$$\sum_{t=1}^{T} \mathcal{E}(f_t)(s_t, a_t) \leq \left[2d_{\mathrm{ABE}} \log T \sum_{t=1}^{T} \sum_{i=1}^{t-1} \|\mathcal{E}(f_t)(s_i, a_i)\|_2^2\right]^{1/2} + (\mathrm{sp}(V^*) + 2) \min\{d_{\mathrm{ABE}}, T\} + T\epsilon,$$

where we take $\phi_t = \mathcal{E}(f_t)$, $\Phi = \mathcal{E}_{\mathcal{H}}$, $\Gamma = \mathcal{D}_\Delta$, $\|\mathcal{E}(f_t)\|_\infty \leq \mathrm{sp}(V^*) + 2$, and $1 + \log T \leq 2 \log T$.$\square$

F Proof of Results for Concrete Examples

In this section, we provide detailed proofs of results for concrete examples in Appendix C.

F.1 Proof of Proposition 5

Proof of Proposition 5. To show that linear FA has low AGEC, we first prove that it is captured by ABE dimension $\dim_{\mathrm{ABE}}(\mathcal{H}, \epsilon)$, and then apply Lemma 2 (low ABE dim $\subseteq$ low AGEC). Suppose that there exists $\epsilon' \geq \epsilon$, $\{\delta_{s_i,a_i}\}_{i\in[m]} \subseteq \mathcal{D}_\Delta$, and $\{f_i\}_{i\in[m]} \subseteq \mathcal{H}$ with length $m \in \mathbb{N}$, such that

$$\sqrt{\sum_{i=1}^{t-1} \big[\mathcal{E}(f_t)(s_i, a_i)\big]^2} \leq \epsilon', \quad \big|\mathcal{E}(f_t)(s_t, a_t)\big| > \epsilon', \quad \forall t \in [m]. \tag{F.1}$$

Based on Definitions 7-8, $\dim_{\mathrm{ABE}}(\mathcal{H}, \epsilon)$ is the largest m. Following this, we provide detailed discussion about linear AMDPs, AMDPs with linear Bellmen completion, and AMDPs with generalized linear completion (see Definition 11-13). Denote $Q_t(s, a) = \phi(s, a)^\top \omega_t$ for all $(s, a, t) \in \mathcal{S} \times \mathcal{A} \times [m]$.

(i). Linear AMDPs. As defined in Definition 11, for any $f_t \in \mathcal{H}$, it holds that

$$\begin{aligned}\mathcal{E}(f_t)(s_i, a_i) &= \phi(s_i, a_i)^\top \omega_t - \phi(s_i, a_i)^\top \theta - \phi(s_i, a_i)^\top \int_S V_t(s) \mathrm{d}\mu(s) + J_t \\ &= \phi(s_i, a_i)^\top \left(\omega_t - \theta + \int_S V_t(s) \mathrm{d}\mu(s) + J_t \mathbf{e}_1\right).\end{aligned} \tag{F.2}$$

(ii). AMDPs with linear Bellmen completion. As a natural extension to the linear AMDPs, linear Bellmen completeness (see Definition 12) suggests that the Bellman error follows

$$\mathcal{E}(f_t)(s_i, a_i) = \phi(s_i, a_i)^\top \omega_t - \phi(s_i, a_i)^\top \mathcal{T}(\omega_t, J_t) = \phi(s_i, a_i)^\top (\omega_t - \mathcal{T}(\omega_t, J_t)). \tag{F.3}$$

(iii). AMDPs with generalized linear completion. Moreover, AMDPs with generalized linear completion further extends the standard linear FA by introducing link functions. Note that

$$\mathcal{E}(f_t)(s_i, a_i) = \sigma\big(\phi(s_i, a_i)^\top \omega_t\big) - \sigma\big(\phi(s_i, a_i)^\top \mathcal{T}(\omega_t, J_t)\big) \tag{F.4}$$

and based on the α-bi-Lipschitz continuity condition in (C.3), it holds that

$$\frac{1}{\alpha} \cdot \phi(s_i, a_i)^\top (\omega_t - \mathcal{T}(\omega_t, J_t)) \leq \mathcal{E}(f_t)(s_i, a_i) \leq \alpha \cdot \phi(s_i, a_i)^\top (\omega_t - \mathcal{T}(\omega_t, J_t)). \tag{F.5}$$

Based on the Lemma 15 and arguments in (i), (ii) and(iii), we are ready to provide a unified proof for linear FA. By substituting the arguments (F.2), (F.3) and (F.5) into (F.1), then

$$\sqrt{\sum_{i=1}^{t-1} [\langle \phi(s_i, a_i), \omega_t - \mathcal{T}(\omega_t, J_t)\rangle]^2} \leq \alpha\epsilon', \quad |\langle \phi(s_t, a_t), \omega_t - \mathcal{T}(\omega_t, J_t)\rangle| > \frac{\epsilon'}{\alpha}, \quad \forall t \in [m]. \tag{F.6}$$

Here, we take $\alpha = 1$ for standard linear FA , and let α be the Lipschitz constant for generalized linear FA. Based on Lemma 15, if we take $\phi_t = \phi(s_t, a_t)$, $\psi_t = \omega_t - \mathcal{T}(\omega_t, J_t)$, $B_\phi = \sqrt{2}$, $B_\psi = \mathrm{sp}(V^*)\sqrt{d}$, $\varepsilon = \epsilon$, $c_1 = \alpha$, $c_2 = \alpha^{-1}$, then $m \leq \mathcal{O}\big(d \log(\mathrm{sp}(V^*)\sqrt{d}/\epsilon)\big)$. As the ABE dimension is defined as the length of the longest sequence satisfying (F.6), thus

$$\dim_{\mathrm{ABE}}(\mathcal{H}, \epsilon) \leq \mathcal{O}\big(d \log\big(\mathrm{sp}(V^*)\sqrt{d}/\epsilon\big)\big).$$

Based on Lemma 2, $d_{\mathrm{G}} \leq \mathcal{O}\big(d \log\big(\mathrm{sp}(V^*)\sqrt{d}\epsilon^{-1}\big) \log T\big)$ and $\kappa_{\mathrm{G}} \leq \mathcal{O}\big(d \log\big(\mathrm{sp}(V^*)\sqrt{d}\epsilon^{-1}\big)\big)$.$\square$

F.2 Proof of Proposition 6

Proof of Proposition 6. For all $f_t \in \mathcal{H}$, the Bellman error can be written as

$$\begin{aligned}\mathcal{E}(f_t)(s_i,a_i) &= \phi(s_i,a_i)^\top \omega_t - \mathbb{E}[\psi(s_{i+1})]^\top \theta_t + J_t - r(s_i,a_i) \\ &= \phi(s_i,a_i)^\top \omega_t - \mathbb{E}[\psi(s_{i+1})]^\top \theta_t + J_t - \left(Q^*(s_i,a_i) - \mathbb{E}[V^*(s_{i+1})] - J^*\right) \\ &= \begin{bmatrix}\phi(s_i,a_i) \\ \mathbb{E}[\psi(s_{i+1})]\end{bmatrix}^\top \left(\begin{bmatrix}\omega_t - \omega^* \\ \theta^* - \theta_t\end{bmatrix} + (J_t - J^*)\cdot \mathbf{e}_1\right),\end{aligned} \tag{F.7}$$

where the second equation results from Bellman optimality equation in (2.1). Following a similar argument in the proof of Proposition 5, we can show that the linear Q^*/V^* AMDPs have a low ABE dimension with an (d_1+d_2)-dimensional compound feature mapping equivalently based on (F.7). Based on Lemma 2 and write $d^+ = d_1 + d_2$, then we have

$$d_{\mathrm{G}} \le \mathcal{O}\big(d^+ \log\big(\mathrm{sp}(V^*)\sqrt{d^+}\epsilon^{-1}\big)\log T\big), \quad \kappa_{\mathrm{G}} \le \mathcal{O}\big(d^+ \log\big(\mathrm{sp}(V^*)\sqrt{d^+}\epsilon^{-1}\big)\big). \qquad \square$$

F.3 Proof of Proposition 7

Proof of Proposition 7. Similar to linear FA, we will show that kernel FA is captured by by ABE dimension $\dim_{\mathrm{ABE}}(\mathcal{H},\epsilon)$, and then apply Lemma 2 (low ABE dim $\subseteq$ low AGEC). Suppose that there exists $\epsilon' \ge \epsilon$, $\{\delta_{s_i,a_i}\}_{i\in[m]} \subseteq \mathcal{D}_\Delta$, and $\{f_i\}_{i\in[m]} \subseteq \mathcal{H}$ with length $m \in \mathbb{N}$, such that

$$\sqrt{\sum_{i=1}^{t-1}\big[\mathcal{E}(f_t)(s_i,a_i)\big]^2} \le \epsilon', \quad \big|\mathcal{E}(f_t)(s_t,a_t)\big| > \epsilon', \quad \forall t \in [m]. \tag{F.8}$$

Suppose the kernel function class has a finite ϵ-effective dimension concerning the feature mapping ϕ. The existence of Bellman error $\mathcal{E}(f_t)$ is equivalent to the one of $W_t \in (\mathcal{W} - \mathcal{W})$:

$$\mathcal{E}(f_t)(\cdot,\cdot) = (Q_{f_t} - \mathcal{T}_{J_t}Q_{f_t})(\cdot,\cdot) = \langle \phi(\cdot,\cdot), \omega_t - \omega_t'\rangle_{\mathcal{K}} := \langle \phi(\cdot,\cdot), W_t\rangle_{\mathcal{K}}, \tag{F.9}$$

where the second equation is based on the self-completeness assumption with kernel FA such that $\mathcal{G} = \mathcal{H}$. Denote $X_t = \phi(s_t,a_t)$, we can rewrite the condition in (F.8) as

$$\sqrt{\sum_{i=1}^{t-1}(X_i^\top W_t)^2} \le \epsilon', \quad |X_t^\top W_t| > \epsilon', \quad \forall t \in [m]. \tag{F.10}$$

Let $\Sigma_t = \sum_{i=1}^{t-1} X_i X_i^\top + (\epsilon'^2/4R^2\cdot \mathrm{sp}(V^*)^2)\cdot \mathbf{I}$, then $\|W_t\|_{\Sigma_t} \le \sqrt{2}\epsilon'$ and $\epsilon' \le \|W_t\|_{\Sigma_t}\|X_t\|_{\Sigma_t^{-1}}$ for all $t\in[m]$ based on Cauchy-Swartz inequlity and $\|\omega_t\|_{\mathcal{K}} \le \mathrm{sp}(V^*)R$. Thus, $\|X_t\|^2_{\Sigma_t^{-1}} \ge 0.5$ and

$$\sum_{t=1}^m \log\left(1 + \|X_t\|^2_{\Sigma_t^{-1}}\right) = \log\left(\frac{\det \Sigma_{m+1}}{\det \Sigma_1}\right) = \log\det\left[\mathbf{I} + \frac{4R^2\mathrm{sp}(V^*)^2}{\epsilon'^2}\sum_{t=1}^m X_t X_t^\top\right], \tag{F.11}$$

based on the matrix determinant lemma. Therefore, (F.10) directly implies that

$$\frac{1}{e} \le \log\frac{3}{2} \le \frac{1}{m}\log\det\left[\mathbf{I} + \frac{4R^2\mathrm{sp}(V^*)^2}{\epsilon'^2}\sum_{t=1}^m X_t X_t^\top\right],$$

and then we have $m \le \dim_{\mathrm{eff}}\big(\mathcal{X}, \epsilon/2\mathrm{sp}(V^*)R\big)$.Recall that the ϵ-effective dimension is the minimum positive integer satisfying the condition. As ABE dimension is defined as the length of the longest sequence satisfying (F.10), thus it holds that

$$\dim_{\mathrm{ABE}}(\mathcal{H},\epsilon) \le \dim_{\mathrm{eff}}\big(\mathcal{X}, \epsilon/2\mathrm{sp}(V^*)R\big).$$

Based on Lemma 2, $d_{\mathrm{G}} \le \dim_{\mathrm{eff}}\big(\mathcal{X}, \epsilon/2\mathrm{sp}(V^*)R\big)\log T$ and $\kappa_{\mathrm{G}} \le \dim_{\mathrm{eff}}\big(\mathcal{X}, \epsilon/2\mathrm{sp}(V^*)R\big)$. $\square$

F.4 PROOF OF PROPOSITION 8

Proof of Proposition 8. Note that expected discrepancy function follows: for any $t \in [T]$

$$\|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\|_2 = \theta_t^\top \left(\psi(s_i, a_i) + \int_{\mathcal{S}} \phi(s_i, a_i, s') V_{f_i}(s') \mathrm{d}s'\right) - r(s_i, a_i) - \mathbb{E}_{\zeta_i}[V_{f_i}(s_{i+1})]$$
$$= (\theta_t - \theta^*)^\top \left(\psi(s_i, a_i) + \int_{\mathcal{S}} \phi(s_i, a_i, s') V_{f_i}(s') \mathrm{d}s'\right). \tag{F.12}$$

Let $W_t = \theta_t - \theta^*$, $X_t = \psi(s_i, a_i) + \int_{\mathcal{S}} \phi(s_i, a_i, s') V_{f_i}(s') \mathrm{d}s'$, and $\Sigma_t = \epsilon \mathbf{I} + \sum_{i=1}^{t-1} X_t X_t^\top$. Note

$$\|W_t\|_{\Sigma_t} = \left[\epsilon \|\theta_t - \theta^*\|_2^2 + \sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\|_2^2\right]^{1/2} \leq 2\sqrt{\epsilon} + \left[\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\|_2^2\right]^{1/2} \tag{F.13}$$

where we use $\|\theta_t\| \leq 1$. Based on the elliptical potential lemma (see Lemma 17), we have

$$\sum_{t=1}^T \|X_t\|_{\Sigma_t^{-1}} \wedge 1 \leq \sum_{t=1}^T 2d \cdot \log\left(1 + \frac{1}{d} \sum_{t=1}^T \|X_t\|_2\right)$$
$$\leq 2d \cdot \log\left(1 + (1 + \mathrm{sp}(V^*)/2) \cdot T \left(\sqrt{d}\epsilon\right)^{-1}\right) := d(\epsilon), \tag{F.14}$$

where the last inequality results from $\|\phi\|_{2,\infty} \leq \sqrt{d}$, $\|\psi\|_{2,\infty} \leq \sqrt{d}$ and $\|V_f\|_\infty \leq \frac{1}{2}\mathrm{sp}(V^*)$. Combine (F.13) and $\mathbb{1}\left(\|X_t\|_{\Sigma_t^{-1}} \geq 1\right) \leq \|X_t\|_{\Sigma_t^{-1}} \wedge 1$, it holds that

$$\sum_{t=1}^T \mathbb{1}\left(\|X_t\|_{\Sigma_t^{-1}} \geq 1\right) \leq \sum_{t=1}^T \|X_t\|_{\Sigma_t^{-1}} \wedge 1 \leq d(\epsilon). \tag{F.15}$$

Step 1: Bound over dominance coefficient.

Note the sum of Bellman errors follows that

$$\sum_{t=1}^T \mathcal{E}(f_t)(s_t, a_t) = \sum_{t=1}^T \left((r_{f_t} + \mathbb{P}_{f_t} V_{f_t})(s_t, a_t) - (r_{f^*} + \mathbb{P}_{f^*} V_{f_t})(s_t, a_t)\right)$$
$$= \sum_{t=1}^T (\theta_t - \theta^*)^\top \left(\psi(s_t, a_t) + \int_{\mathcal{S}} \phi(s_t, a_t, s') V_{f_t}(s') \mathrm{d}s'\right)$$
$$= \sum_{t=1}^T W_t^\top X_t \cdot \left(\mathbb{1}\left(\|X_t\|_{\Sigma_t^{-1}} \leq 1\right) + \mathbb{1}\left(\|X_t\|_{\Sigma_t^{-1}} > 1\right)\right)$$
$$\leq \sum_{t=1}^T W_t^\top X_t \cdot \mathbb{1}\left(\|X_t\|_{\Sigma_t^{-1}} \leq 1\right) + (\mathrm{sp}(V^*) + 2) \cdot \min\{d(\epsilon), T\}$$
$$\leq \sum_{t=1}^T \|W_t\|_{\Sigma_t} \cdot \left(\|X_t\|_{\Sigma_t^{-1}} \wedge 1\right) + (\mathrm{sp}(V^*) + 2) \cdot \min\{d(\epsilon), T\}, \tag{F.16}$$

where the first inequality results from (F.15) and the last inequality arises from the Cauchy-Swartz inequality. Combine (F.13) and (F.14), we have

$$\sum_{t=1}^T \|W_t\|_{\Sigma_t} \cdot \left(\|X_t\|_{\Sigma_t^{-1}} \wedge 1\right) \leq \sum_{t=1}^T \left(2\sqrt{\epsilon} + \left[\sum_{i=1}^{t-1} \|l_{f_i}(f_t, f_t, \zeta_i)\|_2^2\right]^{1/2}\right) \cdot \left(\|X_t\|_{\Sigma_t^{-1}} \wedge 1\right)$$
$$\leq \left[\sum_{t=1}^T 4\epsilon\right]^{1/2} \left[\sum_{t=1}^T \|X_t\|_{\Sigma_t^{-1}} \wedge 1\right]^{1/2} + \left[\sum_{t=1}^T \sum_{i=1}^{t-1} \|l_{f_i}(f_t, f_t, \zeta_i)\|_2^2\right]^{1/2} \left[\sum_{t=1}^T \|X_t\|_{\Sigma_t^{-1}} \wedge 1\right]^{1/2}$$
$$\leq 2\sqrt{T\epsilon \cdot \min\{d(\epsilon), T\}} + \left[d(\epsilon) \sum_{t=1}^T \sum_{i=1}^{t-1} \|l_{f_i}(f_t, f_t, \zeta_i)\|_2^2\right]^{1/2}, \tag{F.17}$$

where the second inequality results from Cauchy-Swartz inequality and the last inequality follows (F.14). Plugging the result back into the (F.16), we conclude that

$$
\begin{aligned}
\sum_{t=1}^{T} \mathcal{E}(f_t)(s_t,a_t) &\le \left[d(\epsilon) \sum_{t=1}^{T}\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t,f_t,\zeta_i)]\|_2^2 \right]^{1/2} \\
&\quad + 2\sqrt{T\epsilon \cdot \min\{d(\epsilon),T\}} + (\mathrm{sp}(V^*)+2)\min\{d(\epsilon),T\} \\
&\le \left[d(\epsilon) \sum_{t=1}^{T}\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t,f_t,\zeta_i)]\|_2^2 \right]^{1/2} + (\mathrm{sp}(V^*)+3)\min\{d(\epsilon),T\} + T\epsilon,
\end{aligned} \tag{F.18}
$$

where the last inequality follows AM-GM inequality. Thus, $d_{\mathrm{G}} \le \mathcal{O}(d\log(\mathrm{sp}(V^*)T/\sqrt{d}\epsilon))$.

Step 2: Bound over transferability coefficient.

Given condition that $\sum_{i=1}^{t-1} \|\mathbb{E}[l_{f_i}(f_t,f_t,\zeta_i)]\|_2^2 \le \beta$ for all $t \in [T]$, we have

$$
\begin{aligned}
\sum_{t=1}^{T} \|\mathbb{E}[l_{f_t}(f_t,f_t,\zeta_t)]\|_2^2 &= \sum_{t=1}^{T} \left[(\theta_i-\theta^*)^\top \left(\psi(s_t,a_t) + \int_{\mathcal{S}} \phi(s_t,a_t,s')V_{f_i}(s')\mathrm{d}s' \right) \right]^2 \\
&\le \sum_{t=1}^{T} (W_t^\top X_t)^2 \cdot \mathbb{1}\left(\|X_t\|_{\Sigma_t^{-1}}^2 \le 1 \right) + (\mathrm{sp}(V^*)+2)^2 \min\{d(\epsilon),T\} \\
&\le \sum_{i=1}^{T} (\beta+4\epsilon)\cdot \left(\|X_t\|_{\Sigma_t^{-1}}^2 \wedge 1 \right) + (\mathrm{sp}(V^*)+2)^2 \min\{d(\epsilon),T\} \\
&\le d(\epsilon)\beta\log T + (\mathrm{sp}(V^*)^2 + 4\mathrm{sp}(V^*)+6)\min\{d(\epsilon),T\} + 2T\epsilon^2,
\end{aligned} \tag{F.19}
$$

where we use a variant of (F.14) and (F.15), following that

$$
\sum_{t=1}^{T} \mathbb{1}\left(\|X_t\|_{\Sigma_t^{-1}}^2 \ge 1 \right) \le \sum_{t=1}^{T} \|X_t\|_{\Sigma_t^{-1}}^2 \wedge 1 \le \sum_{t=1}^{T} \|X_t\|_{\Sigma_t^{-1}} \wedge 1 \le d(\epsilon),
$$

and the last inequality results from a similar proof as (F.17) and (F.18) using Cauchy-Swartz and AM-GM inequality. Thus, we have $\kappa_{\mathrm{G}} \le \mathcal{O}(d\log(\mathrm{sp}(V^*)T/\sqrt{d}\epsilon))$. □

F.5 Discussion about Performance on Concrete Examples

In this subsection, we show performance of Loop for specific problems. Loop achieves an $\tilde{\mathcal{O}}(\sqrt{T})$ regret, which is nearly minimax optimal in T, in linear AMDP and linear mixture AMDP.

Linear AMDP Recall that linear function class is defined as $\mathcal{H} = \big\{(Q,J) : Q(\cdot,\cdot) = \omega^\top\phi(\cdot,\cdot) \big| \|\omega\|_2 \le \frac{1}{2}\mathrm{sp}(V^*)\sqrt{d}, J \in \mathcal{J}_{\mathcal{H}}\big\}$. Consider the ρ-covering number, note that

$$
|Q(s,a) - Q'(s,a)| \le |(\omega-\omega')^\top\phi(s,a)| \le \sqrt{2}\cdot\|\omega-\omega'\|_1.
$$

Based on the Lemma 20, combine the fact that $|\mathcal{J}_{\mathcal{H}}| \le 2$ and its ρ-covering number $\mathcal{N}_\rho(\mathcal{J}_{\mathcal{H}})$ is at most $2\rho^{-1}$, we can get the log covering number of the hypotheses class $\mathcal{H}$ is upper bounded by

$$
\log \mathcal{N}_{\mathcal{H}}(\rho) \le d\log\left(\mathrm{sp}(V^*)2^{\frac{3}{2}}d^{\frac{3}{2}}\rho^{-2}\right), \tag{F.20}
$$

by taking $\alpha = w,\ P = d,\ B = \mathrm{sp}(V^*)\sqrt{d}/2$. Recall that Proposition 5 indicates that

$$
d_{\mathrm{G}} \le \mathcal{O}\big(d\log\left(\mathrm{sp}(V^*)\sqrt{d}\rho^{-1}\right)\log T\big), \quad \kappa_{\mathrm{G}} \le \mathcal{O}\big(d\log\left(\mathrm{sp}(V^*)\sqrt{d}\rho^{-1}\right)\big). \tag{F.21}
$$

Combine (F.20), (F.21) and the regret guarantee in Theorem 3, we get

$$
\mathrm{Reg}(T) \le \mathcal{O}\Big(\mathrm{sp}(V^*)\max\{d_{\mathrm{G}},\kappa_{\mathrm{G}}\}\sqrt{T\log\left(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(1/T)/\delta\right)\mathrm{sp}(V^*)}\Big) \le \tilde{\mathcal{O}}\Big(\mathrm{sp}(V^*)^{\frac{3}{2}}d^{\frac{3}{2}}\sqrt{T}\Big).
$$

For linear AMDPs, our method achieves $\tilde{\mathcal{O}}(\mathrm{sp}(V^*)^{\frac{3}{2}}d^{\frac{3}{2}}\sqrt{T})$ regret for both linear and generalized linear AMDPs. In comparison, the Fopo algorithm (Wei et al., 2021) achieves the best-known $\tilde{\mathcal{O}}(\mathrm{sp}(V^*)d^{\frac{3}{2}}\sqrt{T})$ regret. Our method incurs an additional constant $\mathrm{sp}(V^*)^{\frac{1}{2}}$ in the regret bound.

Linear mixture Recall that the Proposition 8 posits that AGEC of the linear mixture probelm satisfies that $\max\{\sqrt{d_{\mathrm{G}}}, \kappa_{\mathrm{G}}\} \leq \mathcal{O}(d\sqrt{\log T})$. Note that the hypotheses class is defined as

$$\mathcal{H} = \{(\mathbb{P}, r) : \mathbb{P}(\cdot|s,a) = \theta^\top \phi(s,a,\cdot),\ r(s,a) = \theta^\top \psi(s,a) |\ \|\theta\|_2 \leq 1\}.$$

Consider the covering number note that both transition function and reward can be written as

$$\text{(i)}.\ |(\mathbb{P} - \mathbb{P}')(s'|s,a)| = |(\theta - \theta')^\top \phi(s,a,s')| \leq \sqrt{d} \cdot \|\theta - \theta'\|_1,$$
$$\text{(ii)}.\ |(r - r')(s,a)| = |(\theta - \theta')^\top \psi(s,a)| \leq \sqrt{d} \cdot \|\theta - \theta'\|_1.$$

Based on the Lemma 20, the log covering number of $\mathcal{H}_{\mathrm{LM}}$ is upper bounded by

$$\log \mathcal{N}_{\mathcal{H}}(\rho) \leq 2d \log\left(d^{\frac{3}{2}}\rho^{-1}\right),$$

by taking $\alpha = \theta,\ P = d,\ B = 1$. Combine results above and Theorem 3, we get

$$\mathrm{Reg}(T) \leq \mathcal{O}\Big(\mathrm{sp}(V^*) \max\{d_{\mathrm{G}}, \kappa_{\mathrm{G}}\} \sqrt{T \log\left(T\mathcal{N}^2_{\mathcal{H}\cup\mathcal{G}}(1/T)/\delta\right)\mathrm{sp}(V^*)}\Big) \leq \tilde{\mathcal{O}}\Big(\mathrm{sp}(V^*)^{\frac{3}{2}} d^{\frac{3}{2}} \sqrt{T}\Big).$$

At our best knowledge, the UCRL2-VTR (Wu et al., 2022) achieves the best $\tilde{\mathcal{O}}(Dd\sqrt{T})$ regret for linear mixture AMDP, where D is the diameter under communicating AMDP assumption and it is provable that $\mathrm{sp}(V^*) \leq D$ (Wang et al., 2022). We remark that the two algorithms are incomparable under different assumptions and both achieve a near minimax optimal regret at $\tilde{\mathcal{O}}(\sqrt{T})$.

G Technical Lemmas

In this section, we provide useful technical lemmas used in later theoretical analysis. Most are directly borrowed from existing works and proof of modified lemmas is provided in Section G.1.

Lemma 12. Given function class Φ defined on $\mathcal{X}$, and a family of probability measures Γ over $\mathcal{X}$. Suppose sequence $\{\phi_k\}_{k=1}^K \subset \Phi$ and $\{\mu_k\}_{k=1}^K \subset \Gamma$ satisfy that for all $k \in [K]$, $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[\phi_k])^2 \leq \beta$. Then, for all $k \in [K]$, we have

$$\sum_{t=1}^k \mathbb{1}\Big(\big|\mathbb{E}_{\mu_t}[\phi_t]\big| > \epsilon\Big) \leq \Big(\frac{\beta}{\epsilon^2} + 1\Big)\dim_{\mathrm{DE}}(\Phi, \Pi, \epsilon).$$

Proof. See Lemma 43 of Jin et al. (2021) for detailed proof.

Lemma 13 (Pigeon-hole principle)**.** Given function class Φ defined on $\mathcal{X}$ with $|\phi(x)| \leq C$ for all $\phi \in \Phi$ and $x \in \mathcal{X}$, and a family of probability measure over $\mathcal{X}$. Suppose sequence $\{\phi_k\}_{k=1}^K \subset \Phi$ and $\{\mu_k\}_{k=1}^K \subset \Gamma$ satisfy that for all $k \in [K]$, it holds $\sum_{t=1}^{k-1}(\mathbb{E}_{\mu_t}[\phi_k])^2 \leq \beta$. Let $d_{\mathrm{DE}} = \dim_{\mathrm{DE}}(\Phi, \Gamma, \epsilon)$ be the DE dimension, then for all $k \in [K]$ and $\epsilon > 0$, we have

$$\sum_{t=1}^k \Big|\mathbb{E}_{\mu_t}[\phi_t]\Big| \leq 2\sqrt{d_{\mathrm{DE}}\beta k} + \min\{k, d\}C + k\epsilon,$$

and

$$\sum_{t=1}^k \Big[\mathbb{E}_{\mu_t}[\phi_t]\Big]^2 \leq d_{\mathrm{DE}}\beta \log k + \min\{k, d\}C^2 + k\epsilon^2.$$

Proof. See Section G.1.1.

Lemma 14. Given function class Φ defined on $\mathcal{X}$ with $|\phi(x)| \leq C$ for all $\phi \in \Phi$ and $x \in \mathcal{X}$, and a family of probability measure over $\mathcal{X}$. Let $d_{\mathrm{DE}} = \dim_{\mathrm{DE}}(\Phi, \Gamma, \epsilon)$ be the DE dimension, then for all $k \in [K]$ and $\epsilon > 0$, we have

$$\sum_{t=1}^k \Big|\mathbb{E}_{\mu_k}[\phi_k]\Big| \leq \left[d_{\mathrm{DE}}\left(1 + \log K\right) \sum_{k=1}^K \sum_{t=1}^{k-1} (\mathbb{E}_{\mu_t}[\phi_k])^2\right]^{1/2} + \min\{d_{\mathrm{DE}}, k\}C + k\epsilon.$$

Proof. See Section G.1.2.

Lemma 15 (d-upper bound)**.** Let Φ and Ψ be sets of d-dimensional vectors and $\|\phi\|_2 \leq B_\phi$, $\|\psi\|_2 \leq B_\psi$ for any $\phi \in \Phi$ and $\psi \in \Psi$. If there exists set $(\phi_1, \dots, \phi_m)$ and $(\psi_1, \dots, \psi_m)$such that for all $t \in [m]$, $\sqrt{\sum_{k=1}^{t-1} \langle \phi_t, \psi_k \rangle^2} \leq c_1 \varepsilon$ and $|\langle \phi_t, \psi_t \rangle| > c_2 \varepsilon$, where $c_1 \geq c_2 > 0$ is a constant and $\varepsilon > 0$, then the number of elements in set is bounded by $m \leq \mathcal{O}\big(d \log(B_\phi B_\psi / \varepsilon)\big)$.

Proof. See Section G.1.3.

Lemma 16. For any sequence of positive reals $x_1, \dots, x_m$, it holds that $\frac{\sum_{i=1}^m x_i}{\sqrt{\sum_{i=1}^m i x_i^2}} \leq \sqrt{1 + \log n}$.

Proof. See Lemma 6 in Dann et al. (2021) for detailed proof.

Lemma 17. Let $\{x_i\}_{i \in [t]}$ be a sequence of vectors defined over Hilbert space $\mathcal{X}$. Let Λ_0 be a positive definite matrix and $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t-1} x_t x_t^\top$. It holds that

$$\sum_{i=1}^{t} \|x_t\|_{\Lambda_t^{-1}}^2 \wedge 1 \leq 2 \log \left(\frac{\det \Lambda_{t+1}}{\det \Lambda_0} \right).$$

Proof. See Elliptical Potential Lemma (EPL) in Dani et al. (2008) for a detailed proof.

Lemma 18 (Freedman's inequality)**.** Let $X_1, \dots, X_T$ be a real-valued martingale difference sequence adapted to filtration $\{\mathscr{F}_t\}_{t=1}^T$. Assume for all $t \in [T]$ $X_t \leq R$, then for any $\eta \in (0, 1/R)$, with probability greater than $1 - \delta$

$$\sum_{t=1}^{T} X_t \leq \mathcal{O}\Big(\eta \sum_{t=1}^{T} \mathbb{E}\big[X_t^2 | \mathscr{F}_t\big] + \frac{\log(1/\delta)}{\eta} \Big),$$

Proof. See Lemma 7 in Agarwal et al. (2014) for detailed proof.

Lemma 19 (Scaling lemma)**.** Let $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ be a d-dimensional feature mapping, there exists an invertible linear transformation $A \in \mathbb{R}^{d \times d}$ such that for any bounded function $f : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ and $\mathbf{z} \in \mathbb{R}^d$ defined by

$$f(s, a) = \phi(s, a)^\top \mathbf{z},$$

we have $\|A\phi(s, a)\| \leq 1$ and $\|A^{-1} z\| \leq \sup_{s,a} |f| \sqrt{d}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Proof. See Lemma 8 in Wei et al. (2021) for detailed proof.

In Theorem 3, the proved regret contains the logarithmic term of the $1/T$-covering number of the function classes $\mathcal{N}_{\mathcal{H}}(1/T)$, which can be regarded as a surrogate cardinality of the function class $\mathcal{H}$. Here, we provide a formal definition of ρ-covering and the upper bound of ρ-covering number.

Definition 17 (ρ-covering)**.** The ρ-covering number of a function class $\mathcal{F}$ is the minimum integer t satisfying that there exists subset $\mathcal{F}' \subseteq \mathcal{F}$ with $|\mathcal{F}'| = t$ such that for any $f \in \mathcal{F}$ we can find a correspondence $f' \in \mathcal{F}'$ that it holds $\|f - f'\|_\infty \leq \rho$.

Lemma 20 (ρ-covering number)**.** Let $\mathcal{F}$ be a function defined over $\mathcal{X}$ that can be parametrized by $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_P) \in \mathbb{R}^P$ with $|\alpha_i| \leq B$ for all $i \in [P]$. Suppose that for any $f, f' \in \mathcal{F}$ it holds that $\sup_{x \in \mathcal{X}} |f(x) - f'(x)| \leq L \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1$ and let $\mathcal{N}_{\mathcal{F}}(\rho)$ be the ρ-covering number of $\mathcal{F}$, then

$$\log \mathcal{N}_{\mathcal{F}}(\rho) \leq P \log \Big(\frac{2BLP}{\rho} \Big).$$

Proof. See Lemma 12 in Wei et al. (2021) for detailed proof.

G.1 Proof of Technical Lemmas

In this subsection, we present the proofs of technical auxiliary lemmas with modifications.

G.1.1 PROOF OF LEMMA 13

Proof of Lemma 13. The first statement is directly from Lemma 41 in Jin et al. (2021), and the second statement follows a similar procedure as below. Note that Lemma 12 suggests that

$$\sum_{t=1}^{k} \mathbb{1}\Big(\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 > \epsilon^2\Big) \leq \Big(\frac{\beta}{\epsilon^2}+1\Big)\dim_{\mathrm{DE}}(\Phi, \Gamma, \epsilon),$$

and note that the sum of squared expectation can be decomposed as

$$\begin{aligned}\sum_{t=1}^{k}\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 &= \sum_{t=1}^{k}\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2\mathbb{1}\Big(\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 > \epsilon^2\Big) + \sum_{t=1}^{k}\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2\mathbb{1}\Big(\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 \leq \epsilon^2\Big)\\ &\leq \sum_{t=1}^{k}\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2\mathbb{1}\Big(\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 > \epsilon^2\Big) + k\epsilon^2. \qquad \text{(G.1)}\end{aligned}$$

Assume sequence $\big[\mathbb{E}_{\mu_1}[\phi_1]\big]^2, \ldots, \big[\mathbb{E}_{\mu_k}[\phi_k]\big]^2$ are sorted in the decreasing order and consider $t \in [k]$ such that $\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 > \epsilon^2$, there exists a constant $\alpha \in (\epsilon^2, \big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2)$ satisfying

$$t \leq \sum_{i=1}^{k} \mathbb{1}\Big(\big[\mathbb{E}_{\mu_i}[\phi_i]\big]^2 > \alpha\Big) \leq \Big(\frac{\beta}{\alpha}+1\Big)\dim_{\mathrm{DE}}(\Phi,\Gamma,\sqrt{\alpha}) \leq \Big(\frac{\beta}{\alpha}+1\Big)\dim_{\mathrm{DE}}(\Phi,\Gamma,\epsilon),$$

where the last inequality is based on the fact that the DE dimension is monotonically decreasing in terms of ϵ as proposed in Jin et al. (2021). Denote $d_{\mathrm{DE}} = \dim_{\mathrm{DE}}(\Phi,\Gamma,\epsilon)$ and the inequality above implies that $\alpha \leq d_{\mathrm{DE}}\beta/t - d$. Thus, we have $\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 \leq d_{\mathrm{DE}}\beta/t - d$. Beside, based on the definition we also have $\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 \leq C^2$ and thus $\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 \leq \min\{d_{\mathrm{DE}}\beta/t - d, C^2\}$, then

$$\begin{aligned}\sum_{t=1}^{k}\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2\mathbb{1}\Big(\big[\mathbb{E}_{\mu_t}[\phi_t]\big]^2 > \epsilon^2\Big) &\leq \min\{d_{\mathrm{DE}},k\}C^2 + \sum_{t=d+1}^{k}\Big(\frac{d_{\mathrm{DE}}\beta}{t-d_{\mathrm{DE}}}\Big)\\ &\leq \min\{d_{\mathrm{DE}},k\}C^2 + d_{\mathrm{DE}}\cdot\beta\int_0^k \frac{1}{t}\mathrm{d}t\\ &\leq \min\{d_{\mathrm{DE}},k\}C^2 + d_{\mathrm{DE}}\cdot\beta\log k. \qquad \text{(G.2)}\end{aligned}$$

Combine (G.1) and (G.2), then finishes the proof. □

G.1.2 PROOF OF LEMMA 14

We remark that the proof provided in this subsection follows the almost same procedure as Lemma 3.16 in Zhong et al. (2022) with adjustment, and we preserve it for comprehension.

Proof of Lemma 14. Denote $d_{\mathrm{DE}} = \dim_{\mathrm{DE}}(\Phi,\Gamma,\epsilon)$, $\widehat{\epsilon}_{t,k} = |\mathbb{E}_{\mu_t}[\phi_k]|$ and $\epsilon_{t,k} = \widehat{\epsilon}_{t,k}\mathbb{1}(\widehat{\epsilon}_{t,k} > \epsilon)$ for $t,k \in [K]$, $\mu_t \in \Gamma$ and $\phi_k \in \Phi$. The proof follows the procedure below. Consider K empty buckets $B_0, \ldots, B_{K-1}$ as initialization, and we examine $\epsilon_{k,k}$ one by one for all $k \in [K]$ as below:

Case 1 If $\epsilon_{k,k} = 0$, i.e., $\widehat{\epsilon}_{k,k} \leq \epsilon$, then discard it.

Case 2 If $\epsilon_{k,k} > 0$, i.e., $\widehat{\epsilon}_{k,k} > \epsilon$, at bucket j we add k into B_j if $\sum_{t\leq k-1, t\in B_j}(\epsilon_{t,k})^2 \leq (\epsilon_{k,k})^2$, otherwise we continue with the next bucket B_{j+1}.

Denote by b_k the index of bucket that at step k the non-zero $\epsilon_{k,k}$ falls in, i.e. $k \in B_{b_k}$. Based on the rule above, it holds that

$$\sum_{k=1}^{K}\sum_{t=1}^{k-1}(\epsilon_{t,k})^2 \geq \sum_{k=1}^{K}\sum_{0\leq j\leq b_k-1, b_k\geq 1}\ \sum_{t\leq k-1, t\in B_j}(\epsilon_{t,k})^2 \geq \sum_{k=1}^{K} b_k \cdot (\epsilon_{k,k})^2,$$

where the first inequality arises from $\{t \in B_j : t \leq k-1,\ 0 \leq j \leq b_k - 1, b_k \geq 1\} \subseteq [k-1]$ due to the discarding of the b_kth bucket, and the second equality directly follows the allocation rule such that $\sum_{t\leq k-1, t\in B_j}(\epsilon_{t,k})^2 \geq (\epsilon_{k,k})^2$ for any $j \leq b_k - 1$. Recall that based on the definition of distributional eluder (DE) dimension, it is suggested the size $|B_j|$ is no larger than d_{DE}. Then,

$$\sum_{k=1}^{K} b_k(\epsilon_{k,k})^2 = \sum_{j=1}^{K-1} j \sum_{t\in B_j}(\epsilon_{t,t})^2 \qquad \text{(re-summation)}$$

$$\geq \sum_{j=1}^{K-1} \frac{j}{|B_j|}\left(\sum_{t\in B_j}\epsilon_{t,t}\right)^2 \geq \sum_{j=1}^{K-1} \frac{j}{d_{\mathrm{DE}}}\left(\sum_{t\in B_j}\epsilon_{t,t}\right)^2 \qquad (|B_j| \leq d_{\mathrm{DE}})$$

$$\geq \left(d_{\mathrm{DE}}\left(1+\log K\right)\right)^{-1}\left(\sum_{j=1}^{K-1}\sum_{t\in B_j}\epsilon_{t,t}\right)^2 = \left(d_{\mathrm{DE}}\left(1+\log K\right)\right)^{-1}\left(\sum_{t\in[K]\setminus B_0}\epsilon_{t,t}\right)^2, \tag{G.3}$$

where the second inequality follows Lemma 16. Combine the (G.1.2) and (G.3) above, we have

$$\begin{aligned}\sum_{k=1}^{K}\widehat{\epsilon}_{k,k} &\leq \sum_{k=1}^{K}\epsilon_{k,k} + K\epsilon \leq \sum_{t\in[K]\setminus B_0}\epsilon_{t,t} + \min\{d_{\mathrm{DE}}, K\}\|\phi\|_\infty + K\epsilon \\ &\leq \left[d_{\mathrm{DE}}\left(1+\log K\right)\sum_{k=1}^{K}\sum_{t=1}^{k-1}(\epsilon_{t,k})^2\right]^{1/2} + \min\{d_{\mathrm{DE}}, K\}C + K\epsilon \\ &\leq \left[d_{\mathrm{DE}}\left(1+\log K\right)\sum_{k=1}^{K}\sum_{t=1}^{k-1}(\widehat{\epsilon}_{t,k})^2\right]^{1/2} + \min\{d_{\mathrm{DE}}, K\}C + K\epsilon.\end{aligned}$$

Substitute the definition $\widehat{\epsilon}_{t,k} = |\mathbb{E}_{\mu_t}[\phi_k]|$ back into the inequality, then finishes the proof. □

G.1.3 Proof of Lemma 15

Proof of Lemma 15. For notation simplicity, denote $\Lambda_t = \sum_{k=1}^{t-1}\psi_t\psi_t^\top + \frac{\varepsilon^2}{B_\phi^2}\cdot\mathbf{I}$, then for all $t \in [m]$ we have $\|\phi_t\|_{\Lambda_t} \leq \sqrt{\sum_{k=1}^{t-1}(\phi_t^\top\psi_k)^2 + \frac{\varepsilon^2}{B_\phi^2}\|\phi_t\|_2^2} = \sqrt{c_1^2+1}\,\varepsilon$ based on the given condition. Using the Cauchy-Swartz inequality and results above, then it holds $\|\psi_t\|_{\Lambda_t^{-1}} \geq |\langle\phi_t, \psi_t\rangle|/\|\phi_t\|_{\Lambda_t} = c_2/\sqrt{c_1^2+1}$. On one hand, the matrix determinant lemma ensures that

$$\det\Lambda_m = \det\Lambda_0\cdot\prod_{t=1}^{m-1}\left(1+\|\psi_t\|^2_{\Lambda_t^{-1}}\right) \geq \left(1+\frac{c_2^2}{1+c_1^2}\right)^{m-1}\left(\frac{\varepsilon^2}{B_\phi^2}\right)^d. \tag{G.4}$$

On the other hand, according to the definition of Λ_t, we have

$$\det\Lambda_m \leq \left(\frac{\mathrm{Tr}(\Lambda_m)}{d}\right)^d \leq \left(\sum_{k=1}^{t-1}\frac{\|\psi_k\|_2^2}{d} + \frac{\varepsilon^2}{B_\phi^2}\right)^d \leq \left(\frac{B_\psi^2(m-1)}{d} + \frac{\varepsilon^2}{B_\phi^2}\right)^d. \tag{G.5}$$

Combine (G.4) and (G.5), if we take logarithms at both sides, then we have

$$m \leq 1 + d\log\left(\frac{B_\phi^2 B_\psi^2(m-1)}{d\varepsilon^2}+1\right)\Big/\log\left(1+\frac{c_2^2}{1+c_1^2}\right).$$

After simple calculations, we can obtain that m is upper bounded by $\mathcal{O}\big(d\log(B_\phi B_\psi/\varepsilon)\big)$. □

H Supplementary Discussions

H.1 Proof Sketch of MLE-based Results

In this subsection, we provide the proof sketch of Theorem 4. We first introduce several useful lemmas, which is the variant of ones in Appendix D for MLE-based problems, and most have been fully researched in Liu et al. (2022; 2023a); Xiong et al. (2023). As there's no significant technical gap between episodic and average-reward for model-based problems, we only provide a proof sketch.

Lemma 21 (Akin to Lemma 9)**.** Under Assumptions 1-2, MLE-LOOP is an optimistic algorithm such that it ensures $J_t \geq J^*$ for all $t \in [T]$ with probability greater than $1-\delta$.

Proof Sketch of Lemma 21. See Proposition 13 in Liu et al. (2022) with slight modifications. $\square$

Lemma 22 (Akin to Lemma 10)**.** For fixed $\rho > 0$ and a pre-determined optimistic parameter $\beta = c(\log\big(T\mathcal{B}_{\mathcal{H}}(\rho)/\delta\big) + T\rho\big)$ where constant $c > 0$, it holds that

$$\sum_{i=1}^{t-1} \|\mathbb{E}_{\zeta_i}[l_{f_i}(f_t, f_t, \zeta_i)]\|_1 = \sum_{i=1}^{t-1} \mathrm{TV}\big(\mathbb{P}_{f_t}(\cdot|s_i,a_i), \mathbb{P}_{f^*}(\cdot|s_i,a_i)\big) \leq \mathcal{O}(\sqrt{\beta t}), \tag{H.1}$$

for all $t \in [T]$ with probability greater than $1-\delta$.

Proof Sketch of Lemma 22. See Proposition 14 in Liu et al. (2022) with slight modifications. $\square$

Lemma 23 (Akin to Lemma 11)**.** Let $\mathcal{N}(T)$ be the switching cost with time horizon T, given fixed covering coefficient $\rho > 0$ and pre-determined optimistic parameter $\beta = c\big(\log\big(T\mathcal{B}_{\mathcal{H}}(\rho)/\delta\big) + T\rho\big)$ where c is a large enough constant, with probability greater than $1-2\delta$ we have

$$\mathcal{N}(T) \leq \mathcal{O}\big(\kappa_{\mathrm{G}} \cdot \mathrm{poly}(\log T) + \beta^{-1}T\epsilon^2\big),$$

where κ_{G} is the transferability coefficient with respect to MLE-AGEC$(\mathcal{H}, \{l_{f'}\}, \epsilon)$.

Proof Sketch of Lemma 23. The proof is almost the same as Lemma 11.

Step 1: Bound the difference of discrepancy between the minimizer and f^*.

As proposed in Proposition 14, Liu et al. (2022), $0 \leq \sum_{i=1}^{t} \mathrm{TV}\big(\mathbb{P}_{f^*}(\cdot|s_i,a_i), \mathbb{P}_{g_i}(\cdot|s_i,a_i)\big)^2 \leq \beta$ holds with high probability if the update happens at t-th step. Based on the AM-GM inequlaity, we have

$$0 \leq \sum_{i=1}^{t} \mathrm{TV}\big(\mathbb{P}_{f^*}(\cdot|s_i,a_i), \mathbb{P}_{g_i}(\cdot|s_i,a_i)\big) \leq \sqrt{\beta t}. \tag{H.2}$$

Step 2: Bound the expected discrepancy between updates.

Note that for all $t+1 \in [T]$, the update happens only if

$$\sum_{i=1}^{t} \mathrm{TV}\big(\mathbb{P}_{f_t}(\cdot|s_i,a_i), \mathbb{P}_{g_i}(\cdot|s_i,a_i)\big) > 3\sqrt{\beta t}. \tag{H.3}$$

Combine the (H.2) and (H.3) above, and apply the triangle inequality, we have

$$\begin{aligned}
&\sum_{i=1}^{t} \mathrm{TV}\big(\mathbb{P}_{f_t}(\cdot|s_i,a_i), \mathbb{P}_{f^*}(\cdot|s_i,a_i)\big) \\
&\geq \sum_{i=1}^{t} \mathrm{TV}\big(\mathbb{P}_{f_t}(\cdot|s_i,a_i), \mathbb{P}_{g_t}(\cdot|s_i,a_i)\big) - \mathrm{TV}\big(\mathbb{P}_{f^*}(\cdot|s_i,a_i), \mathbb{P}_{g_t}(\cdot|s_i,a_i)\big) \geq 2\sqrt{\beta t}.
\end{aligned}$$

and the construction of confidence set ensures that $\sum_{i=1}^{\tau_t} \mathrm{TV}\big(\mathbb{P}_{f_t}(\cdot|s_i,a_i), \mathbb{P}_{f^*}(\cdot|s_i,a_i)\big) \leq \sqrt{\beta\tau_t}$ with high probability (Liu et al., 2022, Proposition 14). Recall the definition of MLE-transferability coefficient, then the switching cost can be bounded following the same argument in Lemma 11. $\square$

Proof Sketch of Theorem 4. Recall that

$$\mathrm{Reg}(T) \leq \underbrace{\sum_{i=1}^{T} \mathcal{E}(f_t)(s_t,a_t)}_{\text{Bellman error}} + \underbrace{\sum_{t=1}^{T} \Big(\mathbb{E}_{s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}[V_t(s_{t+1})] - V_t(s_t)\Big)}_{\text{Realization error}}, \tag{H.4}$$

where the inequality follows the optimism in Lemma 21. Combine Lemma 22, Lemma 23 and the definition of MLE-AGEC (see Definition 9), then we can finish the proof. $\square$

Algorithm 3 Extended Value Iteration (EVI)

Input: hypothesis $f = (\mathbb{P}_f, r_f)$, desired accuracy level ϵ.
Initialize: $V^{(0)}(s) = 0$ for all $s \in \mathcal{S}$, $J^{(0)} = 0$ and counter $i = 0$.
repeat
  for $s \in \mathcal{S}$ and $a \in \mathcal{A}$ **do**
    Set $Q^{(i)}(s,a) \leftarrow r_f(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_f(s,a)}[V^{(i)}(s')] - J^{(i)}$
    Update $V^{(i+1)}(s) \leftarrow \max_{a \in \mathcal{A}} Q^{(i)}(s,a)$
  Update counter $i \leftarrow i + 1$
until $\max_{s \in \mathcal{S}}\{V^{(i+1)}(s) - V^{(i)}(s)\} - \min_{s \in \mathcal{S}}\{V^{(i+1)}(s) - V^{(i)}(s)\} \leq \epsilon$

H.2 Extended Value Iteration (EVI) for Model-Based Hypotheses

In model-based problems, the discrepancy function sometimes relies on the optimal state bias function V_f and optimal average-reward J_f (see linear mixture model in Section C). In this section, we provide an algorithm, extended value iteration (EVI) proposed in Auer et al. (2008), to output the optimal function and average-reward under given a model-based hypothesis $f = (\mathbb{P}_f, r_f)$. See Algorithm 3 for complete pseudocode. The convergence of EVI is guaranteed by the theorem below.

Theorem 24. UnderAssumption 1, there exists a unique centralized solution pair (Q^*, J^*) to the Bellman optimality equation for any AMDP $\mathcal{M}_f$ characterized by hypothesis $f \in \mathcal{H}$. Then, if the extended value iteration (EVI) is stopped under the condition that

$$\max_{s \in \mathcal{S}}\{V^{(i+1)}(s) - V^{(i)}(s)\} - \min_{s \in \mathcal{S}}\{V^{(i+1)}(s) - V^{(i)}(s)\} \leq \epsilon,$$

then the achieved greedy policy $\pi^{(i)}$ is ϵ-optimal such that $J_{\mathcal{M}_f}^{\pi^{(i)}} \geq J_{\mathcal{M}_f}^{*} + \epsilon$.

Proof Sketch: See Theorem 12 in Auer et al. (2008).