GENERALIZED LINEAR MARKOV DECISION PROCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

The linear Markov Decision Process (MDP) provides a principled basis for reinforcement learning (RL) but assumes that both transitions and rewards are linear in the *same* feature space. This severely limits its applicability when rewards are nonlinear or discrete. We introduce the Generalized Linear MDP (GLMDP), which retains linear transitions while modeling rewards with generalized linear models **under potentially different feature maps**. This separation is crucial: transitions may admit rich representations learned from large unlabeled trajectories, while rewards can be modeled with limited labeled data. We show that GLMDPs are Bellman complete with respect to a new function class, enabling efficient value iteration. Based on this, we develop algorithms with provable guarantees in both **offline** and **online** settings. For offline RL, we design pessimistic and semi-supervised value iteration methods that achieve policy suboptimality bounds and demonstrate significant label-efficiency gains. For online RL, we propose an optimistic algorithm with a near-optimal regret bound. Together, these results broaden the scope of structured and sample-efficient RL to applications with complex reward structures, such as healthcare and e-commerce.

1 INTRODUCTION

Reinforcement learning (RL) has achieved impressive success in domains such as gaming and robotics (Silver et al., 2016; Berner et al., 2019), where abundant online interaction is feasible. However, real-world applications from precision medicine to e-commerce often involve costly, ethically constrained, or risky data collection. In such settings, algorithms must be both *sample-efficient* and capable of learning from limited offline datasets (Levine et al., 2020). Classical deep RL methods, which rely on highly expressive neural networks, can overfit and fail under data scarcity, motivating the study of structured RL frameworks.

Among these, the *linear Markov Decision Processes (MDP)* framework (Jin et al., 2020) provides strong theoretical guarantees and tractable algorithms, and has been applied to healthcare and recommendation systems (Cai et al., 2018; Gao et al., 2024; Trella et al., 2025). Yet linear MDPs assume that both transition dynamics and reward functions are linear in the *same* feature space. This assumption breaks down in practice: outcomes are often binary or count-valued (e.g., treatment adherence, purchase events), while transitions may admit far more complex structure. Consequently, linear MDPs cannot fully capture real-world settings where *reward and transition require distinct feature representations*.

The GLMDP framework. We propose the *Generalized Linear MDP (GLMDP)*, which relaxes this assumption by allowing distinct feature maps for rewards and transitions. Formally, at each step $h \in \{1, \dots, H\}$ in an episodic MDP, the reward and transition satisfy

$$\mathbb{E}[r_h(x_h, a_h) \mid x_h = x, a_h = a] = g(\langle \phi_r(x, a), \theta_h^\star \rangle), \tag{1}$$

$$\mathbb{P}_h(x_{h+1} \mid x_h, a_h) = \langle \phi_p(x_h, a_h), \mu_h(x_{h+1}) \rangle, \tag{2}$$

where $g(\cdot)$ is a known link function, $\phi_r \in \mathbb{R}^{d_r}$ and $\phi_p \in \mathbb{R}^{d_p}$ are (possibly different) feature maps, $\theta_h^\star \in \mathbb{R}^{d_r}$ is an unknown parameter, and μ_h is a measure over next-state distributions. Unlike linear MDPs, GLMDPs capture nonlinear or discrete reward structures while supporting complex transition dynamics, and crucially allow transitions to be learned from large amounts of unlabeled data. When $g(x) = x$ and $\phi_r = \phi_p$, the GLMDP reduces to the standard linear MDP. At first glance, this extension may appear simple—merely introducing a link function for rewards—yet ensuring that the

resulting model class is *Bellman complete* is highly non-trivial. Many natural generalizations of linear MDPs fail to admit any closed Bellman class, making value iteration intractable. Our key contribution is to identify a function family under which GLMDPs are Bellman complete, and to design algorithms that exploit the decoupling between reward and transition estimation. This structural separation is particularly powerful in the semi-supervised setting, where abundant unlabeled trajectories improve transition estimation while only a small fraction require costly reward labels.

Our algorithms and results. Building on this framework, we design algorithms for both offline and online RL. In the offline setting, we introduce *Generalized Pessimistic Value Iteration* (GPEVI) and a semi-supervised variant (SS-GPEVI) that leverages unlabeled trajectories to improve label efficiency. We provide suboptimality bounds showing that SS-GPEVI can substantially outperform fully supervised methods when the transition model is high-dimensional. In the online setting, we propose an optimistic algorithm (GLSVI-UCB) and establish a near-optimal regret bound.

Contributions. Our main contributions are:

- We introduce the **GLMDP framework**, which generalizes linear MDPs by allowing GLM rewards and distinct feature maps for rewards and transitions.
- We prove **Bellman completeness** for a new function class, ensuring tractability under GLMDPs.
- We develop **offline algorithms** (GPEVI, SS-GPEVI) with suboptimality guarantees, showing that unlabeled trajectories can provably accelerate learning when $d_p \gg d_r$.
- We design an **online algorithm** (GLSVI-UCB) with a near-optimal regret bound, extending optimistic exploration to the GLMDP setting.

These results broaden the scope of structured and provably efficient RL, making it applicable to domains with complex reward structures and limited labels.

The remainder of our paper is structured as follows: We explain our GLMDP framework in Section 2, followed by our proposed algorithms in Section 3. Section 4 provides theoretical guarantees that validate our approach's effectiveness. We offer conclusions and future research directions in Section 5. Additional materials are included in the appendices: a comprehensive literature review (Appendix A), algorithm pseudocode (Appendix B), simulation studies (Appendix C), simulation environment studies (Appendix D), discussion about unbounded reward function (Appendix E) and proofs (Appendices F-M).

2 Generalized Linear MDP Framework

We begin by formally defining the *Generalized Linear* MDP (GLMDP) framework. In our framework, we consider an episodic MDP with finite horizon length H. At each time step $h \in \{1, 2, \ldots, H\}$, the reward functions $\{r_h\}_{h=1}^H$ and transition kernels $\{\mathbb{P}_h\}_{h=1}^H$ satisfy equation 1 and equation 2.

Given any policy $\pi = \{\pi_h\}_{h=1}^H$, we denote $\mathcal{S}$ as the state space and $\mathcal{A}$ as the action space and define the state-value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ and the action-value function (Q-function) $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ at time step $h \in [H]$ as follows:

$$V_h^\pi(x) = \mathbb{E}_\pi\Big[\sum_{t=h}^{H} r_t(x_t, a_t) \mid x_h = x\Big], \tag{3}$$

$$Q_h^\pi(x, a) = \mathbb{E}_\pi\Big[\sum_{t=h}^{H} r_t(x_t, a_t) \mid x_h = x, a_h = a\Big]. \tag{4}$$

In equation 3 and equation 4, the expectation $\mathbb{E}_\pi$ is computed over all possible trajectories generated by policy π. Specifically, at each time step $t \in [H]$, we sample action $a_t \sim \pi_t(\cdot \mid x_t)$ at state x_t and observe the subsequent state $x_{t+1} \sim \mathbb{P}_t(\cdot \mid x_t, a_t)$. Here For a positive integer d, we define $[d] = \{1, \ldots, d\}$. Note that in equation 3, we condition on the initial state $x_h = x$, while in equation 4, we condition on both the initial state and action $(x_h, a_h) = (x, a) \in \mathcal{S} \times \mathcal{A}$.

We denote optimal policy, state value function, and Q function by $\pi^* = \{\pi_h^*\}_{h=1}^H$, $V^* = \{V_h^*\}_{h=1}^H$, and $Q^* = \{Q_h^*\}_{h=1}^H$, respectively. Specifically, the optimal state value function $V_h^*(x)$ represents the maximum possible expected return achievable from any state x at step h, defined as $V_h^*(x) = \sup_\pi V_h^\pi(x)$. Similarly, the optimal action value function is defined as $Q_h^*(x,a) = \sup_\pi Q_h^\pi(x,a)$. An optimal policy π^* is any policy that achieves these optimal values. This policy is greedy with respect to the optimal Q function, meaning that it selects an action that maximizes the Q value at each state: $\pi_h^*(\cdot \mid x) \in \arg\max_{a\in\mathcal{A}} Q_h^*(x,a)$.

The fundamental relationships from the Bellman equation are:

$$V_h^\pi(x) = \left\langle Q_h^\pi(x,\cdot), \pi_h(\cdot \mid x)\right\rangle_{\mathcal{A}}, \quad Q_h^\pi(x,a) = (\mathbb{B}_h V_{h+1}^\pi)(x,a),$$

where $\langle\cdot,\cdot\rangle_{\mathcal{A}}$ denotes the inner product over the action space $\mathcal{A}$. In addition, $\mathbb{B}_h$ represents the Bellman operator defined by:

$$(\mathbb{B}_h V)(x,a) = \mathbb{E}\big[r_h(x_h,a_h) + V(x_{h+1}) \mid x_h = x, a_h = a\big]$$

for any function $V : \mathcal{S} \to \mathbb{R}$. The expectation $\mathbb{E}$ is taken over the randomness in both the reward $r_h(x_h,a_h)$ and the next state x_{h+1}, where $x_{h+1} \sim \mathbb{P}_h(x_{h+1} \mid x_h, a_h)$.

The strong structure assumed in Linear MDPs ensures the linear Q-value function class is complete with respect to the Bellman operator, often referred to as Bellman completeness (Xie et al., 2021). Bellman completeness lies at the foundation of the value iteration algorithm over the linear class. We show in Proposition 1 that our extension to the linear MDP retains the Bellman completeness property over the function class $\mathcal{F}$ defined below

$$\mathcal{F} = \left\{(x,a) \mapsto g\left(\langle\phi_r(x,a),\theta\rangle\right) + \langle\phi_p(x,a),\beta\rangle : \theta \in \mathbb{R}^{d_r}, \beta \in \mathbb{R}^{d_p}\right\}. \tag{5}$$

Proposition 1 (Bellman Completeness of GLMDP)**.** *The GLMDP framework satisfies Bellman completeness with respect to the function class $\mathcal{F}$ defined in equation 5. That is, for all $f \in \mathcal{F}$, all policies $\pi = \{\pi_h\}_{h=1}^H$, and all time steps $h \in [H]$, we have $\mathbb{B}_h^\pi f \in \mathcal{F}$ where $\mathbb{B}_h^\pi f(x,a)$ is defined as $\mathbb{B}_h^\pi f(x,a) \coloneqq \mathbb{E}[r_h(x,a) + f(x_{h+1},\pi(x_{h+1})) \mid x_h = x, a_h = a]$.*

Corollary 1. *As a direct consequence of Bellman completeness, the optimal Q-value function satisfies $Q_h^* \in \mathcal{F}$ for all $h \in [H]$. Specifically:*

$$Q_h^*(x,a) = g\left(\langle\phi_r(x,a),\theta_h^*\rangle\right) + \langle\phi_p(x,a),\beta_h^*\rangle, \text{ where } \beta_h^* = \int_{\mathcal{S}} V_{h+1}^*(x')\mu_h(x')dx'. \tag{6}$$

This result connects to Chang et al. (2022) on learning Bellman complete representations for offline reinforcement learning, which is particularly crucial in the offline RL setting. Without this property, error propagation can become uncontrollable with limited offline data. Chang et al. (2022) demonstrated that learning approximately linear Bellman complete representations with good data coverage (i.e., $\lambda_{\min}(\frac{1}{n}\sum_{i=1}^n \phi(x,a)\,\phi(x,a)^\top) > 0$, where $\lambda_{\min}$ is the minimum eigenvalue of the feature covariance matrix.) is essential for sample-efficient offline policy evaluation. Similarly, for GLMDPs, the Bellman completeness property enables provable sample efficiency in offline RL settings where exploration is not possible.

3 Algorithms

In this section, we present algorithmic solutions for the GLMDP framework under both offline and online settings. The offline setting addresses scenarios with pre-collected datasets, while the online setting handles real-time interaction with the environment.

3.1 Offline Reinforcement Learning

We consider a dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{n,H}$ comprising n trajectories with time horizon H. The data is generated as follows: Within each trajectory $\tau \in [n]$ and at each time step $h \in [H]$, an agent executes action $a_h^\tau \in \mathcal{A}$ from state $x_h^\tau \in \mathcal{S}$ according to policy $\pi_h(a_h \mid x_h = x_h^\tau)$, obtains reward $r_h^\tau = r_h(x_h^\tau, a_h^\tau)$, where $r_h : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}$ is a random function, and transitions to the subsequent state x_{h+1}^τ sampled from $\mathbb{P}_h(\cdot|x_h = x_h^\tau, a_h = a_h^\tau)$. The reward functions $\{r_h\}_{h=1}^H$ and transition kernels $\{\mathbb{P}_h\}_{h=1}^H$ are specified in equation 1 and equation 2.

We define the suboptimality of a policy π with an initial state x as

$$\text{SubOpt}\left(\pi; x\right) = V_1^*(x) - V_1^\pi(x).$$

3.1.1 SUPERVISED LEARNING ALGORITHM

While the GLMDP model enjoys the desirable property of Bellman completeness, a central question remains: *Can we design an efficient algorithm that provably learns an optimal policy under this model?* Motivated by this, we propose the GPEVI algorithm, adapted from the pessimism-based approach in Jin et al. (2021), tailored to the GLMDP setting. For simplicity of presentation, we assume that the random reward function is bounded $r_h(x,a) \in [0,1]$. The case where the random reward function $r_h(x,a)$ is unbounded is discussed in Appendix E; this generalization does not affect our main result.

Guided by the Bellman equation equation 6 in Proposition 1, we approximate the optimal action-value function Q_h^* by estimating the parameters θ_h^* and β_h^*, respectively. First, we can obtain the estimator for θ_h^* as

$$\widetilde{\theta}_h = \arg\min_{\theta \in \mathbb{R}^{d_r}} \mathcal{L}_h(\theta) \tag{7}$$

where $\mathcal{L}_h(\theta) = \frac{1}{n}\sum_{\tau=1}^n \big(- r_h^\tau \langle \phi_r(x_h^\tau, a_h^\tau), \theta\rangle + G(\langle \phi_r(x_h^\tau, a_h^\tau), \theta\rangle)\big)$ and $G(a) = \int_0^a g(u)\mathrm{d}u$. The loss function $\mathcal{L}_h(\cdot)$ arises from the negative log-likelihood of a generalized linear model (GLM) with canonical link function (McCullagh and John, 1989).

To estimate the transition component, we define the empirical Bellman error for a value function $V : \mathcal{S} \to \mathbb{R}$ as $M_h(\beta \mid V) = \sum_{\tau=1}^n \big(V(x_{h+1}^\tau) - \langle \phi_p(x_h^\tau, a_h^\tau), \beta\rangle\big)^2$ for $h \in [H]$. Starting with $\widetilde{V}_{H+1}(x) = 0$, we then recursively compute $\widetilde{\beta}_h \in \mathbb{R}^{d_p}$ as

$$\widetilde{\beta}_h = \arg\min_{\beta \in \mathbb{R}^{d_p}} M_h(\beta \mid \widetilde{V}_{h+1}) + \lambda\|\beta\|_2^2 = \sum_{\tau=1}^n (\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \phi_p(x_h^\tau, a_h^\tau) \widetilde{V}_{h+1}(x_{h+1}^\tau)\,, \tag{8}$$

where $\lambda > 0$ is some regularization parameter and $\widetilde{\Lambda}_h = \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau)\phi_p(x_h^\tau, a_h^\tau)^\intercal$. Here we use $\|v\|_2 = \sqrt{\langle v, v\rangle}$ to denote the Euclidean norm of a vector v. An estimate of Q_h^* at time h is $(\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1})(x,a) \coloneqq g\big(\phi_r(x,a)^\intercal \widetilde{\theta}_h\big) + \phi_p(x,a)^\intercal \widetilde{\beta}_h$. To obtain theoretical guarantees, we quantify the deviation between $\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1}$ and the true Bellman operator $\mathbb{B}_h \widetilde{V}_{h+1}$ on the same value function $\widetilde{V}_{h+1}$ using a pessimism-based uncertainty quantification technique (Jin et al., 2021). The pessimism technique deliberately underestimates value functions to ensure conservativeness in learning, which provides robust theoretical guarantees in the presence of uncertainty.

We adopt the notion of a ξ-Uncertainty Quantifier introduced by Jin et al. (2021).

Definition 1 (ξ-Uncertainty Quantifier)**.** *We say* $\{\Gamma_h\}_{h=1}^H$ $(\Gamma_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R})$ *is a* ξ*-uncertainty quantifier of* $\{\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1}\}_{h=1}^H$ *if the event*

$$\mathcal{E} = \big\{ |(\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1})(x,a) - (\mathbb{B}_h \widetilde{V}_{h+1})(x,a)| \le \Gamma_h(x,a) \text{ for all } (x,a) \in \mathcal{S}\times\mathcal{A}, h \in [H] \big\} \tag{9}$$

satisfies $\mathbb{P}_{\mathcal{D}}(\mathcal{E}) \ge 1 - \xi$*, where the probability is taken over the randomness in the generation of the dataset* $\mathcal{D}$*.*

We then construct the uncertainty bound as:

$$\widetilde{\Gamma}_h(x,a) = \widetilde{\Gamma}_{r,h}(x,a) + \widetilde{\Gamma}_{p,h}(x,a), \quad \text{where} \tag{10}$$

$$\widetilde{\Gamma}_{r,h}(x,a) = \alpha_r \sqrt{\dot{g}(\langle \phi_r(x,a), \widetilde{\theta}_h\rangle)^2 \phi_r(x,a)^\intercal \widetilde{\Sigma}_h(\widetilde{\theta}_h)^{-1} \phi_r(x,a)}$$

$$\widetilde{\Gamma}_{p,h}(x,a) = \alpha_p \sqrt{\phi_p(x,a)^\intercal (\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \phi_p(x,a)}$$

with two hyper-parameters α_r and α_p that control the confidence level and $\dot{g}$ representing the first-order derivative of g, and $\widetilde{\Sigma}_h(\widetilde{\theta}_h) = \sum_{\tau=1}^n \dot{g}(\langle \phi_r(x_h^\tau, a_h^\tau), \widetilde{\theta}_h\rangle)\phi_r(x_h^\tau, a_h^\tau)\phi_r(x_h^\tau, a_h^\tau)^\intercal$. We will show later that $\widetilde{\Gamma}_h(x,a)$ is a ξ-Uncertainty Quantifier for $(\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1})(x,a)$ under some mild conditions (Theorem 1). We now define the pessimistically adjusted Q-function and the corresponding value function:

$$\widetilde{Q}_h(x,a) = \min\{(\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1})(x,a) - \widetilde{\Gamma}_h(x,a), H - h + 1\}^+,$$

$$\widetilde{V}_h(x) = \langle \widetilde{Q}_h(x,\cdot), \widetilde{\pi}_h(\cdot \mid x)\rangle_{\mathcal{A}}, \quad \text{where } \widetilde{\pi}_h(\cdot \mid x) = \arg\max_{\pi_h} \langle \widetilde{Q}_h(x,\cdot), \pi_h(\cdot \mid x)\rangle_{\mathcal{A}}.$$

where $\min\{x, y\}^+ = \max\{\min\{x, y\}, 0\}$. The procedure is summarized in Algorithm B.1.

A key novelty of the proposed GPEVI algorithm is the decomposition of the total uncertainty $\widetilde{\Gamma}_h(x, a)$ into two interpretable components: the first part $\widetilde{\Gamma}_{r,h}(x, a)$ captures uncertainty in reward estimation and the second part $\widetilde{\Gamma}_{p,h}(x, a)$ captures uncertainty in transition dynamics. In contrast to prior work such as PEVI (Jin et al., 2021) for linear MDPs, which uses a single aggregated uncertainty bound, our decomposed approach offers three advantages: (1) Interpretability: It provides a clearer understanding of how reward and transition contribute to overall uncertainty; (2) Flexibility in semi-supervised settings: Reward and transition models can be trained using datasets of different sizes or sources; and (3) Adaptivity to GLMs: The reward uncertainty term explicitly includes $\dot{g}$, reflecting the local curvature of the link function and scaling uncertainty appropriately. This decomposition is essential for extending pessimism-based methods beyond linear MDPs to the more expressive GLMDP framework.

3.1.2 Semi-supervised Learning Algorithm

In many practical applications, collecting fully labeled data can be costly and labor-intensive. Reward annotations often require human expertise or specialized instrumentation, making them particularly expensive to acquire. In contrast, state-action-next-state triplets $(x_h^\tau, a_h^\tau, x_{h+1}^\tau)$ are often available at much larger scales (Sonabend et al., 2020; Konyushkova et al., 2020; Hu et al., 2023). This observation motivates a semi-supervised learning approach that leverages both labeled data and more readily available unlabeled data.

The modular structure of our GLMDP framework naturally supports such an approach. Since the reward and transition models are parameterized independently, we can estimate the reward parameters θ_h^* using the labeled dataset $\mathcal{D}$, and estimate the transition parameter β_h^* using both the labeled dataset $\mathcal{D}$ and an unlabeled dataset $\mathcal{D}_u = \{(x_h^\tau, a_h^\tau)\}_{\tau=n+1,h=1}^{n+N,H}$.

Our proposed semi-supervised algorithm, SS-GPEVI, summarized in Algorithm B.2, builds upon the fully supervised GPEVI, but introduces key modifications to incorporate unlabeled data for improved sample efficiency.

Specifically, we estimate β_h^* using both labeled and unlabeled datasets:

$$\widehat{\beta}_h = (\widehat{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \sum_{\tau=1}^{n+N} \phi_p(x_h^\tau, a_h^\tau) \widehat{V}_{h+1}(x_{h+1}^\tau), \tag{11}$$

where $\widehat{\Lambda}_h = \sum_{\tau=1}^{n+N} \phi_p(x_h^\tau, a_h^\tau) \phi_p(x_h^\tau, a_h^\tau)^\top$ includes contributions from both datasets. Similarly, we construct the uncertainty quantifier using information from both datasets:

$$\widehat{\Gamma}_h(x, a) = \widetilde{\Gamma}_{r,h}(x, a) + \widehat{\Gamma}_{p,h}(x, a), \quad \text{where} \tag{12}$$
$$\widehat{\Gamma}_{p,h}(x, a) = \alpha_p \sqrt{\phi_p(x, a)^\top (\widehat{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \phi_p(x, a)}.$$

3.2 Online Reinforcement Learning

While the offline setting is valuable for scenarios with pre-collected data, many applications require real-time learning through environment interaction. In the online setting, the agent sequentially interacts with the GLMDP environment over T episodes where the length of each episode is H, aiming to maximize cumulative reward while learning the optimal policy.

3.2.1 Problem Formulation

In the online RL setting, at episode $t \in [T]$, the agent interacts with the episodic MDP as follows: starting from a fixed initial state $x_{1,t} \in \mathcal{S}$, at step $h \in [H]$ the agent follows policy $\pi_t = \{\pi_{h,t}\}_{h=1}^H$ to select action $a_{h,t}$, receives reward $r_{h,t}$, and transitions to the next state $x_{h+1,t}$. This interaction continues until the terminal step H is reached.

We measure the performance of a T-episode online algorithm with initial state x by its cumulative regret: $\mathcal{R}(x) = TV_1^*(x) - \mathbb{E}\left[\sum_{t=1}^T \sum_{h=1}^H r_{h,t}\right]$ where the expectation is taken over all randomness in the algorithm and environment.

3.2.2 Generalized Least-Squares Value Iteration with UCB

For the online setting, we propose the Generalized Least-Squares Value Iteration with Upper Confidence Bound (GLSVI-UCB) algorithm. Unlike the offline pessimistic approach, the online algorithm employs optimistic exploration through upper confidence bounds to encourage exploration of potentially rewarding state-action pairs.

The key insight is to adapt the principle of optimism in the face of uncertainty to the GLMDP framework. At each episode t, we maintain estimates of both reward parameters $\widehat{\theta}_{h,t}$ and transition parameters $\widehat{\beta}_{h,t}$, along with confidence sets that guide exploration.

For reward estimation, we solve the regularized GLM problem:

$$\widehat{\theta}_{h,t} \coloneqq \arg\min_{\|\theta\|_2 \leq M} \frac{1}{t} \sum_{\tau=1}^{t} (-r_{h,\tau} \langle \phi_r(x_{h,\tau}, a_{h,\tau}), \theta \rangle + G(\langle \phi_r(x_{h,\tau}, a_{h,\tau}), \theta \rangle)) \tag{13}$$

where $M > 0$ is a bound on $\|\theta_h^*\|_2$ and G is the primitive function of the link function g.

For transition estimation, we use least-squares regression:

$$\widehat{\beta}_{h,t} = \sum_{\tau=1}^{t} \Lambda_{h,t}^{-1} \phi_p(x_{h,\tau}, a_{h,\tau}) \max_{a \in \mathcal{A}} \bar{Q}_{h+1,t}(x_{h+1,\tau}, a) \tag{14}$$

where $\Lambda_{h,t} = \sum_{\tau=1}^{t} \phi_p(x_{h,\tau}, a_{h,\tau})^\top \phi_p(x_{h,\tau}, a_{h,\tau}) + \mathbf{I}_{d_p}$ is the empirical covariance matrix.

The algorithm maintains optimistic Q-function estimates:

$$\bar{Q}_{h,t}(x, a) = \min \left\{ H - h + 1, g(\phi_r(x, a)^\top \widehat{\theta}_{h,t}) + \phi_p(x, a)^\top \widehat{\beta}_{h,t} + \Gamma_{r,h,t}(x, a) + \Gamma_{p,h,t}(x, a) \right\} \tag{15}$$

where the confidence bounds are: $\Gamma_{r,h,t}(x, a) = \gamma_r \|\phi_r(x, a)\|_{\Lambda'^{-1}_{h,t}}$ and $\Gamma_{p,h,t}(x, a) = \gamma_p \|\phi_p(x, a)\|_{\Lambda_{h,t}^{-1}}$ with $\Lambda'_{h,t} = \sum_{\tau=1}^{t} \phi_r(x_{h,\tau}, a_{h,\tau})^\top \phi_r(x_{h,\tau}, a_{h,\tau}) + \mathbf{I}_{d_r}$ and appropriate confidence parameters γ_r, γ_p.

3.2.3 Online Algorithm

The GLSVI-UCB algorithm, detailed in Algorithm B.3, seamlessly integrates the structural properties of GLMDPs with the optimistic exploration principle. Its key innovation lies in the decomposed confidence bounds, $\Gamma_{r,h,t}$ and $\Gamma_{p,h,t}$, which separately account for uncertainty in reward and transition estimation. Unlike the pessimistic orientation of our offline algorithms, this approach adds an uncertainty bonus to the estimated Q-values, embodying the principle of "optimism in the face of uncertainty."

This optimistic construction encourages the agent to systematically explore state-action pairs for which its model is uncertain, as these hold the greatest potential for learning. The magnitude of this exploration is dynamically controlled: as more data is gathered through interaction with the environment, as captured by the covariance matrices $\Lambda_{h,t}$ and $\Lambda'_{h,t}$, the confidence bounds shrink. This mechanism ensures an efficient and adaptive transition from an initial, broad exploration to a more focused exploitation of learned high-value actions over time.

4 Theoretical Analysis

In this section, we establish theoretical performance guarantees for our proposed algorithms under both offline and online settings. Our analysis reveals the fundamental trade-offs between sample complexity, model expressiveness, and algorithmic design choices.

4.1 Offline Reinforcement Learning: Theoretical Analysis

We begin by analyzing the performance of our offline algorithms under a set of regularity assumptions that ensure the well-posedness of the GLMDP framework.

Assumption 1. *The link function $g(\cdot)$ has bounded first- and second-order derivatives, denoted $\dot{g}$ and $\ddot{g}$, respectively. In particular, there exists a constant $L > 0$ such that for all $u, v \in \mathbb{R}$, $|\dot{g}(u) - \dot{g}(v)| \leq L|u - v|$. Furthermore, the inequality $|\ddot{g}| \leq \dot{g}$ holds everywhere.*

Assumption 1 imposes smoothness and pseudo self-concordance properties on the link function, which are crucial for controlling approximation errors in GLMs (see, e.g., Ostrovskii and Bach (2021)). Common link functions such as the identity and logistic functions satisfy this assumption. We further define the following matrices:

$$\Sigma_h(\theta_h) = \mathbb{E}_\pi\left[\dot{g}(\langle\phi_r(x_h, a_h), \theta_h\rangle)\phi_r(x_h, a_h)\phi_r(x_h, a_h)^\intercal\right] \text{ and } \Lambda_h = \mathbb{E}_\pi\left[\phi_p(x_h, a_h)\phi_p(x_h, a_h)^\intercal\right].$$

Assumption 2. *We have $\lambda_{\min}\big(\Sigma_h(\theta_h^*)\big) \geq \rho > 0$ for some constant ρ.*

Assumption 2 guarantees sufficient variability in the feature representations by ensuring that the covariance matrix $\Sigma_h(\theta_h^*)$ is well-conditioned. For technical simplicity, we assume that $\max\{\|\phi_r(x, a)\|_2^2, \|\phi_p(x, a)\|_2^2\} \leq 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, $\big\|\mu_h(\mathcal{S})\big\| \leq \sqrt{d_p}$, where we define $\big\|\mu_h(\mathcal{S})\big\| \coloneqq \int_{\mathcal{S}} \big\|\mu_h(x)\big\|_2 \mathrm{d}x$. These regularity assumptions are common in the literature and can be satisfied with suitable normalization.

Theorem 1 (Suboptimality for GPEVI)**.** *Under Assumptions 1-2, we set $\lambda = 1$, $\alpha_r = c_r\sqrt{d_r \log H/\xi}$, $\alpha_p = c_p\left(d_p + d_r\right)H\sqrt{\zeta}$, where $\zeta = \log\left(2\left(d_r + d_p\right)Hn/\xi\right)$, $c_r, c_p > 0$ are absolute constants and $\xi \in (0, 1)$ is the confidence parameter. Then $\widetilde{\Gamma}_h$ in equation 10 is a ξ-uncertainty quantifier of $\widetilde{\mathbb{B}}_h$ w.r.t. value function $\widetilde{V}_{h+1}$. For any $x \in \mathcal{S}$ and n large enough, $\widetilde{\pi} = \{\widetilde{\pi}_h\}_{h=1}^H$ in Algorithm B.1 satisfies*

$$\mathrm{SubOpt}\left(\widetilde{\pi}; x\right) \leq 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\widetilde{\Gamma}_h(x, a) \mid x_1 = x\right]$$

with probability at least $1 - \xi$. Here $\mathbb{E}_{\pi^*}$ is taken with respect to the trajectory induced by π^* in the underlying MDP given the fixed $\widetilde{\Lambda}_h$ and $\widetilde{\Sigma}_h(\widetilde{\theta}_h)$.

This theorem establishes a probabilistic upper bound on the suboptimality of the policy $\widetilde{\pi}$ produced by the GPEVI algorithm. The bound is expressed in terms of the confidence bounds $\widetilde{\Gamma}_h(x, a)$, which quantify the uncertainty in our value function estimates. The suboptimality bound scales with the horizon length H, reflecting the compounding effect of errors across time steps in sequential decision-making problems.

Corollary 2. *Under the assumptions of Theorem 1, if $\lambda_{\min}(\Lambda_h) > 0$, we have for n large enough,*

$$\mathrm{SubOpt}\left(\widetilde{\pi}; x\right) \leq O\left(\sqrt{\frac{d_r H^2 \log(H/\xi)}{n}}\right) + O\left(\sqrt{\frac{(d_p + d_r)^2 H^4 \log\left((d_p + d_r)Hn/\xi\right)}{n}}\right)$$

with probability at least $1 - \xi$. Besides,

$$\max_{h \in [H]} \|\widetilde{\theta}_h - \theta_h^*\|_2 \leq c\sqrt{\frac{d_r \log(H/\xi)}{n}}$$

holds with probability at least $1 - \xi$ for some constant $c > 0$.

The bound decreases at a rate of $O(1/\sqrt{n})$ with respect to the number of labeled samples n, which is optimal in the parametric setting under standard assumptions. The dependence on the dimensions d_r and d_p illustrates the curse of dimensionality inherent in reinforcement learning problems.

Comparison with existing work. Our theoretical bound naturally specializes to the standard linear MDP setting, enabling direct comparison with PEVI (Jin et al., 2021) while maintaining the same suboptimality rate. Here, PEVI operates under the assumption that $d_r = d_p$ with g being the identity mapping. Furthermore, while existing literature explores more general models (Xie et al., 2021; Zanette et al., 2021) that are similar to our GLMDP framework, their proposed algorithms often suffer from either computational intractability or reliance on substantially stronger assumptions. For instance, Xie et al. (2021) proposes an algorithm with detailed theoretical analysis for cases like linear function approximation, but it lacks computational feasibility, whereas Zanette et al. (2021) imposes the restrictive requirement that the Q-function must admit a linear structure.

Theorem 2 (Suboptimality for SS-GPEVI). *Under Assumptions 1-2, we set* $\lambda = 1$, $\alpha_r = c_r\sqrt{d_r \log H/\xi}$, $\alpha_p = c_p\left(d_p + d_r\right) H\sqrt{\zeta}$, *where* $\zeta = \log\left(2\left(d_r + d_p\right) Hn/\xi\right)$, $c_r, c_p > 0$ *are absolute constants and* $\xi \in (0,1)$ *is the confidence parameter. Then* $\widehat{\Gamma}_h$ *in equation 12 is a* ξ*-uncertainty quantifier of* $\widehat{B}_h$ *w.r.t. value function* $\widehat{V}_{h+1}$. *For any* $x \in \mathcal{S}$ *and* n *large enough,* $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ *in Algorithm B.2 satisfies,*

$$\mathrm{SubOpt}(\widehat{\pi}; x) \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\widetilde{\Gamma}_{r,h}(x_h, a_h) + 2\widehat{\Gamma}_h(x_h, a_h) \mid x_1 = x\right] + \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\Delta_{err} \mid x_1 = x\right]$$

with probability at least $1 - \xi$, *where* $\Delta_{err} = \widetilde{O}\left(\frac{d_r^{3/4}}{n^{3/4}}\right)$ *represents the additional error arising from the mismatch between the reward uncertainty quantifiers in the semi-supervised setting.*

Corollary 3. *Under the assumptions of Theorem 2, if* $\lambda_{min}(\Lambda_h) \geq \rho$, *then we have for* n *large enough,*

$$\mathrm{SubOpt}(\widehat{\pi}; x) \leq O\left(\sqrt{\frac{d_r H^2 \log(H/\xi)}{n}}\right) + O\left(\sqrt{\frac{\left(d_p + d_r\right)^2 H^4 \log\left(2\left(d_r + d_p\right) H(n+N)/\xi\right)}{n+N}}\right)$$

with probability at least $1 - \xi$, *which is strictly better than the bound for the supervised approach when* $N > 0$.

Corollary 3 characterizes the performance guarantees of our SS-GPEVI algorithm. The bound consists of two primary components: the first term, scaling as $\widetilde{O}\left(\sqrt{d_r H^2/n}\right)$, captures the uncertainty in reward estimation and depends solely on the size of the labeled dataset n. The second term, scaling as $\widetilde{O}\left(\sqrt{\left(d_p + d_r\right)^2 H^4/(n+N)}\right)$, reflects the uncertainty in transition dynamics estimation and crucially benefits from both labeled and unlabeled data.

A key advantage of our semi-supervised approach arises when $N \gg n$. In particular, when $d_p \gg d_r$ and $N \gg nH^2 d_p^2/d_r$, SS-GPEVI achieves a rate of $\widetilde{O}\left(\sqrt{d_r H^2/n}\right)$, which significantly outperforms the rate of a purely supervised approach, $\widetilde{O}\left(\sqrt{\left(d_p + d_r\right)^2 H^4/n}\right)$. This result rigorously demonstrates the benefits of incorporating unlabeled data in RL, especially in scenarios where labeled data are scarce or costly to obtain.

4.2 Online Reinforcement Learning: Theoretical Analysis

We now turn to the theoretical analysis of our online algorithm, GLSVI-UCB. The online setting presents additional challenges due to the need to balance exploration and exploitation while learning from sequential interactions.

Assumption 3. *There exist constants* $0 < k \leq K < \infty$ *such that* $k \leq \dot{g}(u) \leq K$ *for all* $u \in \mathbb{R}$.

Assumption 4. *For any* $h \in [H]$, *we have* $\|\theta_h^*\|_2 \leq M$ *for some known constant* $M > 0$, *and* $\left\|\mu_h(\mathcal{S})\right\| \leq \sqrt{d_p}$.

Assumption 3 ensures that the link function derivative is bounded away from zero and infinity, which is essential for the stability of the GLM estimation. Assumption 4 provides a known bound on the reward parameters, which is typical in online learning settings to ensure proper regularization.

Theorem 3 (Regret Bound for GLSVI-UCB). *Under Assumptions 3 and 4, for any fixed* $p_0 \in (0,1)$, *if we set*

$$\gamma_r = K \cdot \sqrt{4M^2 + \frac{3 + 16[d_r \ln(2MT) + \ln(3TH/p_0)]}{k}} \tag{16}$$

$$\gamma_p = c_p d_p H \sqrt{\ln(3d_p TH/p_0)} \tag{17}$$

where $c_p > 0$ is a sufficiently large absolute constant, then for any fixed initial state $x \in \mathcal{S}$, the regret of Algorithm B.3 satisfies

$$\begin{aligned}\mathcal{R}(x) &\leq H\sqrt{T}\left(\gamma_r\sqrt{2d_r\ln(1+T/d_r)} + \gamma_p\sqrt{2d_p\ln(1+T/d_p)}\right) + \sqrt{2\ln(6/p_0)TH^3} \\ &= \widetilde{O}\left(d_r + \sqrt{TH^4 d_p^3}\right)\end{aligned}$$

with probability at least $1 - p_0$.

This theorem establishes a regret bound for our online algorithm that scales with $\widetilde{O}(d_r + \sqrt{TH^4 d_p^3})$. The dependence on d_r appears only logarithmically (hidden in the $\widetilde{O}$ notation), while the dependence on d_p is more substantial. This reflects the fundamental difference in complexity between reward and transition estimation in the GLMDP framework.

Comparison with existing online RL results. Our regret bound is comparable to existing results for structured MDPs. For linear MDPs, Jin et al. (2020) achieve $\widetilde{O}(\sqrt{d^3H^4T})$ regret where d is the common feature dimension. Our bound shows that the GLMDP framework, while more expressive, maintains similar regret scaling with respect to the transition feature dimension d_p, with only logarithmic dependence on the reward feature dimension d_r. This suggests that the additional expressiveness of GLMDPs comes at minimal cost in terms of online learning performance.

The key insight from our theoretical analysis is that the modular structure of GLMDPs—separating reward and transition modeling—enables both improved sample efficiency (especially in semi-supervised settings) and maintains favorable regret properties in online learning. This demonstrates the practical value of our framework across different learning paradigms.

5 Discussion and Conclusion

This work introduces the GLMDP framework, which extends classical linear MDPs by incorporating nonlinear link functions into the reward model. This enhancement enables the modeling of a broad class of reward structures, including binary and count-value rewards, thereby addressing a critical limitation of prior linear MDP approaches. Importantly, the GLMDP framework retains the theoretical tractability of linear models while significantly broadening their applicability to real-world domains such as healthcare, recommendation systems, and finance.

A central feature of our approach is the use of **separate feature maps for rewards and transitions**, which increases modeling flexibility and enables an efficient semi-supervised learning strategy. Crucially, our method avoids the need to impute missing rewards—a major challenge in semi-supervised reinforcement learning—by estimating the transition model from both labeled and unlabeled data while using only labeled data for reward learning. Our theoretical analysis establishes that the proposed SS-GPEVI algorithm can achieve performance comparable to fully supervised methods, even when labeled data is limited.

While Assumption 2 provides cleaner theoretical bounds as shown in Theorem 1, we emphasize that analogous results can be established even in its absence. This relaxation, however, necessitates a modified estimation procedure for θ_h^*—specifically, the introduction of a ℓ_2-penalty term. We formalize this extension in Theorem M.3 in Appendix M, where we derive a suboptimality upper bound that depends on the regularization parameter, which is looser than the bound stated in Theorem 1—this represents the trade-off for relaxing this assumption.

The GLMDP framework provides an extensible foundation for generalizing a broad class of linear MDP algorithms, such as model-based (Yang and Wang, 2020), online, or offline methods (Du et al., 2019; Xiong et al., 2022), to accommodate complex reward structures while retaining computational efficiency. A key feature is its support for temporally heterogeneous rewards via step-dependent link functions. This allows for more realistic modeling in domains like clinical decision-making, where outcomes may shift from continuous vital signs to binary survival events.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning, 2019.

Qingpeng Cai, Aris Filos-Ratsikas, Pingzhong Tang, and Yiwei Zhang. Reinforcement mechanism design for e-commerce. In *Proceedings of the 2018 World Wide Web Conference*, pages 1339–1348, 2018.

Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pages 2938–2971. PMLR, 2022.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle, 2019.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20132–20145, 2021.

Daiqi Gao, Hsin-Yu Lai, Predrag Klasnja, and Susan A Murphy. Harnessing causality in reinforcement learning with bagged decision times, 2024.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings, 2018.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012. doi: 10.1214/ECP.v17-2079.

Hao Hu, Yiqin Yang, Qianchuan Zhao, and Chongjie Zhang. The provable benefits of unsupervised data sharing for offline reinforcement learning, 2023.

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

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8248–8254. IEEE, 2019.

Ksenia Konyushkova, Konrad Zolna, Yusuf Aytar, Alexander Novikov, Scott Reed, Serkan Cabi, and Nando de Freitas. Semi-supervised reward learning for offline reinforcement learning, 2020.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

Rui Lu, Gao Huang, and Simon Shaolei Du. On the power of multitask representation learning in linear mdp, 2021. URL https://api.semanticscholar.org/CorpusID:235436204.

Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1711–1724, 2022.

Peter McCullagh and A. Nelder John. *Generalized Linear Models, Section Edition*. Chapman & Hall, 1989.

Aditya Modi and Ambuj Tewari. Contextual markov decision processes using generalized linear models, 2019.

Dmitrii M. Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1):326 – 391, 2021. doi: 10.1214/20-EJS1780. URL https://doi.org/10.1214/20-EJS1780.

Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, 231:120495, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.120495. URL https://www.sciencedirect.com/science/article/pii/S0957417423009971.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Aaron Sonabend, Junwei Lu, Leo Anthony Celi, Tianxi Cai, and Peter Szolovits. Expert-supervised reinforcement learning for offline policy learning and evaluation. In *Advances in Neural Information Processing Systems*, volume 33, pages 18967–18977, 2020. URL https://proceedings.neurips.cc/paper/2020/file/daf642455364613e2120c636b5a1f9c7-Paper.pdf.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

Anna L Trella, Kelly W Zhang, Hinal Jajal, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. A deployed online reinforcement learning algorithm in an oral health clinical trial. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28792–28800, 2025.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Reinforcement learning with general value function approximation: provably efficient approach via bounded eluder dimension. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation, 2019.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game, 2022.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 13626–13640, 2021.

Wei Zhang, Yuanguo Lin, Yong Liu, Huanyu You, Pengcheng Wu, Fan Lin, and Xiuze Zhou. Self-supervised reinforcement learning with dual-reward for knowledge-aware recommendation. *Applied Soft Computing*, 131:109745, 2022.

Qinqing Zheng, Mikael Henaff, Brandon Amos, and Aditya Grover. Semi-supervised offline reinforcement learning with action-free trajectories. In *International Conference on Machine Learning*, pages 42339–42362. PMLR, 2023.

Doudou Zhou, Yufeng Zhang, Aaron Sonabend-W, Zhaoran Wang, Junwei Lu, and Tianxi Cai. Federated offline reinforcement learning. *Journal of the American Statistical Association*, 119 (548):3152–3163, 2024.

A Related works

The linear MDP model has gained substantial attention in RL due to its interpretability and favorable theoretical properties. By employing linear function approximation, this model enables generalization across large state-action spaces under the assumption of linearity in both the transition dynamics and reward functions, as defined via predefined feature maps. This structural simplicity has enabled the development of provably efficient algorithms with sublinear sample complexity (Yang and Wang, 2019; Jin et al., 2020; Duan et al., 2020; Jin et al., 2021, e.g.). Moreover, the framework has been successfully extended to multitask RL (Lu et al., 2021) and federated learning settings (Zhou et al., 2024). A key advantage of linear MDPs lies in their preservation of Q-function linearity under arbitrary policies which facilitates tractable analysis and efficient computation.

Despite these strengths, the expressive power of linear MDPs remains limited, particularly in representing non-continuous rewards, such as binary and count-like outcomes, that frequently arise in real-world applications, including healthcare, recommendation systems, and autonomous driving (Gottesman et al., 2019; Chen et al., 2019; Kendall et al., 2019). To address these limitations, recent studies have sought to enhance the flexibility of linear MDPs while retaining their theoretical benefits.

For example, Wang et al. (2019) proposed a Q-learning algorithm using GLMs to approximate the Bellman operator such that $\mathbb{E}\left[r_h\left(x_h, a_h\right) + V\left(x_h\right) \mid x_h = x, a_h = a\right] = f\left(\langle\phi(x, a), \theta_h\rangle\right)$ for any value function V, where f is a known link function and ϕ is a feature map. Their approach approximates the optimal Q-function using a link function applied to linearly combined state-action features, and maintains optimistic value estimates to encourage exploration. Under a new expressivity assumption called 'optimistic closure', they prove their algorithm achieves a regret bound of $\widetilde{O}(H^2\sqrt{d^3T})$ where d is the dimension of ϕ and T is the number of episodes, and H is the length of an episode. Furthermore, Wang et al. (2020) proposed a provably efficient algorithm with a general value function via bounded Eluder dimension which could extend linear MDP to general function classes. However the regret bound demonstrated in this paper is less tight than Jin et al. (2020) and Wang et al. (2019), where the function class is respectively restricted to linear functions and generalized linear functions.

In a complementary direction, Modi and Tewari (2019) extended GLMs to model transition probabilities while maintaining linearity for rewards, further illustrating the growing interest in structured yet expressive models. These works collectively motivate the development of new frameworks that better balance expressiveness and sample efficiency.

In parallel, deep neural networks have significantly advanced offline RL by capturing complex, non-linear relationships without reliance on hand-crafted features (Shakya et al., 2023). Conservative Q-Learning (CQL) (Kumar et al., 2020) mitigates distributional shift by conservatively estimating out-of-distribution (OOD) Q-values. Subsequent variants, such as Mildly Conservative Q-Learning (MCQ) (Lyu et al., 2022), refine this approach to better balance conservatism and generalization.

However, a critical distinction lies in the sample complexity: while linear methods enjoy explicit theoretical guarantees, including finite-sample performance bounds (Jin et al., 2021), deep networks generally require significantly more data to avoid overfitting, often scaling exponentially with model depth in worst-case scenarios. This contrast has important practical implications. In data-constrained environments, linear models may outperform deep counterparts; conversely, in data-rich scenarios, deep networks can capitalize on their greater representational power.

Hybrid approaches have emerged to bridge this gap through semi-supervised learning. Notably, Konyushkova et al. (2020) introduced one of the first semi-supervised frameworks for reward learning with limited annotations, achieving performance comparable to fully supervised methods. Building on this, Zheng et al. (2023) developed an offline RL method for action-free trajectories, using inverse dynamics models to generate proxy rewards and achieving competitive performance on standard benchmarks with as little as 10% labeled data.

Theoretical support for these methods has been provided by Hu et al. (2023), who established performance guarantees for semi-supervised RL under reduced labeling regimes. Unlike approaches reliant on inverse dynamics or pseudo-labeling (Zhang et al., 2022), our framework decouples the reward and transition models, thereby eliminating the need for reward imputation in unlabeled trajectories.

This design aligns with the minimalist principle advocated by Fujimoto and Gu (2021), which emphasizes that simple modifications to standard RL pipelines can rival complex offline methods. We extend this perspective by integrating the pessimistic value iteration strategy (Jin et al., 2021; Xie and Jiang, 2021) with a semi-supervised learning paradigm, offering a unified solution that is practical, statistically efficient, and algorithmically simple.

B Algorithm Pseudocode

This section provides the detailed pseudocode for the algorithms discussed in Section 3.

Algorithm B.1 Generalized PEssimistic Value Iteration (GPEVI)

Input: Dataset $\mathcal{D} = \big\{(x_h^\tau, a_h^\tau, r_h^\tau)\big\}_{\tau,h=1}^{n,H}$; hyperparameters λ, α_r, α_p, ξ.
Initialization: set $\widetilde{V}_{H+1}(x) \leftarrow 0$.
for step $h = H, H-1, \ldots, 1$ **do**
  Obtain $\widetilde{\theta}_h$ from equation 7 and $\widetilde{\beta}_h$ from equation 8.
  Set $\widetilde{\Gamma}_h(\cdot,\cdot)$ as equation 10.
  Set $\widetilde{Q}_h(x,a) \leftarrow \min\Big\{g\big(\phi_r(x,a)^\intercal\widetilde{\theta}_h\big) + \phi_p(x,a)^\intercal\widetilde{\beta}_h - \widetilde{\Gamma}_h(x,a), H-h+1\Big\}^+$.
  Set $\widetilde{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \big\langle \widetilde{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\big\rangle_{\mathcal{A}}$.
  Set $\widetilde{V}_h(\cdot) \leftarrow \big\langle \widetilde{Q}_h(\cdot,\cdot), \widetilde{\pi}_h(\cdot \mid \cdot)\big\rangle_{\mathcal{A}}$.
Output: $\widetilde{\pi} = \{\widetilde{\pi}_h\}_{h=1}^H$.

Algorithm B.2 Semi-Supervised Generalized PEssimistic Value Iteration (SS-GPEVI)

Input: Labeled dataset $\mathcal{D}$, unlabeled dataset $\mathcal{D}_u$; hyperparameters λ, α_r, α_p, ξ.
Initialization: set $\widehat{V}_{H+1}(x) \leftarrow 0$.
for step $h = H, H-1, \ldots, 1$ **do**
  Obtain $\widetilde{\theta}_h$ from equation 7 using $\mathcal{D}$.
  Obtain $\widehat{\beta}_h$ from equation 11 using both $\mathcal{D}$ and $\mathcal{D}_u$.
  Set $\widehat{\Gamma}_h(\cdot,\cdot)$ as equation 12.
  Set $\widehat{Q}_h(x,a) \leftarrow \min\big\{g\big(\phi_r(x,a)^\intercal\widetilde{\theta}_h\big) + \phi_p(x,a)^\intercal\widehat{\beta}_h - \widehat{\Gamma}_h(x,a), H-h+1\big\}^+$.
  Set $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \big\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\big\rangle_{\mathcal{A}}$.
  Set $\widehat{V}_h(\cdot) \leftarrow \big\langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot \mid \cdot)\big\rangle_{\mathcal{A}}$.
Output: $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$.

Algorithm B.3 Generalized Least Square Value Iteration with UCB (GLSVI-UCB).

Input: hyperparameter γ_r, γ_p.
Initialize estimates $\bar{Q}_{h,0} \equiv H$ for all $h \le H$ and $\bar{Q}_{H+1,t} \equiv 0$ for all $1 \le t \le T$;
for $t = 1, 2, \cdots, T$ **do**
  Commit to policy $\hat{\pi}_{h,t}(x) \coloneqq \arg\max_{a\in\mathcal{A}} \bar{Q}_{h,t-1}(x,a)$;
  Use policy $\hat{\pi}_{\cdot,t}$ to collect one trajectory $\{(x_{h,t}, a_{h,t}, r_{h,t})\}_{h=1}^H$ where we start with the initial state x when $t = 1$;
  for $h = H, H-1, \cdots, 1$ **do**
    Set $\widehat{\theta}_{h,t}$ as equation 13.
    Set $\widehat{\beta}_{h,t}$ as equation 14.
    Set $\bar{Q}_{h,t}(x,a)$ as equation 15.

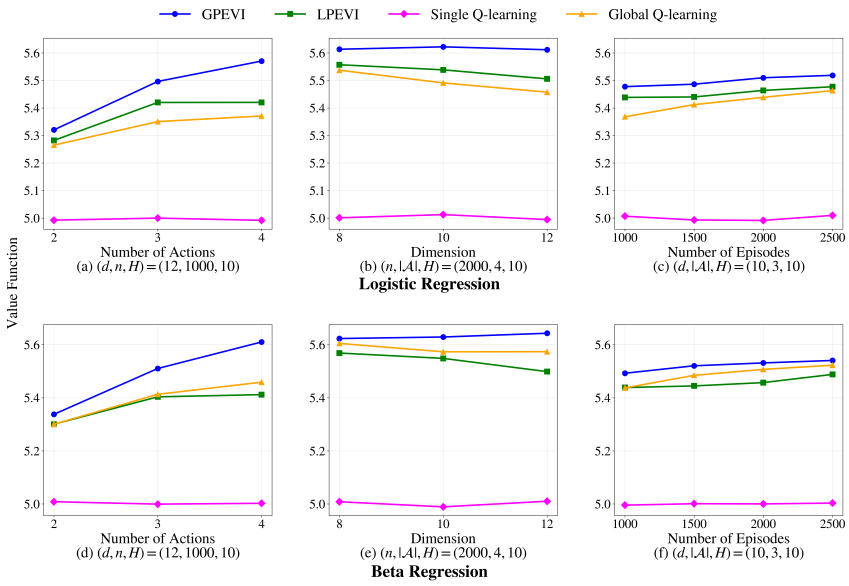

Figure 1: Experimental results for fully labeled data across different parameter configurations

C Simulation Studies

C.1 Full labeled data

We conduct comprehensive experimental evaluations to assess the performance of our proposed methods across varying dimensions, action space cardinalities, and episode counts. Our experiments focus on two fundamental tasks: logistic regression and beta regression.

Logistic regression and beta regression experiments utilize the logit link function and generate simulation data using a consistent Markov Decision Process framework. For each timestep $h \in [H]$, we sample random parameter vectors $\theta_h \in \mathbb{R}^d$ from an element-wise Uniform$(-0.5, 0.5)$ distribution. We generate rewards using two distinct probability distributions: a binomial distribution $r_h \sim \text{Binomial}(1, \text{sigmoid}(\phi(x_h, a_h)^T \theta_h))$ for logistic regression tasks and a beta distribution $r_h \sim \text{Beta}(\text{sigmoid}(\phi(x_h, a_h)^T \theta_h), 1 - \text{sigmoid}(\phi(x_h, a_h)^T \theta_h))$ for beta regression tasks, where $\phi(x_h, a_h)$ represents our feature mapping function that incorporates state-action interactions and normalizes state vectors.

Throughout our simulations, we maintain consistency by using identical mapping functions ϕ for both reward (ϕ_r) and transition probability (ϕ_p) modeling, as well as uniform state dimensions ($d_r = d_p = d$). Our feature mapping pipeline first normalizes states by their L2 norm, then constructs a sparse representation where only elements corresponding to the selected action are non-zero, yielding a feature vector of size $d \cdot |\mathcal{A}|$, where d denotes the state dimension and $|\mathcal{A}|$ represents the cardinality of the action space.

For state transitions, we employ a rejection sampling methodology where candidate next states are sampled from Uniform$(-0.5, 0.5)^d$ and accepted with probability:

$$\alpha = \min\left(1, \frac{\langle x_h \cdot (a_h + 1) + a_h/d, \exp(-x_{h+1}) \rangle}{\sum x_{h+1} \cdot (a_h + 1) + a_h}\right) \tag{C.1}$$

where x_h represents the current state, a_h denotes the selected action, $\sum x_{h+1}$ indicates the scalar value obtained by summing all components of the state vector x_{h+1}, and x_{h+1} represents the proposed next state.

Our experimental design spans multiple parameter configurations: action space cardinalities $|\mathcal{A}| \in \{2, 3, 4\}$, dimensionalities $d \in \{8, 10, 12\}$, and episode counts $n \in \{1000, 1500, 2000, 2500\}$.

We implement and compare the following methods to validate our Algorithm B.1: (1) GPEVI (our proposed method), (2) LPEVI (Linear PEssimistic Value Iteration), (3) single Q-learning, and (4) global Q-learning. The LPEVI method approximates the value function using linear regression following Jin et al. (2021), employing ordinary least squares to estimate Q-functions that are linear in $\phi(x, a)$. Single Q-learning utilizes a single Q-function across all timesteps, while global Q-learning trains a unified Q-function using trajectory data from all timesteps.

Based on our theoretical analysis in Section 4, we set the regularization parameter $\lambda = 1$. The parameter ξ, which defines the probability bounds for suboptimality guarantees, is set to $\xi = 0.01$. For simplicity, we use identical values for the hyperparameters c_r and c_p in both Algorithm B.1 and Algorithm B.2. We employ 5-fold cross-validation to determine the optimal hyperparameter c from the set $\{0.005, 0.001, 0.0005, 0.0001\}$ using the training dataset and the step-importance sampling estimator (Gottesman et al., 2018; Thomas and Brunskill, 2016).

For data generation, we adopt a combined policy approach where actions are selected optimally with 70% probability and randomly with 30% probability, ensuring balanced exploration and exploitation in the training data. For evaluation, we use a test dataset of size 250. Each simulation is repeated 100 times to ensure statistical significance.

Figure 1 presents our comprehensive experimental results for logistic and beta regression. Across all parameter configurations—varying $|\mathcal{A}|$, d, and n—GPEVI consistently demonstrates superior performance in terms of mean value compared to baseline methods.

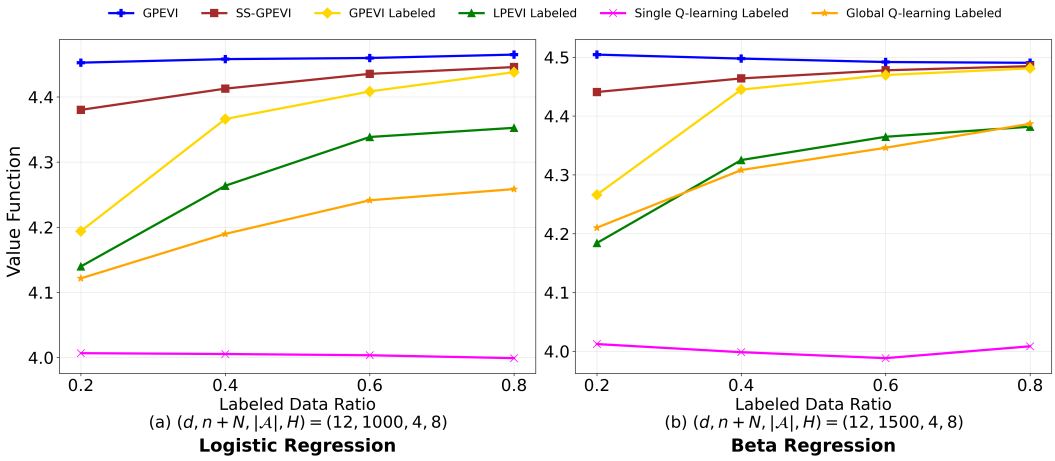

Figure 2: Experimental results for semi-supervised learning across different labeled data ratios

C.2 Semi-Supervised Learning

To evaluate the effectiveness of our proposed Algorithm B.2, we conduct experiments in semi-supervised learning settings. We compare the following methods: (1) GPEVI with the full dataset of $n + N$ samples treated as if all were labeled, (2) SS-GPEVI that properly differentiates between the n labeled and N unlabeled samples, (3) GPEVI trained using only the n labeled samples, (4) LPEVI trained using only the n labeled samples, (5) single Q-learning trained using only the n labeled samples, and (6) global Q-learning trained using only the n labeled samples.

Our experimental configuration for logistic regression sets $d = 12$, total dataset size $n + N = 1000$, action space cardinality $|\mathcal{A}| = 4$, and horizon $H = 8$. For beta regression tasks, we use $d = 12$, $n + N = 1500$, $|\mathcal{A}| = 4$, and $H = 8$. The labeled data ratio is defined as $\frac{n}{n+N}$, where n represents the number of labeled samples and N the number of unlabeled samples. For both data generation and evaluation, we follow the same procedures used in the fully labeled setting.

Figure 2 presents our results across varying labeled data ratios for logistic and beta regression. As expected, GPEVI with complete data (assuming all samples are labeled) achieves the highest performance across all experimental conditions. However, our proposed SS-GPEVI demonstrates remarkably competitive performance, closely approaching that of the fully supervised variant while

substantially outperforming all baseline methods that utilize only labeled data. This validates the efficacy of our semi-supervised approach in effectively leveraging unlabeled data.

D Simulation Environment Study

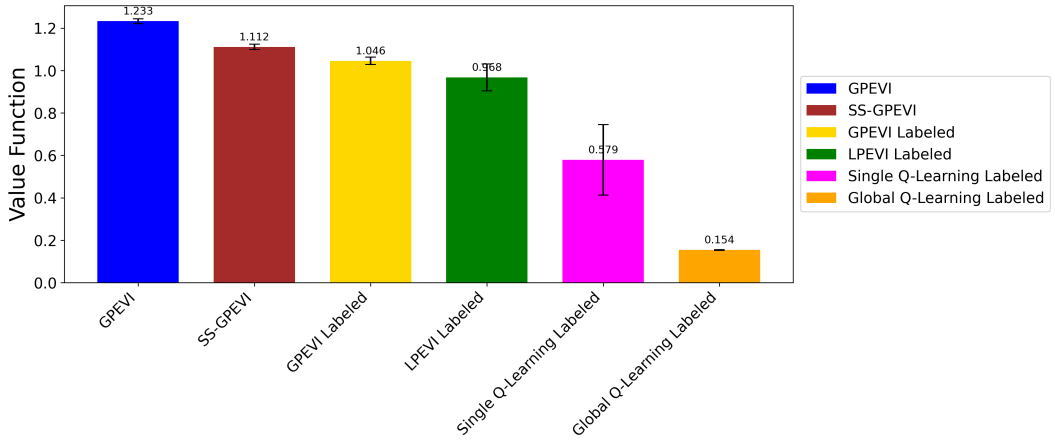

Figure 3: Experimental results on PointMaze dataset with labeled dataset size $n = 1000$ and unlabeled dataset size $N = 1500$. Error bars represent standard deviations across 100 independent runs.

To validate the practical applicability of our proposed methods, we conduct experiments on the PointMaze offline reinforcement learning benchmark datasets. Specifically, we utilize the PointMaze Medium Dense-v3 simulation environment, where an agent follows waypoints generated through Q-Iteration using a PD controller until successfully reaching designated goal locations (Fu et al., 2020).

The simulation environment features a continuous task structure where the agent maintains its current position upon reaching a goal, while the environment generates a new random goal location, creating an ongoing navigation challenge. The reward structure employs a dense reward function, calculated as the negative exponential of the Euclidean distance between the agent's current position and the target goal. To ensure diverse trajectory exploration and increase path variance, random Gaussian noise is injected into the agent's action selection process.

The original dataset comprises $4,752$ episodes with a 2-dimensional continuous action space. To align with our discrete action framework, we discretize the action dimension into 8 distinct actions, as required by our algorithm. For computational efficiency, we truncate episodes to a maximum horizon of $H = 25$ timesteps, retaining only the first 25 steps of longer episodes. The state representation has dimensionality $d = 4$.

Given that the reward values are bounded in the interval $(0, 1)$, we employ beta regression with a logit link function to approximate the value function, which provides a more appropriate probabilistic modeling framework for bounded outcomes compared to traditional linear regression approaches.

For our experimental setup, we allocate $n = 1000$ labeled samples and $N = 1500$ unlabeled samples for training, while reserving a separate test set of size 250 for evaluation. We compare the following approaches: (1) GPEVI with the full dataset of $n + N$ samples treated as if all were labeled, (2) SS-GPEVI that properly differentiates between the n labeled and N unlabeled samples, (3) GPEVI trained using only the n labeled samples, (4) LPEVI trained using only the n labeled samples, (5) single Q-learning trained using only the n labeled samples, and (6) global Q-learning trained using only the n labeled samples. To ensure statistical reliability, all experiments are repeated 100 times.

Performance comparison is based on estimated value functions computed via a step-importance sampling estimator (Gottesman et al., 2018; Thomas and Brunskill, 2016). The results, summarized in Figure 3, demonstrate that our proposed methods consistently outperform baseline approaches. Specifically, GPEVI with all $n + N$ samples treated as labeled (representing an idealized scenario with complete reward knowledge) achieves an average estimated value of 1.233, our SS-GPEVI

(properly using n labeled and N unlabeled samples) achieves 1.112, while GPEVI utilizing only the n labeled samples reaches 1.046. These results substantially exceed the performance of LPEVI and Q-learning baselines. Notably, our SS-GPEVI outperforms the labeled-only GPEVI counterpart, aligning with our theoretical insights on the benefits of incorporating unlabeled data. Additionally, all variants of our method exhibit low standard deviations across runs, demonstrating robustness and consistency in performance.

E Discussion on Unbounded Reward Functions

Assumption E.1. *The reward noise is sub-Gaussian; that is, for all $x \in \mathcal{S}$ and $a \in \mathcal{A}$, the random variable $r_h(x,a) - g(\langle \phi_r(x,a), \theta_h^\star \rangle)$ is sub-Gaussian.*

Assumption E.1 guarantees well-behaved reward noise with desirable concentration properties. Compared to existing literature (e.g., Jin et al. (2021); Xie et al. (2021)) that typically assumes bounded rewards for analytical simplicity, our sub-Gaussian condition represents a strictly weaker requirement. Moreover, when rewards are bounded, Assumption E.1 is naturally satisfied.

In contrast to Jin et al. (2021), which constrains rewards to the interval $[0,1]$, our framework accommodates arbitrary reward ranges, necessitating the standardization of function g in Algorithm B.1. To formalize this extension, we take $g_{\max}$ as an arbitrary constant larger than $\sup_{|x| \le \sup_{h\in[H]} \|\theta_h^*\|_2} g(x)$ and $g_{\min}$ as an arbitrary constant smaller than $\inf_{|x| \le \sup_{h\in[H]} \|\theta_h^*\|_2} g(x)$. We then establish the normalized uncertainty bound:

$$\widetilde{\Gamma}_{h,nrm} = \frac{\widetilde{\Gamma}_h}{g_{\max} - g_{\min}} = \frac{\widetilde{\Gamma}_{r,h} + \widetilde{\Gamma}_{p,h}}{g_{\max} - g_{\min}} = \widetilde{\Gamma}_{r,h,nrm} + \widetilde{\Gamma}_{p,h,nrm} \tag{E.2}$$

This normalization enables us to define the normalized Q-function and its corresponding value function as:

$$\widetilde{Q}_{h,nrm}(x,a) = \min\left\{ \left(\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1}\right)(x,a)_{nrm} - \widetilde{\Gamma}_{h,nrm}(x,a), H-h+1 \right\}^+$$

$$\widetilde{V}_{h,nrm}(x) = \left\langle \widetilde{Q}_{h,nrm}(x,\cdot), \widetilde{\pi}_{h,nrm}(\cdot \mid x) \right\rangle_{\mathcal{A}}$$

where the normalized reward function is defined as:

$$g_{nrm}\left(\phi_r(x,a)^\intercal \widetilde{\theta}_h\right) = \frac{g\left(\phi_r(x,a)^\intercal \widetilde{\theta}_h\right) - g_{\min}}{g_{\max} - g_{\min}}.$$

The normalized Bellman operator is defined as:

$$\left(\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1}\right)(x,a)_{nrm} = g_{nrm}\left(\phi_r(x,a)^\top \widetilde{\theta}_h\right) + \phi_p(x,a)^\top \widetilde{\beta}_{h,nrm},$$

where

$$\widetilde{\beta}_{h,nrm} \coloneqq \sum_{\tau=1}^{n} (\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \phi_p(x_h^\tau, a_h^\tau) \widetilde{V}_{h+1,nrm}(x_{h+1}^\tau). \tag{E.3}$$

and the normalized policy:

$$\widetilde{\pi}_{h,nrm}(\cdot \mid x) = \arg\max_{\pi_h} \left\langle \widetilde{Q}_{h,nrm}(x,\cdot), \pi_h(\cdot \mid x) \right\rangle_{\mathcal{A}}$$

Based on these definitions, we extend the GPEVI algorithm to handle unbounded rewards in Algorithm E.4. Similarly, for the semi-supervised variant (SS-GPEVI), we define the corresponding normalized uncertainty quantifier:

$$\widehat{\Gamma}_{h,nrm} = \frac{\widehat{\Gamma}_h}{g_{\max}} = \frac{\widetilde{\Gamma}_{r,h} + \widehat{\Gamma}_{p,h}}{g_{\max} - g_{\min}} = \widetilde{\Gamma}_{r,h,nrm} + \widehat{\Gamma}_{p,h,nrm} \tag{E.4}$$

and

$$\widehat{\beta}_{h,nrm} \coloneqq \sum_{\tau=1}^{n+N} (\widehat{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \phi_p(x_h^\tau, a_h^\tau) \widehat{V}_{h+1,nrm}(x_{h+1}^\tau)\,, \tag{E.5}$$

The complete procedures for both approaches are systematically presented in Algorithm E.4 and Algorithm E.5, respectively.

Algorithm E.4 GPEVI for Unbounded Rewards

1: Input: Dataset $\mathcal{D} = \left\{(x_h^\tau, a_h^\tau, r_h^\tau)\right\}_{\tau,h=1}^{n,H}$; hyperparameters λ, α_r, α_p, ξ.
2: Initialization: set $\widetilde{V}_{H+1,nrm}(x) \leftarrow 0$.
3: **for** step $h = H, H-1, \ldots, 1$ **do**
4: Obtain $\widetilde{\theta}_h$ from equation 7 and $\widetilde{\beta}_{h,nrm}$ from equation E.3.
5: Set $\widetilde{\Gamma}_{h,nrm}(\cdot,\cdot)$ as equation E.2.
6: Set $\widetilde{Q}_{h,nrm}(x,a) \leftarrow \min\left\{g_{nrm}\left(\phi_r(x,a)^\top \widetilde{\theta}_h\right) + \phi_p(x,a)^\intercal \widetilde{\beta}_{h,nrm} - \widetilde{\Gamma}_{h,nrm}(x,a), H-h+1\right\}^+$.
7: Set $\widetilde{\pi}_{h,nrm}(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \left\langle \widetilde{Q}_{h,nrm}(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$.
8: Set $\widetilde{V}_{h,nrm}(\cdot) \leftarrow \left\langle \widetilde{Q}_{h,nrm}(\cdot,\cdot), \widetilde{\pi}_{h,nrm}(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$.
9: Output: $\widetilde{\pi}_{nrm} = \{\widetilde{\pi}_{h,nrm}\}_{h=1}^H$.

Algorithm E.5 SS-GPEVI for Unbounded Rewards

1: Input: Labeled dataset $\mathcal{D}$, unlabeled dataset $\mathcal{D}_u$; hyperparameters λ, α_r, α_p, ξ.
2: Initialization: set $\widehat{V}_{H+1,nrm}(x) \leftarrow 0$.
3: **for** step $h = H, H-1, \ldots, 1$ **do**
4: Obtain $\widetilde{\theta}_h$ from equation 7 using $\mathcal{D}$.
5: Obtain $\widehat{\beta}_{h,nrm}$ from equation E.5 using both $\mathcal{D}$ and $\mathcal{D}_u$.
6: Set $\widehat{\Gamma}_{h,nrm}(\cdot,\cdot)$ as equation E.4.
7: Set $\widehat{Q}_{h,nrm}(x,a) \leftarrow \min\left\{g_{nrm}\left(\phi_r(x,a)^\top \widetilde{\theta}_h\right) + \phi_p(x,a)^\intercal \widehat{\beta}_{h,nrm} - \widehat{\Gamma}_{h,nrm}(x,a), H-h+1\right\}^+$.
8: Set $\widehat{\pi}_{h,nrm}(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \left\langle \widehat{Q}_{h,nrm}(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$.
9: Set $\widehat{V}_{h,nrm}(\cdot) \leftarrow \left\langle \widehat{Q}_{h,nrm}(\cdot,\cdot), \widehat{\pi}_{h,nrm}(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$.
10: Output: $\widehat{\pi}_{nrm} = \{\widehat{\pi}_{h,nrm}\}_{h=1}^H$.

We could also get similar theory guarantees for these two algorithms as follows:

Theorem E.1. *Under Assumptions 1, 2 and E.1, we set* $\lambda = 1$, $\alpha_r = c_r\sqrt{d_r \log H/\xi}$, $\alpha_p = c_p(g_{\max} - g_{\min})\,(d_p + d_r)\,H\sqrt{\zeta}$, *where* $\zeta = \log\left(2\,(d_r + d_p)\,Hn/\xi\right)$, $c_r, c_p > 0$ *are absolute constants and* $\xi \in (0,1)$ *is the confidence parameter. Then* $\widetilde{\Gamma}_{h,nrm}$ *in equation E.2 is a* ξ*-uncertainty quantifier of* $\widetilde{\mathbb{B}}_h$ *w.r.t. value function* $\widetilde{V}_{h+1,nrm}$. *For any* $x \in \mathcal{S}$ *and* n *large enough,* $\widetilde{\pi}_{nrm} = \{\widetilde{\pi}_{h,nrm}\}_{h=1}^H$ *in Algorithm E.4 satisfies*

$$\mathrm{SubOpt}\left(\widetilde{\pi}_{nrm}; x\right) \le 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\widetilde{\Gamma}_h(x,a) \mid x_1 = x\right]$$

with probability at least $1-\xi$. *Here* $\mathbb{E}_{\pi^*}$ *is taken with respect to the trajectory induced by* π^* *in the underlying MDP given the fixed* $\widehat{\Lambda}_h$ *and* $\widehat{\Sigma}_h(\widetilde{\theta}_h)$.

Corollary E.1. *Under the assumptions of Theorem 1, if* $\lambda_{\min}(\Lambda_h) > 0$, *we have for* n *large enough,*

$$\begin{aligned}\mathrm{SubOpt}\left(\widetilde{\pi}_{nrm}; x\right) \le\; & O\left(\sqrt{\frac{d_r H^2 \log(H/\xi)}{n}}\right) \\ & + O\left(\sqrt{\frac{(g_{\max} - g_{\min})^2 (d_p + d_r)^2 H^4 \log\left((d_p + d_r) Hn/\xi\right)}{n}}\right)\end{aligned}$$

with probability at least $1-\xi$.

Theorem E.2. *Under Assumptions 1, 2 and E.1, we set* $\lambda = 1$, $\alpha_r = c_r\sqrt{d_r \log H/\xi}$, $\alpha_p = c_p(g_{\max} - g_{\min})\,(d_p + d_r)\,H\sqrt{\zeta}$, *where* $\zeta = \log\left(2\,(d_r + d_p)\,Hn/\xi\right)$, $c_r, c_p > 0$ *are absolute constants and* $\xi \in (0,1)$ *is the confidence parameter. Then* $\widehat{\Gamma}_h$ *in equation E.4 is a* ξ*-uncertainty quantifier of* $\widehat{\mathbb{B}}_h$ *w.r.t. value function* $\widetilde{V}_{h+1,nrm}$. *For any* $x \in \mathcal{S}$ *and* n *large enough,* $\widehat{\pi}_{nrm} = \{\widehat{\pi}_{h,nrm}\}_{h=1}^{H}$ *in Algorithm E.5 satisfies,*

$$\mathrm{SubOpt}(\widehat{\pi}_{nrm}; x) \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\widetilde{\Gamma}_{r,h}(x_h, a_h) + 2\widehat{\Gamma}_h(x_h, a_h) \mid x_1 = x\right] + \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}_{nrm}}\left[\Delta_{err} \mid x_1 = x\right]$$

with probability at least $1 - \xi$, *where* $\Delta_{err} = \widetilde{O}\left(\frac{d_r^{3/4}}{n^{3/4}}\right)$ *represents the additional error arising from the mismatch between the reward uncertainty quantifiers in the semi-supervised setting. Specifically,* Δ_{err} *accounts for the difference between using* $\widetilde{\theta}_h$ *(estimated from labeled data) and* θ_h^* *(the true parameter) in the uncertainty quantification when constructing the pessimistic value functions.*

Corollary E.2. *Under the assumptions of Theorem E.2, if* $\lambda_{min}(\Lambda_h) \geq \rho$, *then we have for* n *large enough,*

$$\mathrm{SubOpt}(\widehat{\pi}_{nrm}; x) \leq O\left(\sqrt{\frac{d_r H^2 \log(H/\xi)}{n}}\right) + O\left(\sqrt{\frac{(g_{\max} - g_{\min})^2\,(d_p + d_r)^2\,H^4 \log\left(2\,(d_r + d_p)\,H(n+N)/\xi\right)}{n+N}}\right)$$

with probability at least $1 - \xi$, *which is strictly better than the bound for the supervised approach when* $N > 0$.

Impact of Reward Scale on Theoretical Guarantees. Corollaries E.1 and E.2 reveal a critical insight: the suboptimality bounds for both algorithms exhibit explicit dependence on the range of rewards, $(g_{max} - g_{min})$, in the second term. This dependence emerges from the normalization procedure and has important implications. Particularly, for problems with large reward ranges, the second term in the bound may dominate, potentially resulting in performance degradation. This observation aligns with intuition—in settings where rewards vary dramatically, accurately estimating the transition dynamics becomes more challenging as errors are amplified by the reward scale.

Semi-Supervised Advantage with Unbounded Rewards. The advantage of the semi-supervised approach, as quantified in Corollary E.2, persists in the unbounded reward setting, with the crucial benefit that the term containing $(g_{\max} - g_{\min})$ benefits from the enlarged sample size $(n + N)$. This suggests that semi-supervised learning provides particularly significant advantages in unbounded reward scenarios, as the reduction in uncertainty regarding transition dynamics helps mitigate the amplification effect of large reward ranges. Specifically, when $N \gg n$ and $d_p \gg d_r$, the second term in the bound is substantially reduced compared to the supervised approach, yielding performance improvements that scale with both the reward range and the ratio of unlabeled to labeled data.

F Proof of Proposition 1

GLMDP is Bellman complete with respect to the following function class

$$\mathcal{F} = \{x, a \mapsto g(\langle \phi_r(x,a), \theta\rangle + \langle \phi_p(x,a), \beta\rangle : \theta \in \mathbb{R}^{d_r}, \beta \in \mathbb{R}^{d_p}\}$$

In other words, the optimal Q-value function $Q_h^* \in \mathcal{F}$ *for all* $h \in [H]$.

Proof. We define the optimal Bellman operator w.r.t to some policy π by

$$\mathbb{B}_h^{\pi} Q(x,a) \coloneqq \mathbb{E}[r_h(x,a) + Q(x_{h+1}, \pi(x_{h+1})) \mid x_h = x, a_h = a].$$

Bellman completeness requires for all $f \in \mathcal{F}$ and π, $\mathbb{B}_h^\pi f \in \mathcal{F}$.

$$\begin{aligned}
\mathbb{B}_h^\star f(x,a) &= \mathbb{E}[R_h(x_h,a_h) + f(x_{h+1},\pi(x_{h+1})) \mid x_h = x, a_h = a] \\
&= g(\langle \phi_r(x,a), \theta_h^* \rangle) + \mathbb{E}[f(x_{h+1},\pi(x_{h+1}) \mid x_h = x, a_h = a] \\
&= g(\langle \phi_r(x,a), \theta_h^* \rangle) + \int_{x'} f(x',\pi(x'))P(x' \mid x,a)dx \\
&= g(\langle \phi_r(x,a), \theta_h^* \rangle) + \int_{x'} f(x',\pi(x'))\langle \phi_p(x,a), \mu_h(x') \rangle dx' \\
&= g(\langle \phi_r(x,a), \theta_h^* \rangle) + \left\langle \phi_p(x,a), \int_{x'} f(x',\pi(x'))\mu_h(x')dx' \right\rangle .
\end{aligned}$$

Thus, we have $\mathbb{B}_h^\star f \in \mathcal{F}$ with parameters θ_h^* and $\int_{x'} f(x',\pi(x'))\mu_h(x')dx'$.

The realizability is guaranteed by Bellman completeness. □

G Proof of Theorems 1 and E.1

We only need to prove Theorem $E.1$ since we can choose $g_{\max} = 1$ given rewards are bounded by $[0,1]$. In addition, Without loss of generality, we can assume $g_{\max} = 1$. Henceforth, for notational simplicity, we omit the subscript "nrm". This convention is consistently maintained throughout Sections J, I, J and M.

Proof. By Lemma 3.1 of Jin et al. (2021), we can decompose $\mathrm{SubOpt}\left(\widetilde{\pi}; x\right)$ into three parts:

$$\begin{aligned}
\mathrm{SubOpt}\left(\widetilde{\pi}; x\right) = &\underbrace{-\sum_{h=1}^{H} \mathbb{E}_{\widetilde{\pi}}\left[\iota_h(x_h,a_h) \mid x_1 = x\right]}_{\text{(A): Spurious Correlation}} + \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\iota_h(x_h,a_h) \mid x_1 = x\right]}_{\text{(B): Intrinsic Uncertainty}} \\
&+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\langle \widetilde{Q}_h(x_h,\cdot), \pi_h^*(\cdot \mid x_h) - \widetilde{\pi}_h(\cdot \mid x_h)\rangle_{\mathcal{A}} \mid x_1 = x\right]}_{\text{(C): Optimization Error}},
\end{aligned}$$

where $\iota_h(x,a) = (\mathbb{B}_h \widetilde{V}_{h+1})(x,a) - \widetilde{Q}_h(x,a)$. By the definition of $\widetilde{\pi}_h$, we have (C) ≤ 0. We then show that with probability at least $1-\xi$,

$$0 \leq \iota_h(x,a) \leq 2\widetilde{\Gamma}_h(x,a) \text{ for all } (x,a) \in \mathcal{S} \times \mathcal{A}\,, \tag{G.6}$$

which implies the conclusion of the theorem that

$$\mathrm{SubOpt}\left(\widetilde{\pi}; x\right) \leq 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\Gamma_h(x,a) \mid x_1 = x\right].$$

Note that Lemma 5.1 of Jin et al. (2021) still holds if $g_{\max} = 1$. Hence to show equation G.6, we only need to show that $\{\widetilde{\Gamma}_h\}_{h=1}^H$ are ξ-uncertainty quantifiers such that

$$\left|\left(\mathbb{B}_h \widetilde{V}_{h+1}\right)(x,a) - \left(\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1}\right)(x,a)\right| \leq \widetilde{\Gamma}_h(x,a) \text{ for all } (x,a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$$

with probability at least $1-\xi$. By the definition of $\mathbb{B}_h$, we have

$$\begin{aligned}
\left(\mathbb{B}_h \widetilde{V}_{h+1}\right)(x,a) =& \mathbb{E}\left[r_h\left(x_h,a_h\right) + \widetilde{V}_{h+1}(x_{h+1}) \mid x_h = x, a_h = a\right] \\
=& \mathbb{E}\left[r_h(x_h,a_h) \mid x_h = x, a_h = a\right] + \int_{x' \in \mathcal{S}} \widetilde{V}_{h+1}(x')\mathbb{P}_h(x' \mid x_h = x, a_h = a)\mathrm{d}x' \\
=& g(\langle \phi_r(x,a), \theta_h^\star \rangle) + \langle \phi_p(x,a), \beta_h \rangle\,,
\end{aligned}$$

where $\beta_h = \int_{x' \in \mathcal{S}} \mu_h(x') \widetilde{V}_{h+1}(x') \mathrm{d}x'$. Then we have

$$\big(\mathbb{B}_h \widetilde{V}_{h+1}\big)(x,a) - \big(\widetilde{\mathbb{B}}_h \widetilde{V}_{h+1}\big)(x,a) = \underbrace{g(\langle \phi_r(x,a), \theta_h^\star \rangle) - g(\langle \phi_r(x,a), \widetilde{\theta}_h \rangle)}_{\text{(i)}} + \underbrace{\langle \phi_p(x,a), \beta_h - \widetilde{\beta}_h \rangle}_{\text{(ii)}} .$$

By Lemma L.3, we have $|\text{(i)}| \le \widetilde{\Gamma}_{r,h}(x,a)$ with probability at least $1 - \frac{\xi}{2}$. We then bound (ii). For notional simplicity, we define $\Omega_h = (\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1}$. By the definition of $\widetilde{\beta}_h$, we have

$$\begin{aligned}
\text{(ii)} =& \phi_p(x,a)^\intercal \beta_h - \phi_p(x,a)^\intercal \Omega_h \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau) \widetilde{V}_{h+1}(x_{h+1}^\tau) \\
=& \underbrace{\phi_p(x,a)^\intercal \beta_h - \phi_p(x,a)^\intercal \Omega_h \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau) \phi_p(x_h^\tau, a_h^\tau)^\intercal \beta_h}_{\text{(iii)}} \\
& \underbrace{- \phi_p(x,a)^\intercal \Omega_h \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau) \big(\widetilde{V}_{h+1}(x_{h+1}^\tau) - \phi_p(x_h^\tau, a_h^\tau)^\intercal \beta_h \big)}_{\text{(iv)}} .
\end{aligned}$$

Then by Lemma L.4, we have $\|\beta_h\|_2 \le H\sqrt{d_p}$, and

$$\begin{aligned}
|\text{(iii)}| =& \big| \phi_p(x,a)^\intercal \beta_h - \phi_p(x,a)^\intercal (\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \widetilde{\Lambda}_h \beta_h \big| = \big| \lambda \phi_p(x,a)^\intercal (\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \beta_h \big| \\
\le& H\sqrt{\lambda d_p} \sqrt{\phi_p(x,a)^\intercal (\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \phi_p(x,a)} .
\end{aligned}$$

We then bound $|\text{(iv)}|$. To simplify the notation, for $h \in [H]$ and $\tau \in [n]$, and any value function $V : \mathcal{S} \to [0, H]$, we define

$$\epsilon_h^\tau(V) = V(x_{h+1}^\tau) - \mathbb{E}\big[V(x_{h+1}) \mid x_h = x_h^\tau, a_h = a_h^\tau \big] .$$

We then have

$$\begin{aligned}
|\text{(iv)}| =& \phi_p(x,a)^\intercal \Omega_h \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau) \epsilon_h^\tau(\widetilde{V}_{h+1}) \\
\le& \|\phi_p(x,a)^\intercal\|_{\Omega_h} \underbrace{\Big\| \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau) \epsilon_h^\tau(\widetilde{V}_{h+1}) \Big\|_{\Omega_h}}_{\text{(v)}} .
\end{aligned}$$

We then bound term (v) via concentration inequalities. An obstacle is that $\widetilde{V}_{h+1}$ depends on $\big\{(x_h^\tau, a_h^\tau)\big\}_{\tau=1}^n$ via $\big\{(x_{h'}^\tau, a_{h'}^\tau)\big\}_{\tau \in [n], h' > h}$, as it is constructed based on the dataset $\mathcal{D}$. To this end, we resort to uniform concentration inequalities. Specifically, for all $h \in [H]$, we define the function class

$$\begin{aligned}
&\mathcal{V}_h(R, B, J_r, J_p, \rho, \lambda) = \Big\{ V_h(x; \theta, \beta, \Sigma, \Lambda, \gamma_r, \gamma_p) : \mathcal{S} \to [0, H] \text{ with} \\
&\qquad \|\theta\|_2 \le R, \|\beta\|_2 \le B, \gamma_r \in [0, J_r], \gamma_p \in [0, J_p], \Sigma \succeq \rho \mathbf{I}_{d_r}, \Lambda \succeq \lambda \mathbf{I}_{d_p} \Big\} , \\
&\text{where } V_h(x; \theta, \beta, \Sigma, \Lambda, \gamma_r, \gamma_p) = \max_{a \in \mathcal{A}} \Big\{ \min \big\{ f_r(x, a; \theta, \Sigma, \gamma_r) + f_p(x, a; \beta, \Lambda, \gamma_p), H - h + 1 \big\}^+ \Big\} \\
&\text{with } f_r(x, a; \theta, \Sigma, \gamma_r) = g(\langle \phi_r(x,a), \theta \rangle) - \gamma_r \cdot \sqrt{\phi_r(x,a)^\intercal \Sigma^{-1} \phi_r(x,a)} \\
&\text{and } f_p(x, a; \beta, \Lambda, \gamma_p) = \langle \phi_p(x,a), \beta \rangle - \gamma_p \cdot \sqrt{\phi_p(x,a)^\intercal \Lambda^{-1} \phi_p(x,a)} .
\end{aligned}$$

For all $\epsilon > 0$, let $\mathcal{N}_h(\epsilon; R, B, J_r, J_p, \rho, \lambda)$ be the minimal ϵ-cover of $\mathcal{V}_h(R, B, J_r, J_p, \rho, \lambda)$ with respect to the supremum norm. In other words, for any function $V \in \mathcal{V}_h(R, B, J, \rho, \lambda)$, there exists a function $V^\dagger \in \mathcal{N}_h(\epsilon; R, B, J_r, J_p, \rho, \lambda)$ such that

$$\sup_{x \in \mathcal{S}} \big| V(x) - V^\dagger(x) \big| \le \epsilon .$$

Meanwhile, among all ϵ-covers of $\mathcal{V}_h(R, B, J_r, J_p, \rho, \lambda)$ defined by such a property, we choose $\mathcal{N}_h(\epsilon; R, B, J, \rho, \lambda)$ as the one with the minimal cardinality.

Recall the construction of $\widetilde{V}_h$ in Algorithm B.1. For sufficiently large n, by Lemma L.2, we have $\|\widetilde{\theta}_h\|_2 \le \|\theta_h^\star\|_2 + \|\widetilde{\theta}_h - \theta_h^\star\|_2 \le 2\|\theta_h^\star\|_2 := R_0$ with probability at least $1 - \xi/4$. By equation L.22, we have $\lambda_{\min}\big(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\big) \ge \rho/2$ with probability at least $1 - \xi/4$ where $\rho = \lambda_{\min}\big(\Sigma_h(\theta_h^\star)\big) > 0$. By Lemma L.4, we have $\|\widetilde{\beta}_h\|_2 \le H\sqrt{nd_p/\lambda} := B_0$. Finally, we take $J_r = 2\alpha_r$ and $J_p = 2\alpha_p$. Under these events, we have

$$\widetilde{V}_{h+1} \in \mathcal{V}_{h+1}(R_0, B_0, J_r, J_p, n\rho/2, \lambda).$$

Here $\lambda > 0$ is the regularization parameter and $\alpha_r, \alpha_p > 0$ are the scaling parameters, which are specified in Algorithm B.1. For notational simplicity, we use $\mathcal{V}_{h+1}$ and $\mathcal{N}_{h+1}(\epsilon)$ to denote $\mathcal{V}_{h+1}(R_0, B_0, J_r, J_p, n\rho/2, \lambda)$ and $\mathcal{N}_{h+1}(\epsilon; R_0, B_0, J_r, J_p, n\rho/2, \lambda)$, respectively. As a result, there exists functions $V_{h+1}^\dagger \in \mathcal{N}_{h+1}(\epsilon)$ such that

$$\sup_{x \in \mathcal{S}} |\widetilde{V}_{h+1}(x) - V_{h+1}^\dagger(x)| \le \epsilon.$$

Hence, given $\widetilde{V}_{h+1}$ and $V_{h+1}^\dagger$, we have

$$\mathbb{E}\big[|V_{h+1}^\dagger(x_{h+1}) - \widetilde{V}_{h+1}(x_{h+1})| \mid x_h = x, a_h = a\big] \le \epsilon, \quad \forall (x,a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H].$$

Here the conditional expectation is induced by the transition kernel $\mathbb{P}_h(\cdot \mid x, a)$. As a result, for all $h \in [H]$, we have

$$\big|\epsilon_h^\tau(\widetilde{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)\big| \le 2\epsilon, \forall \tau \in [n].$$

By the Cauchy-Schwarz inequality, for any two vectors $a, b \in \mathbb{R}^d$ and any positive definite matrix $\Lambda \in \mathbb{R}^{d\times d}$, it holds that $\|a+b\|_\Lambda^2 \le 2\|a\|_\Lambda^2 + 2\|b\|_\Lambda^2$. Hence, for all $h \in [H]$, we have

$$|(\mathrm{v})|^2 \le 2 \underbrace{\Big\| \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V_{h+1}^\dagger)\Big\|_{\Omega_h}}_{b(\{V_{h+1}^\dagger\})} + 2 \underbrace{\Big\| \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau)\big(\epsilon_h^\tau(\widetilde{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)\big)\Big\|_{\Omega_h}^2}_{(\mathrm{vi})}$$

Here (vi) can be bounded by

$$(\mathrm{vi}) \le 2\Big\|\Omega_h \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau)\phi_p(x_h^\tau, a_h^\tau)^\intercal\Big\|^2 \sum_{\tau=1}^n \big(\epsilon_h^\tau(\widetilde{V}_{h+1}) - \epsilon_h^\tau(V_{h+1}^\dagger)\big)^2 \le 8n\epsilon^2, \tag{G.7}$$

where we denote $\|A\|$ as the operator norm of a matrix A. We then bound $b(\{V_{h+1}^\dagger\})$ via uniform concentration inequalities. Applying Lemma L.7 and the union bound, for any fixed $h \in [H]$, we have

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}}\Big(\sup_{V \in \mathcal{N}_{h+1}(\epsilon)} \Big\| \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V)\Big\|_{\Omega_h} > H^2(2\log(1/\delta) + d_p\log(1+n/\lambda))\Big)\\ &\le \delta|\mathcal{N}_{h+1}(\epsilon)|.\end{aligned}$$

For all $\xi \in (0,1)$ and all $\epsilon > 0$, we set $\delta = \xi/(4H|\mathcal{N}_{h+1}(\epsilon)|)$. Hence, for any fixed $h \in [H]$, it holds that

$$\begin{aligned}&\sup_{V \in \mathcal{N}_{h+1}(\epsilon)} \Big\| \sum_{\tau=1}^n \phi_p(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V)\Big\|_{\Omega_h} \le H^2(2\log(1/\delta) + d_p\log(1+n/\lambda))\\ &\le H^2\Big(2\log\big(4H|\mathcal{N}_{h+1}(\epsilon)|/\xi\big) + d_p\log(1+n/\lambda)\Big)\end{aligned} \tag{G.8}$$

with probability at least $1 - \xi/(4H)$, which is taken with respect to $\mathbb{P}_{\mathcal{D}}$. Using the union bound again, we have equation G.8 holds for all $h \in [H]$ with probability at least $1 - \xi/4$. Combining equation G.7 and equation G.8, we have

$$|(\mathrm{v})|^2 \le H^2\Big(2\log\big(4H|\mathcal{N}_{h+1}(\epsilon)|/\xi\big) + d_p\log(1+n/\lambda)\Big) + 8n\epsilon^2.$$

Applying Lemma L.6 with $R_0 = 2\|\theta_h^\star\|_2$, $B_0 = H\sqrt{n/\lambda}$, $\epsilon = H\sqrt{d_p}/\sqrt{n}$, $\alpha_r = c_r\sqrt{d_r \log H/\xi}$, $\alpha_p = c_p(d_p + d_r)H\sqrt{\zeta}$, $\zeta = \log\big(2(d_r + d_p)Hn/\xi\big)$, and $\lambda = 1$, when $n, c_p > c_r$ are sufficiently large, we have

$$
\begin{aligned}
&\log\big|\mathcal{N}_h(\epsilon; R_0, B_0, J_r, J_p, n\rho/2, \lambda)\big| \\
&\le d_r \log(1 + 8LR_0/\epsilon) + d_r^2 \log\big(1 + 64{d_r}^{1/2} J_r^2/(n\rho\epsilon^2)\big) \\
&\quad + d_p \log(1 + 8B_0/\epsilon) + d_p^2 \log\big(1 + 32{d_p}^{1/2} J_p^2/(\lambda\epsilon^2)\big) \\
&= d_r \log(1 + 8LR_0 H^{-1}\sqrt{n}/\sqrt{d_p}) + d_r^2 \log\big(1 + 256c_r^2 {d_r}^{3/2}\log(H/\xi)/(d_p\rho H^2)\big) \\
&\quad + d_p \log(1 + 8n/\sqrt{d_p}) + d_p^2 \log\big(1 + 32c_p^2 n {d_p}^{1/2}(d_r + d_p)^2\zeta/d_p\big) \\
&\le 2d_r \log(1 + 8LR_0 H^{-1}\sqrt{n}/\sqrt{d_p}) + 2d_p^2 \log\big(1 + 32c_p^2 n(d_r + d_p)^2\zeta/d_p^{1/2}\big) \\
&\le 2d_r\zeta + 2d_p^2 \log\big(64c_p^2 n(d_r + d_p)^2\zeta/d_p^{1/2}\big) \\
&\le 2d_r\zeta + 2d_p^2(5 + 2\log c_p + 3\zeta) \\
&\le 2(d_r + d_p)^2(\log c_p + 5\zeta)
\end{aligned}
$$

which implies that

$$
\begin{aligned}
|(\mathrm{v})|^2 &\le H^2\Big(2\log\big(4H|\mathcal{N}_{h+1}(\epsilon)|/\xi\big) + d_p\log(1 + n/\lambda)\Big) + 8n\epsilon^2 \\
&\le 2H^2\Big(\log(4H/\xi) + 2(d_r + d_p)^2(\log c_p + 5\zeta) + d_p\log(1 + n)\Big) + 8H^2 d_p \\
&\le 2H^2\Big(\zeta + 2(d_r + d_p)^2(\log c_p + 5\zeta) + d_p\zeta + 4d_p\Big) \\
&\le 20(d_r + d_p)^2 H^2\zeta(1 + \log c_p) \le (d_r + d_p)^2 H^2 c_p^2\zeta/4
\end{aligned}
$$

when $c_p \ge 1$ sufficiently large. We then have $|(\mathrm{iv})| \le \alpha_p/2\|\phi_p(x,a)^\intercal\|_{\Omega_h}$. In addition, $H\sqrt{\lambda d_p} \le \alpha_p/2$, we have

$$
|(\mathrm{ii})| = \big|\langle\phi_p(x,a), \beta_h - \widetilde{\beta}_h\rangle\big| \le \alpha_p\|\phi_p(x,a)\|_{\Omega_h} = \widetilde{\Gamma}_{p,h}(x,a),
$$

which finishes the proof. □

H Proof of Corollaries 2 and E.1

Proof. Note that using the matrix Bernstein inequality(Tropp, 2015) , for n large enough, we have

$$
\lambda_{\min}(\frac{1}{n}\widetilde{\Lambda}_h) \ge \rho/2
$$

holds for any $h \in [H]$ with probability at least $1 - \xi/2$. Also using equation L.22, we have for n large enough,

$$
\lambda_{\min}\big(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\big) \ge \rho/2
$$

holds for any $h \in H$ with probability at least $1 - \xi/2$. Conditioning on the two events, the first claim of the corollary immediately follows from Theorem 1.
The second claim immediately comes from Lemma L.2. □

I Proof of Theorems 2 and E.2

Proof. Following the definition of $\mathrm{SubOpt}(\widehat{\pi}, x)$, we decompose the suboptimality gap as:

$$
\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}, x) &= V^*_{1,\theta^*}(x) - V^{\widehat{\pi}}_{1,\theta^*}(x) \\
&= V^*_{1,\theta^*}(x) - \widehat{V}_{1,\theta^*}(x) + \widehat{V}_{1,\theta^*}(x) - V^{\widehat{\pi}}_{1,\theta^*}(x) \\
&\le V^*_{1,\theta^*}(x) - \widehat{V}_{1,\theta^*}(x) \\
&= \underbrace{V^*_{1,\theta^*}(x) - V^*_{1,\widetilde{\theta}}(x)}_{(i)} + \underbrace{V^*_{1,\widetilde{\theta}}(x) - \widehat{V}_{1,\widetilde{\theta}}(x)}_{(ii)} + \underbrace{\widehat{V}_{1,\widetilde{\theta}}(x) - \widehat{V}_{1,\theta^*}(x)}_{(iii)}
\end{aligned}
$$

where the footnote θ^* and $\widetilde{\theta}$ mean that we use true and estimated θ in each time step. The first inequality follows from Lemma 1 of Jin et al. (2021). We analyze each term separately.

For term (i), we have:

$$\begin{aligned}(i) &= V^{*}_{1,\theta^*}(x) - V^{*}_{1,\widetilde{\theta}}(x) \\ &= \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[g\left(\langle \phi_r(x,a), \theta_h^* \rangle\right) - g\left(\left\langle \phi_r(x,a), \widetilde{\theta}_h \right\rangle\right) \mid x_1 = x\right] \\ &\leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\widetilde{\Gamma}_{r,h}(x,a) \mid x_1 = x\right]\end{aligned} \tag{I.9}$$

where the last inequality follows from Lemma L.3, since $\widetilde{\Gamma}_{r,h}(x,a)$ bounds the difference in the reward function approximation.

For term (ii), we have:

$$\begin{aligned}(ii) &= V^{*}_{1,\widetilde{\theta}}(x) - \widehat{V}_{1,\widetilde{\theta}}(x) \\ &= \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left(\mathbb{B}_h \widehat{V}_{h+1}\right)(x_h,a_h) - \widehat{Q}_h(x_h,a_h) \mid x_1 = x\right] \\ &- \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left\langle \widehat{Q}_h(x_h,\cdot), \widehat{\pi}_h(\cdot \mid x_h) - \pi_h^*(\cdot \mid x_h)\right\rangle_{\mathcal{A}} \mid x_1 = x\right] \\ &\leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\left(\mathbb{B}_h \widehat{V}_{h+1}\right)(x_h,a_h) - \widehat{Q}_h(x_h,a_h) \mid x_1 = x\right] \\ &= \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\iota_h(x_h,a_h) \mid x_1 = x\right] \\ &\leq 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\widehat{\Gamma}_h(x_h,a_h) \mid x_1 = x\right]\end{aligned} \tag{I.10}$$

where $\iota_h(x,a) = (\mathbb{B}_h \widehat{V}_{h+1})(x,a) - \widehat{Q}_h(x,a)$ and the inequalities follow from a similar analysis as in Theorem 1.

For term (iii), we have:

$$\begin{aligned}(iii) &= \widehat{V}_{1,\widetilde{\theta}}(x) - \widehat{V}_{1,\theta^*}(x) \\ &= \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[g\left(\phi_r(x,a)^\top \widetilde{\theta}_h\right) + \phi_p(x,a)^\top \widehat{\beta}_h - \widehat{\Gamma}_h(x,a,\widetilde{\theta}_h) \mid x_1 = x\right] \\ &- \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[g\left(\phi_r(x,a)^\top \theta_h^*\right) + \phi_p(x,a)^\top \widehat{\beta}_h - \widehat{\Gamma}_h(x,a,\theta_h^*) \mid x_1 = x\right] \\ &= \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[g\left(\phi_r(x,a)^\top \widetilde{\theta}_h\right) - g\left(\phi_r(x,a)^\top \theta_h^*\right) \mid x_1 = x\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\widehat{\Gamma}_h(x,a,\theta_h^*) - \widehat{\Gamma}_h(x,a,\widetilde{\theta}) \mid x_1 = x\right]\end{aligned}$$

By Lemma L.3, we have

$$\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}} \left[g\left(\phi_r(x,a)^\top \widetilde{\theta}_h\right) - g\left(\phi_r(x,a)^\top \theta_h^*\right) \mid x_1 = x \right] \le \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}} \left[\widehat{\Gamma}_{r,h}(x,a) \mid x_1 = x \right]$$

Now, let's analyze the difference in the uncertainty quantifiers:

$$\begin{aligned}
&\widehat{\Gamma}_h(x,a,\theta_h^*) - \widehat{\Gamma}_h(x,a,\widetilde{\theta}_h) \\
&= \alpha_r \mathbb{E}_{\widehat{\pi}} \left[\dot{g}\left(\langle \phi_r(x,a), \theta_h^* \rangle\right) \|\phi_r(x,a)\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}} - \dot{g}\left(\left\langle \phi_r(x,a), \widetilde{\theta}_h \right\rangle\right) \|\phi_r(x,a)\|_{\widehat{\Sigma}_h(\widetilde{\theta}_h)^{-1}} \mid x_1 = x \right]
\end{aligned}$$

Define Δ as:

$$\Delta \coloneqq \alpha_r \left(\dot{g}\left(\langle \phi_r(x,a), \theta_h^* \rangle\right) \|\phi_r(x,a)\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}} - \dot{g}\left(\left\langle \phi_r(x,a), \widetilde{\theta}_h \right\rangle\right) \|\phi_r(x,a)\|_{\widehat{\Sigma}_h(\widetilde{\theta}_h)^{-1}} \right)$$

We decompose Δ into:

$$\begin{aligned}
\Delta = &\underbrace{\alpha_r \left[\dot{g}\left(\langle \phi_r, \theta_h^* \rangle\right) - \dot{g}\left(\left\langle \phi_r, \widetilde{\theta}_h \right\rangle\right) \right] \|\phi_r\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}}}_{\Delta_1} \\
&+ \underbrace{\alpha_r \, \dot{g}\left(\left\langle \phi_r, \widetilde{\theta}_h \right\rangle\right) \left[\|\phi_r\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}} - \|\phi_r\|_{\widehat{\Sigma}_h(\widetilde{\theta}_h)^{-1}} \right]}_{\Delta_2}
\end{aligned}$$

Under Assumption 2, we have:

$$\|\phi_r\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}} \le \frac{1}{\sqrt{\lambda_{\min}\left(\widehat{\Sigma}_h\left(\theta_h^*\right)\right)}} \cdot \|\phi_r\|_2 \le \frac{1}{\sqrt{n\rho/2}} \cdot 1 = \sqrt{\frac{2}{n\rho}}$$

By Assumption 1 and Lemma L.2, we bound Δ_1:

$$\begin{aligned}
\Delta_1 &\le \alpha_r L \left| \dot{g}\left(\langle \phi_r, \theta_h^* \rangle\right) - \dot{g}\left(\left\langle \phi_r, \widetilde{\theta}_h \right\rangle\right) \right| \|\phi_r\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}} \\
&\le \alpha_r L \left| \left\langle \phi_r, \theta_h^* - \widetilde{\theta}_h \right\rangle \right| \|\phi_r\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}} \\
&\le \alpha_r L \|\phi_r\|_2 \|\widetilde{\theta}_h - \theta_h^*\|_2 \|\phi_r\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}} \\
&\le \alpha_r L \sqrt{\frac{2}{n\rho}} \sqrt{\frac{c d_r \log 1/\xi}{n}} \\
&= L \sqrt{\frac{2c(\log H/\xi)(\log 1/\xi)}{\rho}} \frac{d_r}{n}
\end{aligned} \tag{I.11}$$

For Δ_2, let $L = \sup_x \dot{g}(x)$, then:

$$\begin{aligned}
\Delta_2 &\le \alpha_r L \left| \|\phi_r\|_{\widehat{\Sigma}_h(\theta_h^*)^{-1}} - \|\phi_r\|_{\widehat{\Sigma}_h(\widetilde{\theta}_h)^{-1}} \right| \\
&\le \alpha_r L \sqrt{\left| \phi_r^\top \left(\widehat{\Sigma}_h\left(\theta_h^*\right)^{-1} - \widehat{\Sigma}_h\left(\widetilde{\theta}_h\right)^{-1} \right) \phi_r \right|} \\
&\le \alpha_r L \sqrt{\|\phi_r\|_2^2 \cdot \|\widehat{\Sigma}_h\left(\theta_h^*\right)^{-1} - \widehat{\Sigma}_h\left(\widetilde{\theta}_h\right)^{-1}\|}
\end{aligned}$$

Using the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, we get:

$$\begin{aligned}\|\widehat{\Sigma}_h\left(\theta_h^*\right)^{-1} - \widehat{\Sigma}_h\left(\widetilde{\theta}_h\right)^{-1}\| &\leq \|\widehat{\Sigma}_h\left(\theta_h^*\right)^{-1}\| \cdot \|\widehat{\Sigma}_h\left(\theta_h^*\right) - \widehat{\Sigma}_h\left(\widetilde{\theta}_h\right)\| \cdot \|\widehat{\Sigma}_h\left(\widetilde{\theta}_h\right)^{-1}\| \\ &\leq \frac{1}{n}\frac{2}{\rho} \cdot L\|\widetilde{\theta}_h - \theta_h^*\|_2 \cdot \frac{2}{\rho} \\ &\leq \frac{4L}{n\rho^2}\sqrt{\frac{cd_r \log 1/\xi}{n}}\end{aligned}$$

Therefore:

$$\begin{aligned}\Delta_2 &\leq \alpha_r L\sqrt{\frac{4L}{n\rho^2}\sqrt{\frac{cd_r \log 1/\xi}{n}}} \\ &= L\sqrt{\frac{d_r 4L\log(H/\xi)}{n\rho^2}} \cdot \left(\frac{cd_r \log 1/\xi}{n}\right)^{1/4}\end{aligned} \tag{I.12}$$

Combining I.11 and I.12, and define Δ_{err}:

$$\Delta_{err} \coloneqq L\sqrt{\frac{2c(\log H/\xi)(\log 1/\xi)}{\rho}}\frac{d_r}{n} + G\sqrt{\frac{d_r 4L\log(H/\xi)}{n\rho^2}} \cdot \left(\frac{cd_r \log 1/\xi}{n}\right)^{1/4} \tag{I.13}$$

We could get:

$$\begin{aligned}\Delta &\leq \Delta_{err} \\ &= L\sqrt{\frac{2c(\log H/\xi)(\log 1/\xi)}{\rho}}\frac{d_r}{n} + G\sqrt{\frac{d_r 4L\log(H/\xi)}{n\rho^2}} \cdot \left(\frac{cd_r \log 1/\xi}{n}\right)^{1/4} \\ &= \widetilde{O}\left(\frac{d_r}{n}\right) + \widetilde{O}\left(\frac{d_r^{3/4}}{n^{3/4}}\right) \\ &= \widetilde{O}\left(\frac{d_r^{\frac{3}{4}}}{n^{\frac{3}{4}}}\right)\end{aligned}$$

for sufficiently large n.

Combining I.9, I.10, and I.13, we have:

$$\mathrm{SubOpt}(\widehat{\pi}; x) \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\widetilde{\Gamma}_{r,h}(x_h, a_h) + 2\widehat{\Gamma}_h(x_h, a_h) \mid x_1 = x\right] + \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\Delta_{err} \mid x_1 = x\right]$$

□

J Proof of Corollaries 3 and E.2

Proof. From Theorem 2, we get:

$$\begin{aligned}\Delta &\leq \Delta_{err} \\ &= L\sqrt{\frac{2c(\log H/\xi)(\log 1/\xi)}{\rho}}\frac{d_r}{n} + G\sqrt{\frac{d_r 4Llog(H/\xi)}{n\rho^2}} \cdot \left(\frac{cd_r \log 1/\xi}{n}\right)^{1/4} \\ &= \widetilde{O}\left(\frac{d_r^{\frac{3}{4}}}{n^{\frac{3}{4}}})\right)\end{aligned} \tag{J.14}$$

for sufficiently large n.

Now, for the semi-supervised estimator $\widehat{\beta}_h$, we benefit from the additional unlabeled data. By using a similar analysis as in the proof of Theorem 1, and if $\Lambda_h \geq \rho$, we can show that:

$$\begin{aligned}\widehat{\Gamma}_{p,h}(x,a) &= \alpha_p \sqrt{\phi_p(x,a)^\top (\widehat{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1} \phi_p(x,a)} \\ &= O\left(\sqrt{\frac{(d_p+d_r)^2 H^2 \log\left(2\left(d_r+d_p\right)H(n+N)/\xi\right)}{n+N}}\right)\end{aligned} \tag{J.15}$$

From Lemma L.3, we could also get similar results that:

$$\begin{aligned}\widetilde{\Gamma}_{r,h}(x,a) &= c_0 \sqrt{d_r \log H/\xi} \times \sqrt{\dot{g}(\langle \phi(x,a), \widetilde{\theta}_h \rangle)^2 \phi_r(x,a)^\intercal \widehat{\Sigma}_h (\widetilde{\theta}_h)^{-1} \phi_r(x,a)} \\ &= O\left(\sqrt{\frac{d_r \log(H/\xi)}{n}}\right)\end{aligned} \tag{J.16}$$

Therefore, the increased sample size from n to $n+N$ leads to a reduction in the uncertainty quantifier related to the transition dynamics.

Combining J.14, J.15, and J.16, when n is large enough, we have:

$$\begin{aligned}\mathrm{SubOpt}(\widehat{\pi}; x) &\leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\widetilde{\Gamma}_{r,h}(x_h,a_h) + 2\widehat{\Gamma}_h(x_h,a_h) \mid x_1 = x\right] + \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\Delta_{err} \mid x_1 = x\right] \\ &\leq O\left(\sqrt{\frac{d_r H^2 \log H/\xi}{n}}\right) + O\left(\sqrt{\frac{(d_p+d_r)^2 H^4 \log\left(2\left(d_r+d_p\right)H(n+N)/\xi\right)}{n+N}}\right) \\ &+ \widetilde{O}\left(\frac{d_r^{\frac{3}{4}})}{n^{\frac{3}{4}}}\right) \\ &= O\left(\sqrt{\frac{d_r H^2 \log(H/\xi)}{n}}\right) \\ &+ O\left(\sqrt{\frac{(d_p+d_r)^2 H^4 \log\left(2\left(d_r+d_p\right)H(n+N)/\xi\right)}{n+N}}\right)\end{aligned} \tag{J.17}$$

This result shows that the semi-supervised approach benefits from unlabeled data in improving the estimation of transition dynamics while the reward estimation is limited by the size of the labeled dataset n. When $N \gg n$, this approach can significantly reduce the overall suboptimality compared to using only labeled data.

□

K Proof of Theorem 3

Proof. By Lemmas L.10, L.11 and L.12, taking $\mathrm{conf}_{h,t} = \min\{H-h+1, \gamma_r \|\phi_r\|_{\Lambda'^{-1}_{h,t}} + \gamma_p \|\phi_p\|_{\Lambda^{-1}_{h,t}}\}$ in Lemma 7 in Wang et al. (2019), and using Lemma 6 in Wang et al. (2019), Lemma L.10 and Lemma L.11, we have

$$\mathcal{R}(x) \leq H\sqrt{T}\Big(\gamma_r \sqrt{2d_r \ln(1+T/d_r)} + \gamma_p \sqrt{2d_r \ln(1+T/d_p)}\Big) + \sum_{t=1}^{T} \zeta_t$$

holds with probability at least $1 - 2p_0/3$, where $\zeta_t = \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}_{\cdot,t}} \mathrm{conf}_{h,t-1}(x_h, \widehat{\pi}_{h,t}(x_h)) - \mathrm{conf}_{h,t-1}(x_{h,t}, a_{h,t})$. Hence using the Azuma inequality we have

$$\mathcal{R}(x) \le H\sqrt{T}\Big(\gamma_r\sqrt{2d_r\ln(1+T/d_r)} + \gamma_p\sqrt{2d_r\ln(1+T/d_p)}\Big) + \sqrt{2\ln(6/p_0)TH^3}$$

holds with probability at least $1 - p_0$. □

L Technical Lemmas

Lemma L.1. *If $\lambda_{\min}\big(\Sigma_h(\theta_h^\tau)\big) \ge \rho > 0$, then for $\xi \in (0,1)$, with sufficiently large n, we have*

$$\lambda_{\min}\big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big) \ge \frac{3\rho}{4}$$

with probability at least $1 - \xi$.

Proof. By the matrix Bernstein inequality (Tropp, 2015) and $\|\phi_r(x,a)\|_2 \le 1$, we have

$$\|\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star) - \Sigma_h(\theta_h^\star)\| \le C\sqrt{\log(d_r/\xi)/n} \tag{L.18}$$

with probability at least $1 - \xi/2$ for an absolute constant $C > 0$. Hence if n is sufficiently large, we have

$$\lambda_{\min}\big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big) \ge \lambda_{\min}\big(\Sigma_h(\theta_h^\star)\big) - \|\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star) - \Sigma_h(\theta_h^\star)\| \ge \rho - \frac{\rho}{4} = \frac{3\rho}{4}.$$

□

Lemma L.2. *Suppose that $\lambda_{\min}\big(\Sigma_h(\theta_h^\star)\big) \ge \rho > 0$. Under Assumption 1, for $\xi \in (0,1)$, with sufficiently large n and $h \in [H]$ fixed, we have*

$$\|\widetilde{\theta}_h - \theta_h^\star\|^2_{\frac{1}{n}\widehat{\Sigma}_h(\theta_h^\star)} \le \frac{cd_r\log 1/\xi}{n\rho},\ \|\widetilde{\theta}_h - \theta_h^\star\|_2^2 \le \frac{cd_r\log 1/\xi}{n\rho^2} \text{ and } \|\nabla\mathcal{L}_h(\theta_h^\star)\|^2_{\Sigma_h^{-1}(\theta_h^\star)} \le \frac{cd_r\log 1/\xi}{n\rho}$$

with probability at least $1 - \xi$ for some absolute constant $c > 0$.

Proof. Apply Theorem 2.1 in Hsu et al. (2012) to $A = \Sigma_h^{-1/2}(\theta_h^*)$ and $x = \sqrt{n}\nabla\mathcal{L}_h(\theta_h^*)$. Since

$$\nabla l_h(\theta_h^*) = \big(r_h - g(\langle\phi_h, \theta_h^*\rangle)\big)\phi_h$$

is subguassian since we assume $r_h - g(\langle\phi_h, \theta^*\rangle)$ is subguassian, there exists $\sigma > 0$, independent with n, such that for any $t > 0$,

$$\mathbb{P}\Big(\|x\|^2_{\Sigma_h(\theta_h^\star)^{-1}} - \sigma\big(\mathrm{Tr}(\Sigma_h(\theta_h^\star)^{-1}) + 2\|\Sigma_h(\theta_h^\star)^{-1}\|t + 2\sqrt{\mathrm{Tr}(\Sigma_h(\theta_h^\star)^{-2})t}\,\big) > 0\Big) \le e^{-t}.$$

Hence

$$\|\nabla\mathcal{L}_h(\theta_h^*)\|^2_{\Sigma_h(\theta_h^\star)^{-1}} \le \frac{3\sigma d_r\|\Sigma_h(\theta_h^\star)^{-1}\|\log(1/\xi)}{n}$$

holds with probability at least $1 - \xi$, for any small $\xi > 0$.

Similarly, for any $1 \le \tau \le n$, applying Theorem 2.1 in Hsu et al. (2012) to $A = \Sigma_h(\theta_h^\star)^{-1/2}$ and $x = \phi_h^\tau$, there exists $\sigma' > 0$ independent with n, such that

$$\max_{1\le\tau\le n}\|\phi_h^\tau\|^2_{\Sigma_h^{-1}} \le 3\sigma' d_r\|\Sigma_h^{-1}\|\log(n/\xi).$$

holds with probability at least $1 - \xi$, for any (small) $\xi > 0$. Using Theorem A.2 in Ostrovskii and Bach (2021), we have

$$\max_{1\le\tau\le n}\|\phi_h^\tau\|_{\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)^{-1}}\|\nabla\mathcal{L}_h(\theta_h^*)\|^2_{\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)^{-1}} \le \frac{1}{4}$$

holds with probability at least $1 - 3\xi$ if n is sufficiently large.

It's easy to check $\mathcal{L}_h$ falls into the case (a) of Proposition B.3 in Ostrovskii and Bach (2021) with $\theta_0 = \theta_h^*$, $H_0 = \widetilde{\Sigma}_h$, $W(\theta) = \phi^{\tau(\theta)}$ where $\tau(\theta) \coloneqq \operatorname{argmin}_{1\le\tau\le n} |\langle \phi^\tau, \theta - \theta_h^*\rangle|$. Then using Proposition B.4 in Ostrovskii and Bach (2021), we have

$$\|\widetilde{\theta}_h - \theta_h^*\|^2_{\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^*)} \le 4\|\nabla\mathcal{L}_h(\theta_h^*)\|^2_{\frac{1}{n}\widetilde{\Sigma}_h^{-1}(\theta_h^*)} \le \frac{12\sigma d_r \log(1/\xi)}{\rho n} \tag{L.19}$$

holds with probability at least $1-3\xi$ if n is large enough, for some constant $K > 0$, any small $\xi > 0$ and $\epsilon > 0$. The bound for $\|\theta_h^\star - \widetilde{\theta}_h\|_2$ immediately comes from equation L.19 and Lemma L.1. □

Lemma L.3. *Suppose that $\lambda_{\min}\big(\Sigma_h(\theta_h^\star)\big) \ge \rho > 0$. Under Assumptions 1, for $\xi \in (0,1)$, with sufficiently large n, we have,*

$$\begin{aligned}\big|g(\langle\phi_r(x,a),\widetilde{\theta}_h\rangle) - g(\langle\phi_r(x,a),\theta_h^\star\rangle)\big| \le& c_0\sqrt{d_r\log H/\xi} \\ &\times\sqrt{\dot g(\langle\phi(x,a),\widetilde{\theta}_h\rangle)^2\phi_r(x,a)^\intercal\widetilde{\Sigma}_h(\widetilde{\theta}_h)^{-1}\phi_r(x,a)}\,.\end{aligned}$$

for all $(x,a)\in\mathcal{S}\times\mathcal{A}$ and $h\in[H]$ with probability at least $1-\xi$ for some absolute constant $c_0>0$.

Proof. By Taylor's theorem,

$$\mathbf{0} = \nabla\mathcal{L}_h(\widetilde{\theta}_h) = \nabla\mathcal{L}_h(\theta_h^\star) - \frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)(\widetilde{\theta}_h - \theta_h^\star) + o(\|\widetilde{\theta}_h - \theta_h^*\|_2)$$

Hence, we have

$$\widetilde{\theta}_h - \theta_h^\star = \big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big)^{-1}\big(-\nabla\mathcal{L}_h(\theta_h^\star) + o(\|\widetilde{\theta}_h - \theta_h^*\|_2)\big)\,. \tag{L.20}$$

By Taylor's theorem again, we have

$$\begin{aligned}g(\langle\phi_r(x,a),\widetilde{\theta}_h\rangle) - g(\langle\phi_r(x,a),\theta_h^\star\rangle) =&\dot g(\langle\phi_r(x,a),\widetilde{\theta}_h\rangle)\langle\phi_r(x,a),\widetilde{\theta}_h - \theta_h^\star\rangle \\ &+\frac{1}{2}\ddot g(\langle\phi_r(x,a),\breve{\theta}_h\rangle)\langle\phi_r(x,a),\widetilde{\theta}_h - \theta_h^\star\rangle^2 \\ :=&e_1 + e_2\end{aligned}$$

for some $\breve{\theta}_h$ on the line segment between $\theta_h^\star$ and $\widetilde{\theta}_h$. We then bound $|e_1|$ and $|e_2|$ separately. First, by equation L.20, we have

$$e_1 = \big\langle\phi_r(x,a), \big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big)^{-1}\big(-\nabla\mathcal{L}_h(\theta_h^\star) + o(\|\widetilde{\theta}_h - \theta_h^*\|_2)\big)\big\rangle\,.$$

By the matrix Bernstein inequality (Tropp, 2015), we have

$$\|\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star) - \Sigma_h(\theta_h^\star)\| \le C\sqrt{\log(Hd_r/\xi)/n} \tag{L.21}$$

with probability at least $1-\xi/2$ for an absolute constant $C>0$. As a result,

$$\lambda_{\min}\big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big) \ge \lambda_{\min}\big(\Sigma_h(\theta_h^\star)\big) - \|\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star) - \Sigma_h(\theta_h^\star)\| \ge \rho - \frac{\rho}{4} = \frac{3\rho}{4}$$

when n is sufficiently large. Besides, we have

$$\begin{aligned}\|\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star) - \frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\| &\le \frac{L}{n}\sum_{\tau=1}^n |\langle\phi_r(x_h^\tau,a_h^\tau),\theta_h^\star - \widetilde{\theta}_h\rangle|\|\phi_r(x_h^\tau,a_h^\tau)\|_2^2 \\ &\le L\|\theta_h^\star - \widetilde{\theta}_h\|_2\,.\end{aligned}$$

For n sufficiently large and any $h\in[H]$, by Lemma L.2, we have

$$\|\widetilde{\theta}_h - \theta_h^\star\|_2^2 \le \frac{cd_r\log H/\xi}{\rho^2 n} \le \frac{\rho}{4L}$$

with probability at least $1-\xi$, which implies that

$$\lambda_{\min}\big(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\big) \ge \lambda_{\min}\big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big) - \|\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h) - \frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\| \ge \frac{3\rho}{4} - \frac{\rho}{4} = \frac{\rho}{2}. \tag{L.22}$$

Note that we have

$$
\begin{aligned}
&\big|\langle \phi_r(x,a), \big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big)^{-1}\big(-\nabla\mathcal{L}_h(\theta_h^\star) + o(\|\widetilde{\theta}_h - \theta_h^*\|_2)\rangle\big| \\
&\le \big|\langle \phi_r(x,a), \big(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\big)^{-1}\big(-\nabla\mathcal{L}_h(\theta_h^\star) + o(\|\widetilde{\theta}_h - \theta_h^*\|_2)\rangle\big| \\
&+ \big|\langle \phi_r(x,a), \Big(\big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big)^{-1} - \big(\frac{1}{n}\widetilde{\Sigma}_h(\theta_h^\star)\big)^{-1}\Big)\big(-\nabla\mathcal{L}_h(\theta_h^\star) + o(\|\widetilde{\theta}_h - \theta_h^*\|_2)\rangle\big| \\
&\le \|\phi_r(x,a)\|_{\left(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\right)^{-1}} \Big\|\nabla\mathcal{L}_h(\theta_h^\star)\Big\|_{\left(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\right)^{-1}} + \frac{2}{\rho}\|\phi_r(x,a)\|_{\left(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\right)^{-1}}\|\widetilde{\theta}_h - \theta_h^\star\|_2 \\
&+ \frac{4L}{\rho^2}\|\phi_r(x,a)\|_{\left(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\right)^{-1}}\|\widetilde{\theta}_h - \theta_h^\star\|_2^2
\end{aligned}
$$

by the Cauchy-Schwarz inequality. By equation L.22 and Lemma L.2, we have

$$
\Big\|\nabla\mathcal{L}_h(\theta_h^\star)\Big\|_{\left(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\right)^{-1}} \le \frac{2}{\rho}\|\nabla\mathcal{L}_h(\theta_h^\star)\Big\|_2 \le \frac{2}{\rho}\sqrt{\frac{cd_r\log H/\xi}{\rho n}}\,.
$$

Thus, we get

$$
|e_1| \le \frac{6}{\rho}\sqrt{\frac{d_r\log H/\xi}{n}}\|\phi_r(x,a)\|_{\widetilde{\Sigma}_h(\widetilde{\theta}_h)^{-1}}\,.
$$

Finally,

$$
\begin{aligned}
|e_2| &= \big|\frac{1}{2}\ddot{g}(\langle\phi_r(x,a),\breve{\theta}_h\rangle)\langle\phi_r(x,a),\widetilde{\theta}_h - \theta_h^\star\rangle^2\big| \\
&\le \frac{L}{2}\|\phi_r(x,a)\|^2_{\left(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\right)^{-1}}\|\theta_h^\star - \widetilde{\theta}_h\|^2_{\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)} \\
&\le \frac{c'Ld_r\log H/\xi}{2n\rho}\|\phi_r(x,a)\|^2_{\left(\frac{1}{n}\widetilde{\Sigma}_h(\widetilde{\theta}_h)\right)^{-1}} \\
&\le c_0\sqrt{d_r\log H/\xi}\dot{g}(\langle\phi(x,a),\widetilde{\theta}_h\rangle)\|\phi_r(x,a)\|_{\widetilde{\Sigma}_h(\widetilde{\theta}_h)^{-1}}
\end{aligned}
$$

for some constant $c' > 0$ for sufficiently large n, where the first inequality comes from the Cauchy-Schwarz inequality, Assumption 1, the second inequality comes from Lemma L.2, and the last inequality comes from the fact that with high probability, $\|\widetilde{\theta}_h\|_2$ lies in a fixed compact interval, say $[0, D]$ and $\dot{g}_{\min} := \inf_{y\in[0,D]}\dot{g}(y) > 0$ since if $\dot{g}_{\min} = 0$, there exists $x \in R$ such that $\dot{g}(x) = 0$ otherwise g must be a constant function since $\dot{g}$ is continuous and $|\ddot{g}| \le \dot{g}$.

Combining these derivations and using the union bounds, we finish the proof. □

Lemma L.4 (Bounded Coefficients). *For any functions $V : \mathcal{S} \to [0, V_{\max}]$ where $V_{\max} > 0$ is an absolute constant, the vector $\beta_h = \int_{x'\in\mathcal{S}}\mu_h(x')V(x')\mathrm{d}x'$ satisfies $\|\beta_h\|_2 \le V_{\max}\sqrt{d_p}$. Besides, we have*

$$
\|\widetilde{\beta}_h\|_2 \le H\sqrt{nd_p/\lambda}\,.
$$

Proof. First, we have

$$
\|\beta_h\| = \Big\|\int_{x'\in\mathcal{S}}\mu_h(x')V(x')\mathrm{d}x'\Big\| \le V_{\max}\|\mu_h^j(\mathcal{S})\| \le V_{\max}\sqrt{d_p}.
$$

Besides, note that $|\widetilde{V}_{h+1}(x_{h+1}^\tau)| \le H$, we have

$$\begin{aligned}
\|\widetilde{\beta}_h\|_2 &= \|\sum_{\tau=1}^{n}(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1}\phi_p(x_h^\tau, a_h^\tau)\widetilde{V}_{h+1}(x_{h+1}^\tau)\|_2 \\
&\le H\sum_{\tau=1}^{n}\|\sqrt{\phi(x_h^\tau, a_h^\tau)(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1/2}(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1}(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1/2}\phi(x_h^\tau, a_h^\tau)}. \\
&\le \frac{H}{\sqrt{\lambda}}\sum_{\tau=1}^{n}\sqrt{\phi(x_h^\tau, a_h^\tau)^\top(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1}\phi(x_h^\tau, a_h^\tau)} \\
&\le H\sqrt{\frac{n}{\lambda}}\sqrt{\sum_{\tau=1}^{n}\phi(x_h^\tau, a_h^\tau)^\top\widetilde{\Lambda}_h^{-1}\phi(x_h^\tau, a_h^\tau)} \\
&\le H\sqrt{\frac{n}{\lambda}}\sqrt{\mathrm{Tr}[(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1}\sum_{\tau=1}^{n}\phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top]} \\
&= H\sqrt{\frac{n}{\lambda}}\sqrt{\mathrm{Tr}[(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p})^{-1}\widetilde{\Lambda}_h]} \\
&\le H\sqrt{\frac{nd_p}{\lambda}}
\end{aligned}$$

□

Lemma L.5 (Covering Number of Euclidean Ball)**.** *For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.*

Proof. See Lemma 5.2 in Vershynin (2010). □

Lemma L.6. *Let $\mathcal{V}_h(R, B, J_r, J_p, \rho, \lambda)$ denote a class of functions from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form*

$$\begin{aligned}
V_h(x; \theta, \beta, \Sigma, \Lambda, \gamma_r, \gamma_p) = \max_{a\in\mathcal{A}}\Big\{ &\min\big\{f_r(x, a; \theta, \Sigma, \gamma_r) + f_p(x, a; \beta, \Lambda, \gamma_p), H - h + 1\big\}^+\Big\} \\
&\textit{with } f_r(x, a; \theta, \Sigma, \gamma_r) = g(\langle\phi_r(x, a), \theta\rangle) - \gamma_r \cdot \sqrt{\phi_r(x, a)^\top\Sigma^{-1}\phi_r(x, a)} \\
&\textit{and } f_p(x, a; \beta, \Lambda, \gamma_p) = \langle\phi_p(x, a), \beta\rangle - \gamma_p \cdot \sqrt{\phi_p(x, a)^\top\Lambda^{-1}\phi_p(x, a)}\,,
\end{aligned}$$

where the parameters $(\theta, \beta, \Sigma, \Lambda, \gamma_r, \gamma_p)$ satisfy $\|\theta\|_2 \le R, \|\beta\|_2 \le B, \gamma_r, \gamma_p \in [0, J], \Sigma \succeq \rho\mathbf{I}_{d_r}, \Lambda \succeq \lambda\mathbf{I}_{d_p}$. Suppose that the first-order derivative $\dot{g}(\cdot)$ of the link function $g(\cdot)$ is bounded by $L > 0$. Assume that $\max\{\|\phi_r(x, a)\|_2, \|\phi_p(x, a)\|_2\} \le 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, and let $\mathcal{N}_h(\epsilon; R, B, J, \rho, \lambda)$ be the ϵ-covering number of $\mathcal{V}_h(R, B, J, \rho, \lambda)$ with respect to the distance $\mathrm{dist}(V, V') = \sup_{x\in\mathcal{S}}|V(x) - V'(x)|$. Then

$$\begin{aligned}
\log\mathcal{N}_h(\epsilon; R, B, L, \rho, \lambda) \le\ & d_r\log(1 + 8LR/\epsilon) + d_p\log(1 + 8B/\epsilon) + d_r^2\log\big(1 + 32{d_r}^{1/2}J^2/(\rho\epsilon^2)\big) \\
&+ d_p^2\log\big(1 + 32{d_p}^{1/2}J^2/(\lambda\epsilon^2)\big).
\end{aligned}$$

Proof. Equivalently, we can reparametrize the function class $\mathcal{V}_h(R, B, J, \rho, \lambda)$ by setting $M_r = \gamma_r^2\Sigma^{-1}$ and $M_p = \gamma_p^2\Lambda^{-1}$, so we have

$$f_r(x, a; \theta, \Sigma, \gamma_r) = f_r(x, a; \theta, M_r) = g(\langle\phi_r(x, a), \theta\rangle) - \sqrt{\phi_r(x, a)^\top M_r\phi_r(x, a)}$$

and

$$f_p(x, a; \beta, \Lambda, \gamma_p) = f_p(x, a; \beta, M_p) = \langle\phi_p(x, a), \beta\rangle - \sqrt{\phi_p(x, a)^\top M_p\phi_p(x, a)}$$

for $\|M_r\| \le J^2\rho^{-1}$ and $\|M_p\| \le J^2\lambda^{-1}$. Then for any two function V and $V' \in \mathcal{V}_h(R, B, L, \rho, \lambda)$ with parameters $(\theta, \beta, M_r, M_p)$ and $(\theta', \beta', M_r', M_p')$, we have

$$\begin{aligned}
\operatorname{dist}(V, V') \le& \sup_{x,a} \left| f_r(x,a;\theta,M_r) + f_p(x,a;\beta,M_p) - f_r(x,a;\theta',M_r') - f_p(x,a;\beta',M_p') \right| \\
\le& \sup_{x,a} \left| g(\langle \phi_r(x,a),\theta\rangle) - g(\langle \phi_r(x,a),\theta'\rangle) \right| + \sup_{x,a} \left| \langle \phi_p(x,a),\beta\rangle - \langle \phi_p(x,a),\beta'\rangle \right| \\
&+ \sup_{x,a} \left| \sqrt{\phi_r(x,a)^\top M_r \phi_r(x,a)} - \sqrt{\phi_r(x,a)^\top M_r' \phi_r(x,a)} \right| \\
&+ \sup_{x,a} \left| \sqrt{\phi_p(x,a)^\top M_p \phi_p(x,a)} - \sqrt{\phi_p(x,a)^\top M_p' \phi_p(x,a)} \right| \\
\le& \sup_{\phi_r:\|\phi_r\|_2\le 1} L_1 \left| \langle \phi_r, \theta - \theta'\rangle) \right| + \sup_{\phi_p:\|\phi_p\|_2\le 1} \left| \langle \phi_p, \beta - \beta'\rangle) \right| \\
&+ \sup_{\phi_r:\|\phi_r\|_2\le 1} \sqrt{|\phi_r^\top (M_r - M_r')\phi_r|} + \sup_{\phi_p:\|\phi_p\|_2\le 1} \sqrt{|\phi_p^\top (M_p - M_p')\phi_p|} \\
=& L\|\theta - \theta'\|_2 + \|\beta - \beta'\|_2 + \sqrt{\|M_r - M_r'\|} + \sqrt{\|M_p - M_p'\|} \\
\le& L\|\theta - \theta'\|_2 + \|\beta - \beta'\|_2 + \sqrt{\|M_r - M_r'\|_{\mathrm{F}}} + \sqrt{\|M_p - M_p'\|_{\mathrm{F}}},
\end{aligned}$$

where the third inequality follows from the fact that $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$ holds for any $x, y \ge 0$.

Let $\mathcal{C}_\theta$ be an $\epsilon/4L$-cover of $\{\theta \in \mathbb{R}^{d_r} \mid \|\theta\| \le R\}$, $\mathcal{C}_\beta$ be an $\epsilon/4$-cover of $\{\beta \in \mathbb{R}^{d_p} \mid \|\beta\| \le B\}$ with respect to the ℓ_2-norm, and $\mathcal{C}_{M_r}$ be an $\epsilon^2/16$-cover of $\{M_r \in \mathbb{R}^{d_r \times d_r} \mid \|M_r\|_{\mathrm{F}} \le d_r^{1/2} J^2 \rho^{-1}\}$, $\mathcal{C}_{M_p}$ be an $\epsilon^2/16$-cover of $\{M_p \in \mathbb{R}^{d_p \times d_p} \mid \|M_p\|_{\mathrm{F}} \le d_p^{1/2} J^2 \lambda^{-1}\}$ with respect to the norm. By Lemma L.5, we have

$$\begin{aligned}
&|\mathcal{C}_\theta| \le (1 + 8LR/\epsilon)^{d_r}, \quad |\mathcal{C}_\beta| \le (1 + 8B/\epsilon)^{d_p}, \\
&|\mathcal{C}_{M_r}| \le \left[1 + 32 {d_r}^{1/2} J^2/(\rho\epsilon^2)\right]^{d_r^2}, \quad |\mathcal{C}_{M_p}| \le \left[1 + 32 {d_p}^{1/2} J^2/(\beta\epsilon^2)\right]^{d_p^2}.
\end{aligned}$$

Since $\operatorname{dist}(V, V') \le L\|\theta - \theta'\|_2 + \|\beta - \beta'\|_2 + \sqrt{\|M_r - M_r'\|_{\mathrm{F}}} + \sqrt{\|M_p - M_p'\|_{\mathrm{F}}}$, for any $V \in \mathcal{V}_h(R, B, J, \rho, \lambda)$, we can find $\theta' \in \mathcal{C}_\theta$, $\beta' \in \mathcal{C}_\beta$, $M_r' \in \mathcal{C}_{M_r}$, $M_p' \in \mathcal{C}_{M_p}$ such that $\operatorname{dist}(V, V') \le \epsilon$. Hence, it holds that $\mathcal{N}_h(\epsilon; R, B, J, \rho, \lambda) \le |\mathcal{C}_\theta| \cdot |\mathcal{C}_\beta| \cdot |\mathcal{C}_{M_r}| \cdot |\mathcal{C}_{M_p}|$, which gives:

$$\begin{aligned}
\log \mathcal{N}_h(\epsilon; R, B, J, \rho, \lambda) \le& \log |\mathcal{C}_\theta| + \log |\mathcal{C}_\beta| + \log |\mathcal{C}_{M_r}| + \log |\mathcal{C}_{M_p}| \\
\le& d_r \log(1 + 8LR/\epsilon) + d_p \log(1 + 8B/\epsilon) \\
&+ d_r^2 \log\left(1 + 32 {d_r}^{1/2} J^2/(\rho\epsilon^2)\right) + d_p^2 \log\left(1 + 32 {d_p}^{1/2} J^2/(\lambda\epsilon^2)\right).
\end{aligned}$$

This concludes the proof. □

Lemma L.7 (Concentration of Self-Normalized Processes)**.** *Let $V : \mathcal{S} \to [0, H-1]$ be any fixed functions. For any fixed $h \in [H]$ and any $\delta \in (0,1)$, we have*

$$\mathbb{P}_{\mathcal{D}}\left(\Big\| \sum_{\tau=1}^{n} \phi_p(x_h^\tau, a_h^\tau) \epsilon_h^\tau(V) \Big\|_{\Omega_h} > H^2 (2\log(1/\delta) + d_p \log(1 + n/\lambda)) \right) \le \delta.$$

Proof. For the fixed $h \in [H]$, we define the σ-algebra

$$\mathcal{F}_h^\tau = \sigma\big(\{(x_h^i, a_h^i)\}_{i=1}^n \cup \{x_{h+1}^i\}_{i=1}^\tau \big),$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by a set of random variables. For all $\tau \in [n]$, we have $\phi_p(x_h^\tau, a_h^\tau) \in \mathcal{F}_h^{\tau-1}$, as (x_h^τ, a_h^τ) is $\mathcal{F}_h^{\tau-1}$-measurable. Also, for the fixed function V and all $\tau \in [n]$, we have

$$\epsilon_h^\tau(V) = V(x_{h+1}^\tau) - \mathbb{E}\big[V(x_{h+1}) \mid x_h = x_h^\tau, a_h = a_h^\tau\big] \in \mathcal{F}_h^\tau$$

as x_{h+1}^τ is $\mathcal{F}_h^\tau$-measurable. Hence, $\{\epsilon_h^\tau(V)\}_{\tau=1}^n$ is a stochastic process adapted to the filtration $\{\mathcal{F}_{h,j,\tau}\}_{j,\tau=0}^{n_j}$. We have

$$\mathbb{E}_{\mathcal{D}}[\epsilon_h^\tau(V) \mid \mathcal{F}_h^{\tau-1}] = \mathbb{E}[\epsilon_h^\tau(V) \mid \mathcal{F}_h^{\tau-1}] = 0.$$

As a result, $\epsilon_h^\tau(V)$ is mean-zero and H-sub-Gaussian conditioning on $\mathcal{F}_h^{\tau-1}$.

We invoke Lemma L.8 with $M_0 = \lambda \mathbf{I}_{d_p}$ and $M_n = (\Omega_h)^{-1} = \widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p}$. For the fixed function V and fixed $h \in [H]$, we have

$$\mathbb{P}_{\mathcal{D}}\left(\Big\| \sum_{\tau=1}^{n} \phi_p(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V)\Big\|_{\Omega_h} > 2H^2 \cdot \log\big(\frac{\det\left(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p}\right)^{1/2}}{\delta \cdot \det(\lambda \mathbf{I}_{d_p})^{1/2}}\big)\right) \le \delta.$$

for all $\delta \in (0,1)$. Note that $\|\phi_p(x,a)\| \le 1$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$. We have $\|\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p}\|_{\mathrm{op}} \le \lambda + n$. Hence, it holds that $\det\left(\widetilde{\Lambda}_h + \lambda \mathbf{I}_{d_p}\right) \le (\lambda + n)^{d_p}$ an $\det(\lambda \mathbf{I}_{d_p}) = \lambda^{d_p}$, which implies

$$\mathbb{P}_{\mathcal{D}}\left(\Big\| \sum_{\tau=1}^{n} \phi_p(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V)\Big\|_{\Omega_h} > H^2(2\log(1/\delta) + d_p \log(1 + n/\lambda))\right) \le \delta.$$

Therefore, we finish the proof. □

Lemma L.8 (Concentration of Self-Normalized Processes)**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration and $\{\epsilon_t\}_{t=1}^{\infty}$ be an $\mathbb{R}$-valued stochastic process such that ϵ_t is $\mathcal{F}_t$-measurable for all $t \ge 1$. Moreover, suppose that conditioning on $\mathcal{F}_{t-1}$, ϵ_t is a zero-mean and σ-sub-Gaussian random variable for all $t \ge 1$, that is,*

$$\mathbb{E}[\epsilon_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}\big[\exp(\lambda \epsilon_t) \mid \mathcal{F}_{t-1}\big] \le \exp(\lambda^2 \sigma^2/2), \qquad \forall \lambda \in \mathbb{R}.$$

Meanwhile, let $\{\phi_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that ϕ_t is $\mathcal{F}_{t-1}$-measurable for all $t \ge 1$. Also, let $\mathbf{M}_0 \in \mathbb{R}^{d \times d}$ be a deterministic positive-definite matrix and

$$\mathbf{M}_t = \mathbf{M}_0 + \sum_{s=1}^{t} \phi_s \phi_s^{\mathsf{T}}$$

for all $t \ge 1$. For all $\delta > 0$, it holds that

$$\Big\| \sum_{s=1}^{t} \phi_s \epsilon_s \Big\|_{\mathbf{M}_t^{-1}}^2 \le 2\sigma^2 \cdot \log\Big(\frac{\det(\mathbf{M}_t)^{1/2} \det(\mathbf{M}_0)^{-1/2}}{\delta}\Big)$$

for all $t \ge 1$ with probability at least $1 - \delta$.

Proof. See Theorem 1 of Abbasi-yadkori et al. (2011) for a detailed proof. □

Lemma L.9 (Concentration)**.** *When n is sufficiently large, for any $h \in [H]$, it holds for any $h \in [H]$ that*

$$\big\|n^{-1}\widetilde{\Lambda}_h - \mathcal{C}_h\big\| \le C\sqrt{\log(dHK/\xi)/n}$$

with probability at least $1 - \xi$, where $C > 0$ is an absolute constant and the expectation $\mathbb{E}_{\mathcal{D}}$ is with respect to the data collecting process.

Proof. We notice that

$$\frac{\widetilde{\Lambda}_h}{n} - \mathcal{C}_h = \frac{1}{n}\sum_{\tau=1}^{n}\Big(\phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top - \mathbb{E}_{\mathcal{D}}\big[\phi(s_h, a_h)\phi(s_h, a_h)^\top\big]\Big) = \frac{\sum_{\tau=1}^{n} Z_h^\tau}{n},$$

where we write

$$Z_h^\tau = \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top - \mathbb{E}_{\mathcal{D}}\big[\phi(x_h, a_h)\phi(x_h, a_h)^\top \mid x_1 = x\big].$$

We notice that $\|Z_h^\tau\| \le 2$. Then, applying the matrix Bernstein inequality (Tropp, 2015), we have that

$$\|n^{-1}\sum_{\tau=1}^{n} Z_h^\tau\| \le C\sqrt{\log(d/\xi)/n}$$

with probability at least $1 - \xi$ for an absolute constant $C > 0$. Applying the union bound, by the definition of Z_h^τ, we have for any $h \in [H]$ that

$$\Big\|n^{-1}\widetilde{\Lambda}_h - \mathbb{E}_{\mathcal{D}}\big[\phi(x_h, a_h)\phi(x_h, a_h)^\top\big]\Big\| \le C\sqrt{\log(dHK/\xi)/n}$$

with probability at least $1 - \xi$. Thus, we complete the proof of Lemma L.9. □

Lemma L.10. *Under Assumptions 3 and 4, we have for any* $p_1 \in (0,1)$

$$\begin{aligned}&\left|g(\phi_r(x,a)^\top \widehat{\theta}_{h,t}) - |g(\phi_r(x,a)^\top \theta_h^*)\right| \\ &\leq K \cdot \sqrt{4M^2 + \frac{3 + 16[d_r \ln(2Mt) + \ln(TH/p_1)]}{k}} \|\phi_r(x,a)\|_{\Lambda_{h,t}'^{-1}}\end{aligned}$$

holds with probability at least $1 - p_1$ *for any* $h \in [H]$ *and* $t \in [T]$.

Proof. By the definition of $\widehat{\theta}_{h,t}$, we have

$$\sum_{\tau=1}^{t} r_{h,\tau} \langle \phi_r(x_{h,\tau}, a_{h,\tau}), \theta_h^* - \widehat{\theta}_{h,t} \rangle \leq \sum_{\tau=1}^{t} \int_{\langle \phi_r(x_{h,\tau}, a_{h,\tau}), \widehat{\theta}_{h,t} \rangle}^{\langle \phi_r(x_{h,\tau}, a_{h,\tau}), \theta_h^* \rangle} g(u) du,$$

Then we have

$$\begin{aligned}&\sum_{\tau=1}^{t} (r_{h,\tau} - g(\langle \phi_r(x_{h,\tau}, a_{h,\tau}), \theta_h^* \rangle) \langle \phi_r(x_{h,\tau}, a_{h,\tau}), \widehat{\theta}_{h,t} - \theta_h^* \rangle \\ &\geq \sum_{\tau=1}^{t} \int_{\langle \phi_r(x_{h,\tau}, a_{h,\tau}), \theta_h^* \rangle}^{\langle \phi_r(x_{h,\tau}, a_{h,\tau}), \widehat{\theta}_{h,t} \rangle} g(u) - g(\langle \phi_r(x_{h,\tau}, a_{h,\tau}), \theta_h^* \rangle) du,\end{aligned} \tag{L.23}$$

The right side is larger than

$$\sum_{\tau=1}^{t} k \langle \phi_r(x_{h,\tau}, a_{h,\tau}), \widehat{\theta}_{h,t} - \theta_h^* \rangle^2 := k V_{t,h}.$$

Also we can bound the left side of equation L.23 by using Lemma 9 in Wang et al. (2019), that is , for any $\delta \in (0,1)$ with probability at least $1 - \delta$, we have

$$\begin{aligned}&\sum_{\tau=1}^{t} (r_{h,\tau} - g(\langle \phi_r(x_{h,\tau}, a_{h,\tau}), \theta_h^* \rangle) \langle \phi_r(x_{h,\tau}, a_{h,\tau}), \widehat{\theta}_{h,t} - \theta_h^* \rangle \\ &\leq 1 + 2\Big(1 + \sqrt{V_{h,t}}\Big) \sqrt{d_r \ln(2Mt) + \ln 1/\delta}.\end{aligned}$$

Hence

$$\begin{aligned}&k V_{h,t} \leq 1 + 2\Big(1 + \sqrt{V_{h,t}}\Big) \sqrt{d_r \ln(2Mt) + \ln 1/\delta} \\ &\leq \max\Big\{2 + 4\sqrt{\ln\Big(\frac{2Mt}{\delta}\Big)}, \sqrt{V_{h,t}} \sqrt{d_r \ln 2Mt + \ln 1/\delta}\Big\}.\end{aligned}$$

Then we have

$$V_{h,t} \leq \frac{3 + 16[d_r \ln 2Mt + \ln(TH/p_1)]}{k}$$

with probability at least $1 - p_1$ for any $h \in [H]$, $t \in [T]$. Then

$$\begin{aligned}&\left|g(\phi_r(x,a)^\top \widehat{\theta}_{h,t}) - |g(\phi_r(x,a)^\top \theta_h^*)\right| \\ &\leq K \|\phi_r(x,a)\|_{\Lambda_{h,t}'^{-1}} \sqrt{V_{h,t} + 4M^2} \\ &\leq K \cdot \|\phi_r(x,a)\|_{\Lambda_{h,t}'^{-1}} \sqrt{\frac{3 + 16[d_r \ln(2Mt) + \ln(TH/p)]}{k} + 4M^2}\end{aligned}$$

□

Lemma L.11. *Under the assumptions of Theorem 3, we have*

$$|\langle \phi_p(x,a), \widehat{\beta}_{h,t} \rangle - \mathbb{E}(V_{h+1,t}|x,a)| \leq \gamma_p \|\phi_p(x,a)\|_{\Lambda_{h,t}^{-1}}$$

holds for any $h \in [H]$, $t \in [T]$ *with probability at least* $1 - p_0/3$.

Proof. We can easily follow the proof of Lemmas B.2 and B.3 in Jin et al. (2020) to obtain the next two properties of $\widehat{\beta}_{h,t}$:

$$\|\widehat{\beta}_{h,t}\| \leq H\sqrt{d_p t}$$

and if we define $V_{h+1,t} = \max_{a\in\mathcal{A}} \bar{Q}_{h+1}(x_{h+1,t}, a)$ we have for any $p_2 \in (0,1)$, there exists a universal constant C such that

$$\begin{aligned}&\|\sum_{\tau=1}^{t} \phi_p(x_{h,\tau}, a_{h,\tau})[V_{h+1,\tau}(x_{h+1,\tau}) - \mathbb{E}(V_{h+1,\tau}|x_{h,\tau}, a_{h,\tau})]\|_{\Lambda_{h,t}^{-1}}\\ &\leq Cd_pH\sqrt{\ln[3(c_p+1)d_pTH/p_0]}\end{aligned}$$

holds for any $h \in [H]$, $t \in [T]$ with probability at least $1 - p_0/3$. Define $\beta_{h,t} = \int V_{h,t}(x')d\mu_h(x')$ which is smaller than $H\sqrt{d}$, and then

$$\begin{aligned}&|\langle\phi_p(x,a), \widehat{\beta}_{h,t}\rangle - \mathbb{E}(V_{h+1,t}|x,a)|\\ &= |\langle\phi_p(x,a), \widehat{\beta}_{h,t} - \beta_{h,t}\rangle|\\ &= |\langle\phi_p(x,a)\Lambda_{h,t}^{-1}[-\beta_{h,t} + \sum_{\tau=1}^{t}\phi_p(x_{h,\tau}, a_{h,\tau})[V_{h+1,t}(x_{h+1,\tau}) - \mathbb{E}(V_{h+1,t}|x_{h,\tau}, a_{h,\tau})]\rangle|\\ &\leq \left(H\sqrt{d_p} + Cd_pH\sqrt{\ln[3(c_r+1)d_pTH/p_0]}\right)\|\phi_p(x,a)\|_{\Lambda_{h,t}^{-1}}\\ &\leq \gamma_p\|\phi_p(x,a)\|_{\Lambda_{h,t}^{-1}}.\end{aligned}$$

The last inequality holds if c_p is not too small. □

Lemma L.12. *Under the assumptions of Theorem 3, with probability* $1 - 2p_0/3$*,*

$$\bar{Q}_{h,t}(x,a) \geq Q_h^*(x,a)$$

holds for all h, t, s, a*.*

Proof. We prove this lemma by induction on h. When $h = H$, it is trivial. Suppose $\bar{Q}_{h+1,t}(x,a) \geq Q_{h+1}^*(x,a)$, then we have

$$\bar{Q}_h(x,a) \geq \mathbb{E}(r_h + V_{h+1,t}|x,a) \geq \mathbb{E}(r_h + \max_{a'\in\mathcal{A}} Q_{h+1}^*|x,a) \geq Q_h^*(x,a).$$

holds with probability at least $1 - 2p_0/3$. □

M Relaxing Assumption 2

Given a regularization parameter $\lambda' = \lambda'_n \gg n^{-1/2}$, we define

$$\mathcal{L}_{h,\lambda'}(\theta) \coloneqq \frac{1}{n}\sum_{\tau=1}^{n}\big(-r_h^\tau\langle\phi_r(x_h^\tau, a_h^\tau), \theta\rangle) + G(\langle\phi_r(x_h^\tau, a_h^\tau), \theta\rangle) + \lambda'\|\theta\|_2^2\big),$$

and a new estimator for $\theta_h^\star$ is defined as

$$\widetilde{\theta}_{h,\lambda'} \coloneqq \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}_{h,\lambda'}(\theta).$$

We also define

$$\begin{aligned}l_{h,\lambda'}(\theta) &\coloneqq -r_h\langle\phi_r(x_h, a_h), \theta\rangle) + G(\langle\phi_r(x_h, a_h), \theta\rangle) + \lambda'\|\theta\|_2^2,\\ \theta_{h,\lambda'}^* &\coloneqq \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}_\pi l_{h,\lambda'}(\theta),\\ \Sigma_{h,\lambda'}(\theta) &\coloneqq \nabla^2[\mathbb{E}_\pi l_{h,\lambda'}(\theta)],\end{aligned}$$

and

$$\widetilde{\Sigma}_{h,\lambda'}(\theta) \coloneqq n\nabla^2\mathcal{L}_{h,\lambda'}(\theta) = \sum_{\tau=1}^{n}\dot{g}(\langle\phi_r(x_h^\tau, a_h^\tau), \theta\rangle)\phi_r(x_h^\tau, a_h^\tau)\phi_r(x_h^\tau, a_h^\tau)^\intercal + 2\lambda'\mathbf{I}_{d_r}.$$

Lemma M.13. *Under Assumptions E.1 and 1, We have*

$$\|\theta^\star_{h,\lambda'} - \theta^\star_h\|_2 \le \|\theta^\star_h\|_2$$

holds for any $h \in [H]$.

Proof. By the definitions of $\theta^\star_h$ and $\theta^\star_{\lambda,h}$, we have

$$0 = \mathbb{E}_\pi\big[\big(g(\langle\phi_r, \theta^\star_{h,\lambda'}\rangle) - g(\langle\phi_r, \theta^\star_h\rangle)\big)\phi_r\big] + 2\lambda'\theta^\star_{h,\lambda'}$$

Using Taylor's expansion for $F_v(\theta) \coloneqq \mathbb{E}_\pi\big[g(\langle\phi_r, \theta\rangle)\langle\phi_r, v\rangle\big]$ where $v \in R^{d_r}$, we have that there exists $t \in [0,1]$ and $\bar\theta = t\theta^\star_h + (1-t)\theta^\star_{h,\lambda'}$, such that

$$-2\lambda'\langle v, \theta^\star_{h,\lambda'}\rangle = F_v(\theta^\star_{h,\lambda'}) - F_v(\theta^\star_h) = \langle\mathbb{E}_\pi\big[\dot g(\langle\phi_r, \bar\theta\rangle)\phi_r\phi_r^\intercal\big]v, \theta^\star_{h,\lambda'} - \theta^\star_h\rangle \tag{M.24}$$

We take $v = \theta^\star_h - \theta^\star_{h,\lambda'}$, then

$$\begin{aligned}2\lambda'\|v\|_2\|\theta^\star_h\|_2 \ge -2\lambda'\langle v, \theta^\star_h\rangle &= \langle\big(2\lambda'\mathbf{I}_{d_r} + \mathbb{E}_\pi\big[\dot g(\langle\phi_r, \bar\theta\rangle)\phi_r\phi_r^\intercal\big]\big)v, \theta^\star_{h,\lambda'} - \theta^\star_h\rangle \\ &\ge 2\lambda'\|\theta^\star_{h,\lambda'} - \theta^\star_h\|_2^2.\end{aligned}$$

Hence $\|\theta^\star_{h,\lambda'} - \theta^\star_h\|_2 \le \|\theta^\star_h\|_2$. □

Lemma M.14. *Under Assumption E.1 and 1, for $\xi \in (0,1)$, with sufficiently large n and $h \in [H]$ fixed, we have*

$$\|\widetilde\theta_{h,\lambda'} - \theta^\star_{h,\lambda'}\|^2_{\frac{1}{n}\widetilde\Sigma_{h,\lambda'}} \le \frac{C_1 d_r \log 1/\xi}{n\lambda'} \text{ and } \|\nabla\mathcal{L}_h(\theta^\star_h)\|^2_{\Sigma^{-1}_{h,\lambda'}} \le \frac{C_1 d_r \log 1/\xi}{n\lambda'}$$

with probability at least $1-\xi$ for some absolute constant $C_1 > 0$. Here we abbreviate $\Sigma_{h,\lambda'}(\theta^\star_{h,\lambda'})$ and $\widetilde\Sigma_{h,\lambda'}(\theta^\star_{h,\lambda'})$ to $\Sigma_{h,\lambda'}$ and $\widetilde\Sigma_{h,\lambda'}$.

Proof. Apply Theorem 2.1 in Hsu et al. (2012) to $A = \Sigma^{-1/2}_{h,\lambda'}$ and $x = \sqrt{n}\nabla\mathcal{L}_{h,\lambda'}(\theta^*_{h,\lambda'})$. Since

$$\nabla l_{h,\lambda'}(\theta^*_{h,\lambda'}) = \big(r_h - g(\langle\phi_h, \theta^*_{h,\lambda'}\rangle)\big)\phi_h + 2\lambda'\theta^*_{h,\lambda'}$$

is subguassian since we assume $r_h - g(\langle\phi_h, \theta^*\rangle)$ is subguassian, and $\theta^*_{h,\lambda'}$, θ^*_h lies in a compact space, there exists $\sigma > 0$ independent with n such that for any $t > 0$,

$$\mathbb{P}\Big(\|x\|^2_{\Sigma^{-1}_{h,\lambda'}} - \sigma\big(\operatorname{Tr}(\Sigma^{-1}_{h,\lambda'}) + 2\|\Sigma^{-1}_{h,\lambda'}\|t + 2\sqrt{\operatorname{Tr}(\Sigma^{-2}_{h,\lambda'})t}\,\big) > 0\Big) \le e^{-t}.$$

Hence

$$\|\nabla\mathcal{L}_h(\theta^*_h)\|^2_{\Sigma^{-1}_{h,\lambda'}} \le \frac{3\sigma d_r\|\Sigma^{-1}_{h,\lambda'}\|\log(1/\xi)}{n}$$

with

holds with probability at least $1-\xi$, for any (small) $\xi > 0$.

Similarly, for any $1 \le \tau \le n$, applying Theorem 2.1 in Hsu et al. (2012) to $A = \Sigma^{-1/2}_{h,\lambda'}$ and $x = \phi^\tau_h$, we have there exists $\sigma' > 0$, such that

$$\max_{1\le\tau\le n}\|\phi^\tau_h\|^2_{\Sigma^{-1}_{h,\lambda'}} \le 3\sigma' d_r\|\Sigma^{-1}_{h,\lambda'}\|\log(n/\xi).$$

holds with probability at least $1-\xi$, for any (small) $\xi > 0$. Using Theorem A.2 in Ostrovskii and Bach (2021), we have

$$\max_{1\le\tau\le n}\|\phi^\tau_h\|_{\widetilde\Sigma^{-1}_{h,\lambda'}}\|\nabla\mathcal{L}_h(\theta^*_h)\|^2_{\widetilde\Sigma^{-1}_{h,\lambda'}} \le \frac{1}{4}$$

holds with probability at least $1-3\xi$ if n satisfies $n \ge 36\sigma\sigma' d_r^3 L\lambda'^{-2}[\log(n) - \log(\xi)]\log(1/\xi)$ and $n \ge C_3\lambda'^{-2}[d_r + \log(1/\xi)]$ for some $C_3 > 0$. It's easy to check $\mathcal{L}_{h,\lambda'}$ falls into the case (a) of Proposition B.3 in Ostrovskii and Bach (2021) with $\theta_0 = \theta^*_{h,\lambda'}$, $H_0 = \Sigma_{h,\lambda',n}$, $W(\theta) = \phi^{\tau(\theta)}$ where $\tau(\theta) \coloneqq \operatorname{argmin}_{1\le\tau\le n}|\langle\phi^\tau, \theta - \theta^*_{h,\lambda'}\rangle|$. Then using Proposition B.4 in Ostrovskii and Bach (2021), we have

$$\|\widetilde\theta_{h,\lambda'} - \theta^*_{h,\lambda'}\|^2_{\frac{1}{n}\widetilde\Sigma_{h,\lambda'}} \le 4\|\nabla\mathcal{L}_h(\theta^*_h)\|^2_{\left(\frac{1}{n}\Sigma_{h,\lambda'}\right)^{-1}} \le \frac{12\sigma\log(1/\xi)}{n\lambda'}$$

holds with probability at least $1-3\xi$. if $n \ge K_1\lambda'^{-2}[\log(1/\xi)]^{1+\epsilon}$, for some constant $K_1 > 0$, any small $\xi > 0$ and $\epsilon > 0$. □

Lemma M.15. *We denote $\mathbb{E}_\pi \phi_r(x_h, a_h)\phi_r(x_h, a_h)^\top$ and $\sum_{\tau=1}^n \phi_r(x_h^\tau, a_h^\tau)\phi_r(x_h^\tau, a_h^\tau)^\top$ as $\Lambda_{r,h}$ and $\widetilde{\Lambda}_{r,h}$. Under Assumptions E.1 and 1, for $\xi \in (0,1)$, with sufficiently large n, we have,*

$$\begin{aligned}
&\big|g(\langle\phi_r(x,a), \widetilde{\theta}_{h,\lambda'}\rangle) - g(\langle\phi_r(x,a), \theta_h^\star\rangle)\big| \\
&\le C_2|\dot{g}(\langle\phi_r(x,a), \widetilde{\theta}_{h,\lambda'}\rangle)|\Big(\sqrt{\frac{d_r \log H d_r/\xi}{\lambda'^2}} + \frac{d_r \log H/\xi}{n^{1/2}\lambda'^{7/2}}\Big)\|\phi_r(x,a)\|_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}} \qquad \text{(M.25)} \\
&+ C_2 L\sqrt{\lambda'}\big(\|\phi_r(x,a)\|_{(2\lambda'\mathbf{I}_{d_r} + \frac{1}{n}\widetilde{\Lambda}_{r,h})^{-1}} + \frac{[\log(Hd_r/\xi)]^{1/4}}{\lambda' n^{1/4}}\big),
\end{aligned}$$

for all $(x,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$ with probability at least $1-\xi$ for some absolute constant $C_2 > 0$.

Proof. Note that

$$\begin{aligned}
\big|g(\langle\phi_r(x,a), \widetilde{\theta}_{h,\lambda'}\rangle) - g(\langle\phi_r(x,a), \theta_h^\star\rangle)\big| &\le \big|g(\langle\phi_r(x,a), \theta_{h,\lambda'}^\star\rangle) - g(\langle\phi_r(x,a), \theta_h^\star\rangle)\big| + \\
&\quad \big|g(\langle\phi_r(x,a), \theta_{h,\lambda'}^\star\rangle) - g(\langle\phi_r(x,a), \widetilde{\theta}_h\rangle)\big| \\
&\coloneqq k_1 + k_2\,.
\end{aligned}$$

First we bound k_1. Similar to the argument in the end of the proof of Lemma L.3, we have $\inf_{x\in\mathcal{S}, a\in\mathcal{A}, \lambda'>0, h\in[H], t\in[0,1]} \dot{g}(\langle\phi_r(x,a), t\theta_h^\star + (1-t)\theta_{h,\lambda'}^\star\rangle) > 0$. We arbitrarily take a lower bound of this term, say $c_1 \in (0,1)$. Taking $v = \theta_{h,\lambda'}^\star - \theta_h^\star$ in equation M.24, we have

$$\begin{aligned}
c_2\lambda' &\ge -2\lambda'\langle\theta_{h,\lambda'}^\star - \theta_h^\star, \theta_h^\star\rangle \\
&= \langle\big(2\lambda'\mathbf{I}_{d_r} + \mathbb{E}_\pi\big[\dot{g}(\langle\phi_r, \bar{\theta}\rangle)\phi_r\phi_r^\top\big]\big)\theta_{h,\lambda'}^\star - \theta_h^\star, \theta_{h,\lambda'}^\star - \theta_h^\star\rangle \\
&\ge \|\theta_{h,\lambda'}^\star - \theta_h^\star\|^2_{2\lambda'\mathbf{I}_{d_r} + c_1\Lambda_h}
\end{aligned}$$

for some constant $c_2 > 0$. Hence

$$\begin{aligned}
k_1 &\le L|\langle\phi_r(x,a), \theta_{h,\lambda'}^\star - \theta_h^\star\rangle| \\
&\le L\|\phi_r(x,a)\|_{(2\lambda'\mathbf{I}_{d_r} + c_1\Lambda_{r,h})^{-1}}\|\theta_{h,\lambda'}^\star - \theta_h^\star\|_{2\lambda'\mathbf{I}_{d_r} + c_1\Lambda_{r,h}} \\
&\le L\sqrt{c_2\lambda'/c_1}\big(\|\phi_r(x,a)\|_{(2\lambda'\mathbf{I}_{d_r} + \frac{1}{n}\widetilde{\Lambda}_{h,r})^{-1}} + |\|\phi_r(x,a)\|_{(2\lambda'\mathbf{I}_{d_r} + \frac{c_1}{n}\widetilde{\Lambda}_{r,h})^{-1}} \\
&\quad - \|\phi_r(x,a)\|_{(2\lambda'\mathbf{I}_{d_r} + c_1\Lambda_{r,h})^{-1}}|\big) \\
&\le L\sqrt{c_2\lambda'/c_1}\big(\|\phi_r(x,a)\|_{(2\lambda'\mathbf{I}_{d_r} + \frac{1}{n}\widetilde{\Lambda}_{r,h})^{-1}} + \frac{c_3[\log(Hd_r/\xi)]^{1/4}}{\lambda' n^{1/4}}\big),
\end{aligned}$$

where the last inequality comes from

$$\begin{aligned}
&|\|\phi_r(x,a)\|_{(2\lambda'\mathbf{I}_{d_r} + c_1\widetilde{\Lambda}_h)^{-1}} - \|\phi_r(x,a)\|_{(2\lambda'\mathbf{I}_{d_r} + c_1\Lambda_h)^{-1}}| \\
&\le \sqrt{|\phi_r(x,a)^\top\big((2\lambda'\mathbf{I}_{d_r} + \Lambda_{r,h})^{-1} - (2\lambda'\mathbf{I}_{d_r} + \widetilde{\Lambda}_{r,h})^{-1}\big)\phi_r(x,a)|} \\
&\le \|(2\lambda'\mathbf{I}_{d_r} + \Lambda_{r,h})^{-1} - (2\lambda'\mathbf{I}_{d_r} + \widetilde{\Lambda}_{r,h})^{-1}\|^{1/2} \qquad \text{(M.26)} \\
&\le \frac{1}{2\lambda'}\|\Lambda_{r,h} - \widetilde{\Lambda}_{r,h}\|^{1/2} \\
&\le \frac{c_3[\log(Hd_r/\xi)]^{1/4}}{\lambda' n^{1/4}}
\end{aligned}$$

holds with probability $1-\xi$ for some constant $c_3 > 0$. The last inequality uses the matrix Bernstein inequality Tropp (2015).

Then we bound k_2. By Taylor's theorem, there exists $t \in [0,1]$ and $\bar{\theta} = t\theta_{h,\lambda'}^\star + (1-t)\widetilde{\theta}_{h,\lambda'}$ such that

$$\begin{aligned}
0 &= \langle\phi_r(x,a), \big(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\big)^{-1}\nabla\mathcal{L}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\rangle \\
&= \langle\phi_r(x,a), \big(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\big)^{-1}\nabla\mathcal{L}_{h,\lambda'}(\theta_{h,\lambda'}^\star)\rangle + \langle\phi_r(x,a), \widetilde{\theta}_{h,\lambda'} - \theta_{h,\lambda'}^\star\rangle \\
&+ \frac{1}{2}\langle\widetilde{\theta}_{h,\lambda'} - \theta_{h,\lambda'}^*, \big[\frac{1}{n}\sum_{\tau=1}^n \ddot{g}(\langle\phi_r(x_h^\tau, a_h^\tau), \bar{\theta}_{h,\lambda'}\rangle) \qquad \text{(M.27)} \\
&\langle\big(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\big)^{-1}\phi_r(x,a), \phi_r(x_h^\tau, a_h^\tau)\rangle\phi_r(x_h^\tau, a_h^\tau)\phi_r(x_h^\tau, a_h^\tau)^\top\big](\widetilde{\theta}_{h,\lambda'} - \theta_{h,\lambda'}^*)\big\rangle\,.
\end{aligned}$$

By Taylor's theorem again, we have

$$\begin{aligned} g(\langle\phi_r(x,a),\widetilde{\theta}_{h,\lambda'}\rangle) - g(\langle\phi_r(x,a),\theta^\star_{h,\lambda'}\rangle) =& \dot{g}(\langle\phi_r(x,a),\widetilde{\theta}_{h,\lambda'}\rangle)\langle\phi_r(x,a),\widetilde{\theta}_{h,\lambda'}-\theta^\star_{h,\lambda'}\rangle \\ &+ \frac{1}{2}\ddot{g}(\langle\phi_r(x,a),\breve{\theta}_{h,\lambda'}\rangle)\langle\phi_r(x,a),\widetilde{\theta}_{h,\lambda'}-\theta^\star_{h,\lambda'}\rangle^2 \\ :=& \dot{g}(\langle\phi_r(x,a),\widetilde{\theta}_{h,\lambda'}\rangle)e_3 + e_4 \end{aligned}$$

for some $\breve{\theta}_{h,\lambda'}$ on the line segment between $\theta^\star_{h,\lambda'}$ and $\widetilde{\theta}_{h,\lambda'}$. We then bound $|e_3|$ and $|e_4|$ separately. First, by equation M.27, we have

$$\begin{aligned} e_3 =& \big\langle\phi_r(x,a), -\big(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\theta^\star_{h,\lambda'})\big)^{-1}\nabla\mathcal{L}_h(\theta^\star_{h,\lambda'})\big\rangle \\ &+ \frac{1}{2}\big\langle\widetilde{\theta}_{h,\lambda'}-\theta^*_{h,\lambda'}, \big[\frac{1}{n}\sum_{\tau=1}^n \ddot{g}(\langle\phi_r(x_h^\tau,a_h^\tau),\bar{\theta}_{h,\lambda'}\rangle) \\ &\big\langle\big(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\big)^{-1}\phi_r(x,a),\phi_r(x_h^\tau,a_h^\tau)\rangle\phi_r(x_h^\tau,a_h^\tau)\phi_r(x_h^\tau,a_h^\tau)^\intercal\big](\widetilde{\theta}_{h,\lambda'}-\theta^*_{h,\lambda'})\big\rangle\,. \end{aligned}$$

Besides, we have

$$\begin{aligned} \|\big(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\theta^\star_{h,\lambda'}) - \frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\big)\| &\le \frac{L}{n}\sum_{\tau=1}^n |\langle\phi_r(x_h^\tau,a_h^\tau),\theta^\star_{h,\lambda'}-\widetilde{\theta}_{h,\lambda'}\rangle|\|\phi_r(x_h^\tau,a_h^\tau)\|_2^2 \\ &\le L\|\theta^\star_{h,\lambda'}-\widetilde{\theta}_{h,\lambda'}\|_2\,. \end{aligned} \tag{M.28}$$

By Lemma M.14, we have

$$\|\widetilde{\theta}_{h,\lambda'}-\theta^\star_{h,\lambda'}\|_2^2 \le \frac{1}{\lambda'}\|\widetilde{\theta}_{h,\lambda'}-\theta^\star_{h,\lambda'}\|^2_{\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}} \le \frac{C_1 d_r\log H/\xi}{n\lambda'^2} \tag{M.29}$$

with probability at least $1-\xi$, for any $h\in[H]$. Then we have

$$\begin{aligned} e_3 &\le \big|\big\langle\phi_r(x,a),\big(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\theta^\star_{h,\lambda'})\big)^{-1}\big(-\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\big\rangle\big| \\ &+ \frac{LC_1d_r\|\phi_r(x,a)\|_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}}\log H/\xi}{n^{1/2}\lambda'^{5/2}} \\ &\le \big|\big\langle\phi_r(x,a),\big(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\big)^{-1}\big(-\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\big\rangle\big| \\ &+ \big|\big\langle\phi_r(x,a),n\big(\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}-\widetilde{\Sigma}_{h,\lambda'}(\theta^\star_{h,\lambda'})^{-1}\big)\big(-\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\big\rangle\big| \\ &+ \frac{LC_1d_r\|\phi_r(x,a)\|_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}}\log H/\xi}{n^{1/2}\lambda'^{5/2}} \\ &\le \|\phi_r(x,a)\|_{\left(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda})^{-1}\right)}\Big\|\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\Big\|_{\left(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\right)^{-1}} \\ &+ \frac{L}{\lambda'}\|\phi_r(x,a)\|_{\left(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\right)^{-1}}\|\widetilde{\theta}_{h,\lambda'}-\theta^\star_{h,\lambda'}\|_2\Big\|\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\Big\|_{\left(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\theta^\star_{h,\lambda'})\right)^{-1}} \\ &+ \frac{LC_1d_r\|\phi_r(x,a)\|_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}}\log H/\xi}{n^{1/2}\lambda'^{5/2}} \end{aligned}$$

by equation M.29 and the Cauchy-Schwarz inequality. By the matrix Bernstein inequality (Tropp, 2015) and $\|\phi_r(x,a)\|_2\le 1$, we have

$$\|\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\theta^\star_{h,\lambda'}) - \Sigma_{h,\lambda'}(\theta^\star_{h,\lambda'})\| \le C\sqrt{\log(Hd_r/\xi)/n} \tag{M.30}$$

with probability at least $1-\xi/2$ for an absolute constant $C>0$. By equation M.29 and Lemma M.14, similar toequation M.28, we have

$$\begin{aligned}
&\left\|\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\right\|_{\left(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\right)^{-1}} \le \left\|\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\right\|_{\Sigma_{h,\lambda'}(\theta^*_{h,\lambda'})^{-1}} \\
&+\sqrt{K\lambda'^{-3/2}\left\|\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})-\Sigma_{h,\lambda'}(\theta^\star_{h,\lambda'})\right\|\left\|\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\right\|_{\Sigma_{h,\lambda'}(\theta^*_{h,\lambda'})^{-1}}} \\
&\le \sqrt{\frac{C_1 d_r\log H/\xi}{n\lambda'}}+\sqrt{K\Big(L\frac{C_1 d_r\log H/\xi}{n\lambda'^3}+C\sqrt{\frac{C_1 d_r\log(Hd_r/\xi)\log H/\xi}{n^2\lambda'^4}}\Big)},
\end{aligned}$$

where K is chosen to be a upper bound of $\|\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\|_2$ with high probability. Such a upper bound exists since the noise is subguassian and $\theta^\star_{h,\lambda'}$ is close to $\theta^\star_h$. Similarly,

$$\left\|\nabla\mathcal{L}_{h,\lambda'}(\theta^\star_{h,\lambda'})\right\|_{\left(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\theta^\star_{h,\lambda'})\right)^{-1}} \le \sqrt{\frac{C_1 d_r\log H/\xi}{n\lambda'}}+\sqrt{KL\frac{C_1 d_r\log H/\xi}{n\lambda'^3}}.$$

Thus, we get

$$\begin{aligned}
|e_3| &\le \left(\sqrt{\frac{C_1 d_r\log H/\xi}{n\lambda'}}+\sqrt{K\Big(L\frac{C_1 d_r\log H/\xi}{n\lambda'^3}+C\sqrt{\frac{C_1 d_r\log(Hd_r/\xi)\log H/\xi}{n^2\lambda'^4}}\Big)}\right.\\
&\quad\left.+\frac{L}{\lambda'}\left[\frac{C_1 d_r\log H/\xi}{n\lambda'^{3/2}}+\sqrt{KL}\frac{C_1 d_r\log H/\xi}{n\lambda'^{5/2}}\right]\right)\|\phi_r(x,a)\|_{\left(\frac{1}{n}\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})\right)^{-1}}\\
&\quad+\frac{LC_1 d_r\|\phi_r(x,a)\|_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}}\log H/\xi}{n^{1/2}\lambda'^{5/2}}\\
&\le C'\left(\sqrt{\frac{KCC_1 d_r\log Hd_r/\xi}{\lambda'^2}}+\frac{C_1 d_r\log H/\xi}{\sqrt{n}\lambda'^{7/2}}\right)\|\phi_r(x,a)\|_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}}.
\end{aligned}$$

Finally,

$$\begin{aligned}
|e_4| &= \left|\frac{1}{2}\ddot{g}(\langle\phi_r(x,a),\breve{\theta}_h\rangle)\langle\phi_r(x,a),\widetilde{\theta}_{h,\lambda'}-\theta^\star_{h,\lambda'}\rangle^2\right|\\
&\le \frac{L}{2}\|\phi_r(x,a)\|^2_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}}\|\theta^\star_{h,\lambda'}-\widetilde{\theta}_{h,\lambda'}\|^2_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})}\\
&\le \frac{c'Ld_r\log H/\xi}{2\lambda'^2}\|\phi_r(x,a)\|^2_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}}\\
&\le \frac{C'}{\sqrt{n}\lambda'^3}\sqrt{d_r\log H/\xi}\dot{g}(\langle\phi(x,a),\widetilde{\theta}_h\rangle)\|\phi_r(x,a)\|_{\widetilde{\Sigma}_{h,\lambda'}(\widetilde{\theta}_{h,\lambda'})^{-1}}
\end{aligned}$$

for some constant $c',C'>0$ with sufficiently large n, where the first inequality comes from the Cauchy-Schwarz inequality, Assumption 1, the second inequality comes from Lemma M.14, and the last inequality comes from an argument similar to the end of the proof of Lemma L.3. Combining these derivations and using the union bounds, we finish the proof. □

Then we claim a theorem analogous to Theorem 1 with respect to Algorithm B.1, where $\widetilde{\Gamma}_h$ is replaced by $\widetilde{\Gamma}_{h,\lambda'}$. We define

$$\widetilde{\Gamma}_{h,\lambda'}(x,a) \coloneqq \widetilde{\Gamma}_{r,h,\lambda'}(x,a)+\widetilde{\Gamma}_{p,h}(x,a)$$

with $\widetilde{\Gamma}_{r,h,\lambda'}$ equal to the right side of equation M.25.

Theorem M.3 (Suboptimality for GPEVI without Assumption 2)**.** *Under Assumptions E.1 and 1, we set $\lambda=1$, $\alpha_p=c_p\left(d_p+d_r\right)H\sqrt{\zeta}$, where $\zeta=\log\left(2\left(d_r+d_p\right)Hn/\xi\right)$, $c_r,c_p>0$ are absolute constants and $\xi\in(0,1)$ is the confidence parameter. Then $\{\widetilde{\Gamma}_{h,\lambda'}\}_{h=1}^H$ in equation 10 is a ξ-uncertainty quantifier of $\widetilde{\mathbb{B}}_h$ w.r.t. value function $\{\widetilde{V}_{h+1}\}_{h=1}^H$. For any $x\in\mathcal{S}$ and n large enough, $\widetilde{\pi}=\{\widetilde{\pi}_h\}_{h=1}^H$,*

$$\mathrm{SubOpt}\left(\widetilde{\pi};x\right)\le 2\sum_{h=1}^H\mathbb{E}_{\pi^*}\left[\widetilde{\Gamma}_{h,\lambda'}(x,a)\mid x_1=x\right]$$

holds with probability at least $1-\xi$*. Here* $\mathbb{E}_{\pi^*}$ *is taken with respect to the trajectory induced by* π^* *in the underlying MDP given the fixed* Λ_h*.*

Proof. The proof is similar to that of Theorem 1, where Lemmas L.2 and L.3 are replaced by Lemmas M.14 and M.15, and ρ is replaced by λ'. Note that

$$\log\left|\mathcal{N}_h(\epsilon; R_0, B_0, J_r, J_p, n\lambda'/2, \lambda)\right| \leq 2(d_r+d_p)^2(\log c_p + 5\zeta)$$

still holds due to a similar argument in the proof of Theorem 1 as we choose $\lambda' \gg n^{-1/2}$. □

LLM USAGE STATEMENT

In the spirit of transparency and in accordance with the ICLR 2026 policy, we disclose that Large Language Models (LLMs) were utilized to aid in the preparation of this manuscript. The use of these models was limited to improving the quality of the writing. Specifically, LLMs were mainly employed for the following purposes:

- **Proofreading:** To identify and correct grammatical mistakes and typographical errors, thereby enhancing the clarity and readability of the text.
- **Notation Consistency Check:** To assist in reviewing the mathematical sections for potential issues, such as ensuring that all mathematical notations were defined before use and applied consistently throughout the paper.