Optimal Estimation of Policy Gradient via Double Fitted Iteration

Chengzhuo Ni [1] **Ruiqi Zhang** [2] **Xiang Ji** [1] **Xuezhou Zhang** [1] **Mengdi Wang** [1]

Abstract

Policy gradient (PG) estimation becomes a challenge when we are not allowed to sample with the target policy but only have access to a dataset generated by some unknown behavior policy. Conventional methods for off-policy PG estimation often suffer from either significant bias or exponentially large variance. In this paper, we propose the double Fitted PG estimation (FPG) algorithm. FPG can work with an arbitrary policy parameterization, assuming access to a Bellman-complete value function class. In the case of linear value function approximation, we provide a tight finite-sample upper bound on policy gradient estimation error, that is governed by the amount of distribution mismatch measured in feature space. We also establish the asymptotic normality of FPG estimation error with a precise covariance characterization, which is further shown to be statistically optimal with a matching Cramer-Rao lower bound. Empirically, we evaluate the performance of FPG on both policy gradient estimation and policy optimization, using either softmax tabular or ReLU policy networks. Under various metrics, our results show that FPG significantly outperforms existing off-policy PG estimation methods based on importance sampling and variance reduction techniques.

1. Introduction

Policy gradient plays a key role in policy-based reinforcement learning (RL). We focus on the estimation of policy gradient in off-policy reinforcement learning. In the off-policy setting, we are given episodic trajectories that were generated by some unknown behavior policy. Our goal is to estimate the *single* policy gradient of a target policy π_θ, i.e., $\nabla_\theta v^{\pi_\theta}$, based on the off-policy data only. This is motivated by applications such as medical diagnosis and ICU management, in which sampling data with a proposed policy is prohibitive or extremely costly. In these applications, one may not expect to learn the full optimal policy from limited data, but rather learn a single gradient vector for directions of improvement. To handle the distribution mismatch between behavior and target policy, a classic approach is importance sampling (IS) (Jie & Abbeel, 2010). However, IS is known to be sample-expensive and unstable, as the importance sampling weight can grow exponentially with respect to time horizon and causing uncontrollably large variances.

In this work, we design an algorithm to avoid the high variance of importance sampling by utilizing a *good* (to be defined in Sec. 4) value function approximation should they be available. The key idea is to perform PG estimation in an iterative way, similar to the well-known Fitted Q Iteration (FQI) algorithm. We propose the double Fitted Policy Gradient (FPG) estimation algorithm, which conducts iterative regression to estimate Q functions and $\nabla_\theta Q$ functions jointly. The FPG algorithm is able to provide an accurate estimation under mild data coverage assumption and without the knowledge of the behavior policy, in contrast to vanilla IS which must know the behavior policy.

When the function approximator is linear, we show that FPG is equivalent to a model-based plugin estimator and can give an ε-close PG estimator using a sample size of $N = O\left(CH^5/\varepsilon^2\right)$, where H is the horizon length and C is a constant to be specified that measures the distribution shift between behavior policy and target policy. Notably, this distribution shift $C = O(1+\chi^2_{\mathcal{F}}(\mu,\bar{\mu}))$ can be bounded by a form of relative condition number or a restricted chi-square divergence, measuring the mismatch between the behavior and target policy in feature space. We additionally establish the asymptotic normality of our FPG estimator with closed form variance expression. We also provide a matching information-theoretic Cramer-Rao lower bound, showing that our estimator is in fact asymptotically optimal. See Table 1 for a summary of theoretical results for off-policy PG estimation. FPG can be easily applied as a plug-in PG estimator in any off-policy PG algorithm. Under standard assumptions, a PG algorithm with FPG estimator can find an ε-stationary policy using at most $N = \tilde{O}\left(\dim(\Theta)^2/\varepsilon^2\right)$

[1]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA [2]School of Mathematical Science, Peking University, Beijing, China. Correspondence to: Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

samples. If the policy optimization landscape happens to satisfy the *Polyak-lojasiewicz condition* (Polyak, 1963; Bhandari & Russo, 2019), the sample complexity can be further improved to $N = \tilde{O}\left(\dim(\Theta)^2/\varepsilon\right)$ for finding an ε-optimal policy.

2. Problem Definitions

Markov Decision Process An instance of MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, p, r, \xi, H)$ where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $H \in \mathbb{N}_+$ is the horizon, $p_h : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}, h \in [H]$ is the transition probability (where $\Delta_{\mathcal{S}}$ denotes the probability simplex over $\mathcal{S}$), $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1], h \in [H]$ is the reward function and $\xi \in \Delta_{\mathcal{S}}$ is the initial state distribution. Given an MDP, a policy $\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}, h \in [H]$ is a distribution over the action space given the state s and time step h. At each time step h, the agent observes s_h and action a_h according to its behavior policy π. The agent then observes a reward $r_h(s_h, a_h)$ and the next state s_{h+1} sampled according to $s_{h+1} \sim p_h(\cdot|s_h, a_h)$. A policy π is measured by the Q function Q^π and the value v^π, defined by $Q_h^\pi(s, a) = \mathbb{E}^\pi[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})|s_h = s, a_h = a], \forall h \in [H], (s, a) \in \mathcal{S} \times \mathcal{A}$ and $v^\pi = \mathbb{E}^\pi[\sum_{h=1}^H r_h(s_h, a_h)|s_1 \sim \xi]$, where $\mathbb{E}^\pi$ denotes the expectation over trajectories by following policy π. The optimal policy of the MDP is defined as $\pi^* := \arg\max_\pi v^\pi$.

Off-Policy Policy Gradient Estimation Direct policy optimization methods are popular in RL due to their effectiveness and generalizability. Among them, the classic Policy Gradient (PG) method represents policies via a parametric function approximation and perform gradient ascent on the policy parameters (Sutton et al., 2000).

Denote a parametrized policy as π_θ, where $\theta \in \Theta$ is the policy parameters. Policy Gradient is defined as the gradient of policy value v_θ with respect to the policy parameter θ: $\nabla_\theta v_\theta = \nabla_\theta \mathbb{E}^{\pi_\theta}[\sum_{h=1}^H r_h(s_h, a_h)|s_1 \sim \xi]$. With policy gradients, one may directly search in the policy parameter space Θ using gradient ascent iterations, giving rise to the class of PG algorithms. However, directly differentiating through the value function is very difficult, especially when we do not have access to the transition probability of the MDP. The policy gradient theorem (Sutton et al., 2000) provides a convenient formula for estimating PG using Monte Carlo sampling:

$$\nabla_\theta v_\theta = \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^H \left(\sum_{h'=h}^H r_{h'}\right) \nabla_\theta \log \pi_{\theta,h}\left(a_h|s_h\right) \middle| s_1 \sim \xi\right].$$

In the online RL setting, one can interact with the environment directly with target policy π_θ and directly estimate the PG by averaging over sample trajectories (Degris et al., 2012; Kakade, 2001; Peters & Schaal, 2008; Sutton et al., 2000; Williams, 1992).

We focus on the more challenging offline RL setting, where we are not allowed to interact with the environment with the target policy π_θ. Instead, we only have access to offline logged data, $\mathcal{D} = \{(s_h^{(k)}, a_h^{(k)}, s_{h+1}^{(k)}, r_h^{(k)})\}_{h\in[H],k\in[K]}$, which consists of K *i.i.d.* trajectories, each of length H and is generated from an unknown behavior policy $\bar{\pi}$. The goal of off-policy PG estimation is to construct an estimator $\widehat{\nabla_\theta v_\theta}$ based solely on the off-policy data $\mathcal{D}$ that approximates the true gradient with low sample and computational complexity.

Notations Let π_θ be a policy parameterized by $\theta \in \Theta \subseteq \mathbb{R}^m$, where Θ is compact and $m = \dim(\theta)$. Let $\theta^* = \arg\max_{\theta\in\Theta} v_\theta$. Denote for short that $Q_h^\theta := Q_h^{\pi_\theta}, v_\theta := v^{\pi_\theta}$. Define the transition operator $\mathcal{P}_\theta$ by

$$\left(\mathcal{P}_{\theta,h} f\right)(s, a) := \mathbb{E}^{\pi_\theta}\left[f(s_{h+1}, a_{h+1})|s_h = s, a_h = a\right], \quad \forall f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, h \in [H], s \in \mathcal{S}, a \in \mathcal{A}.$$

where $[N]$ is the set of integer $1, 2, \ldots, N$. Given a real-valued function class $\mathcal{F}$ and vector-valued function $u : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$, we say $u \in \mathcal{F}$ if $u_j \in \mathcal{F}, \forall j \in [m]$. For any matrix $E \in \mathbb{R}^{d_1\times d_2}$ (which includes scalars and vectors as special cases), we define its Jacobian as $\nabla_\theta E_\theta = (\nabla_\theta^1 E_\theta, \nabla_\theta^2 E_\theta, \ldots, \nabla_\theta^m E_\theta) \in \mathbb{R}^{d_1\times md_2}$, where ∇_θ^j is the partial derivative w.r.t. the jth entry, i.e., $\nabla_\theta^j := \frac{\partial}{\partial\theta_j}$.

3. Related Work

When it comes to off-policy PG estimation, one demanding challenge is the distribution shift between the possibly unknown behavior policy and target policy (Agarwal et al., 2021). The basic Importance Sampling (IS) estimator for off-policy PG, which is still the most common approach used in practice, is

$$\widehat{\nabla_\theta v_\theta^{IS}} := \frac{1}{K}\sum_{k=1}^K w_k \sum_{h=1}^H \left(\sum_{h'=h}^H r_{h'}^{(k)}\right) \nabla_\theta \log \pi_{\theta,h}\left(a_h^{(k)} \middle| s_h^{(k)}\right),$$

where $w_k = \prod_{h=1}^H \frac{\pi_{\theta,h}\left(a_h^{(k)}\middle|s_h^{(k)}\right)}{\bar{\pi}_h\left(a_h^{(k)}\middle|s_h^{(k)}\right)}$ is the IS weight. Classical PG methods including REINFORCE and GPOMDP (Sutton et al., 2000; Williams, 1992) are all based on this idea or its modifications (Degris et al., 2012; Kakade, 2001; Peters & Schaal, 2008). A severe drawback of the IS method is its huge variance that can be as large as $2^{\Theta(H)}$, resulting in ill-behaved gradient steps in practice. IS also requires prior knowledge of $\bar{\pi}$ to compute the IS weights, which often is not available. (Kallus & Uehara, 2020) proposes a meta-algorithm called (EOOPG) that performs doubly robust off-policy PG estimation, assuming access to a number of nuisance estimators including Q function and density estimators. They show that if the state-action density ratio function $\mu^\pi/\bar{\mu}$ can be estimated with error rate $K^{-1/2}$, the EOOPG would be asymptotically efficient with a limit variance $\Theta(H^4/K)$. (Xu et al., 2021) extends the doubly

Algorithm	Variance	Require Known Behavior Policy?	Required Estimators	Finite Sample
REINFORCE (Kakade, 2001)	$2^{\Theta(H)}\Theta(\frac{1}{K})$	Yes	None	Yes
GPOMDP (Kakade, 2001; Shelton, 2013)	$2^{\Theta(H)}\Theta(\frac{1}{K})$	Yes	$\widehat{Q}_h^\theta$	Yes
EOOPG (Kallus & Uehara, 2020)	$\Theta(\frac{H^4}{K}\max_{s,a,h}\frac{\mu_h^\theta(s,a)}{\bar{\mu}_h(s,a)})$	Yes	$\widehat{\mu}_h^\theta, \widehat{\nabla_\theta \mu_h^\theta}, \widehat{Q}_h^\theta, \widehat{\nabla_\theta Q_h^\theta}$	No
FPG (This Paper, with d-dim Features)	$\Theta(\frac{H^4}{K}\min\{C_1 d, H\}(1+\chi^2_{\mathcal{F}}(\mu^\theta,\bar{\mu})))$	No	None	Yes

Table 1. **Comparison of Off-Policy PG Estimation Methods.** Both REINFORCE and GPOMDP suffers from exponential variance in the worst case. EOOPG's bound scales with the maximum density ratio $\max_{s,a,h}\frac{\mu_h^\theta(s,a)}{\bar{\mu}_h(s,a)}$), whereas our bound only scales with the χ^2-divergence which can be much smaller. For example, the density ratio between two standard Gaussian distribution is infinite, but the χ^2-divergence is always bounded. In addition, our method also does not require the knowledge of the behavior policy or assume access to any high-performing value or gradient estimators, which may not be available in practice.

robust approach to the case of discounted MDP and with a finite sample guarantee. However, both work require the density ratio be precisely estimated, which is arguably an even harder problem. Note that density ratio estimation requires learning a function that maps from the raw state space, which can be arbitrarily high dimensional and complex, whereas policy gradient estimation only requires estimating a vector of length being the number of policy parameters. They did not provide a guarantee on the error of such an estimator and leave the estimation error in the final result as a irreducible term. (Morimura et al., 2010) proposes a temporal difference method and estimates the policy gradient via linear function approximation of the stationary state distribution, but does not provide a formal statistical guarantee. Several other methods for off-policy PG, including Non-parametric OPPG (Tosatto et al., 2020) and Q-Prop (Gu et al., 2016), are found to be empirically effective but no theoretical guarantee is provided. In general, theoretical understanding for off-policy PG remains rather limited. We summarize known variance bounds for off-policy PG estimation in Table 1.

Off-policy PG estimation is closely related to PG-based policy optimization. For example, even in online policy optimization, one can use past data for more efficient PG estimation. Several works (Papini et al., 2018; Xu et al., 2020; 2019) combines IS with variance reduction technique, but their theories are based on the assumption that the variance of the IS estimator is bounded at some controllable level instead of grow exponentially (Jiang & Li, 2016; Degris et al., 2012; Kallus & Uehara, 2020) or that Lipschitz continuity holds (Zhang et al., 2021a). (Tosatto et al., 2020) provides a non-parametric OPPG method with some error analysis. (Liu et al., 2019; Gu et al., 2016) combine off policy PG estimation with actor-critic/policy gradient schemes. (Zhang et al., 2020) generalizes the notion of policy gradient to RL with general utilities and shows that such PG can be estimated by solving a stochastic saddle point.

Another closely related topic is the Offline Policy Evaluation (OPE), i.e., to estimate the target policy's value given offline data generated by some behavior policy $\bar{\pi}$. Various methods, from importance sampling to doubly robust estimators have been proposed (Tokdar & Kass, 2010; Precup et al., 2000; Jiang & Li, 2016; Thomas & Brunskill, 2016). A marginalized importance sampling (Xie et al., 2019) for tabular MDP and a fitted Q evaluation (Duan et al., 2020) approach for linear MDP provably achieve minimax-optimal error bound with matching information-theoretic lower bounds. The fitted Q evaluation method was later shown to work with bootstrapping (Hao et al., 2021), kernel function approximation (Duan et al., 2021), third-order differentiable function approximation (Zhang et al., 2022) and ReLU networks (Ji et al., 2022).

Another line of works use the pessimism principle to design algorithms that can perform stable estimation even under weaker coverage assumption (Jin et al., 2020; Rashidinejad et al., 2021; Zanette et al., 2021; Zhang et al., 2021b; Chang et al., 2021; Yin et al., 2022). However, to the best of our knowledge, these algorithms do not achieve the minimax optimal rate for OPE and it's unclear how to apply them to PG estimation.

4. Assumptions

In this paper, we focus on a setting where Q^θ and $\nabla_\theta Q^\theta$ can both be represented within a function class $\mathcal{F}$. Assume without loss of generality that $\mathbf{1} \in \mathcal{F}$.

Assumption 4.1 (Bellman Completeness)**.** For any $f \in \mathcal{F}$ and $h \in [H]$, we have $\mathcal{P}_{\theta,h} f \in \mathcal{F}$, and we suppose $r_h \in \mathcal{F}, \forall h \in [H]$. It follows that $Q_h^\theta \in \mathcal{F}, \forall h \in [H], \theta \in \Theta$.

The *Bellman Completeness* assumption has been commonly made in the theoretical offline RL literature (Xie et al., 2021; Duan et al., 2020). It requires $\mathcal{F}$ to be closed under the transition operator $\mathcal{P}_\theta$, so that the function approximation incurs zero Bellman error. In fact, it is known both theoretically (Wang et al., 2020) and empirically (Wang et al., 2021) that without such assumption FQI can diverge. We similarly assume that the gradient map also belongs to $\mathcal{F}$.

Assumption 4.2. $\nabla_\theta Q_h^\theta \in \mathcal{F}, \forall h \in [H], \theta \in \Theta$.

In the theoretical results, we will focus on the tractable case where $\mathcal{F}$ is a linear function class, since even OPE with general nonlinear function class remains an open problem. However, we remark that our algorithm (see Alg. 1) applies to any function class, including neural networks.

Linear function approximation Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a state-action feature map. Let $\mathcal{F}$ be the class of linear functions given by $\mathcal{F} = \{\phi(\cdot,\cdot)^\top w | w \in \mathbb{R}^d\}$. Then for any policy π_θ and $h \in [H]$, Assumption 4.1 implies there exist $w_r \in \mathbb{R}^d$ and $w_h^\theta \in \mathbb{R}^d$ such that

$$r_h(s,a) = \phi(s,a)^\top w_{r,h}, \quad Q_h^\theta(s,a) = \phi(s,a)^\top w_h^\theta.$$

Furthermore, we show that Assumption 4.1 alone is sufficient to ensure the expressiveness of $\mathcal{F}$ for PG estimation in case of the linear function class.

Proposition 4.3. *If $\mathcal{F} = \{\phi(\cdot,\cdot)^\top w | w \in \mathbb{R}^d\}$, Assumption 4.1 implies Assumption 4.2. In particular, we have w_h^θ is differentiable w.r.t. θ and*

$$\nabla_\theta Q_h^\theta(s,a) = \phi(s,a)^\top \nabla_\theta w_h^\theta, \quad \forall h \in [H].$$

In other word, as long as one can use linear function approximation for policy evaluation, the same feature map automatically allows linear function approximation of $\nabla_\theta Q_h^\theta$.

5. Algorithm

In this section, we describe our double Fitted Policy Gradient iteration (FPG) algorithm, designed to estimate the policy gradient $\nabla_\theta v_\theta$ from an arbitrary batch data $\mathcal{D}$.

5.1. Policy Gradient Bellman Equation

Notice that by Bellman's equation, we have

$$Q_h^\theta(s,a) = r_h(s,a) + \int_{\mathcal{S}\times\mathcal{A}} p_h(s'|s,a)\pi_{\theta,h+1}(a'|s') Q_{h+1}^\theta(s',a')\mathrm{d}s'\mathrm{d}a'.$$

Differentiating on both sides w.r.t. θ, we get

$$\begin{aligned}
&\nabla_\theta Q_h^\theta(s,a) \\
=& \int_{\mathcal{S}\times\mathcal{A}} p_h(s'|s,a)\left(\nabla_\theta \pi_{\theta,h+1}(a'|s')\right) Q_{h+1}^\theta(s',a')\mathrm{d}s'\mathrm{d}a' \\
&+ \int_{\mathcal{S}\times\mathcal{A}} p_h(s'|s,a)\pi_{\theta,h+1}(a'|s')\nabla_\theta Q_{h+1}^\theta(s',a')\mathrm{d}s'\mathrm{d}a' \\
=& \mathbb{E}^{\pi_\theta}\left[\left(\nabla_\theta \log \pi_{\theta,h+1}(a_{h+1}|s_{h+1})\right) Q_{h+1}^\theta(s_{h+1},a_{h+1}) \right.\\
&\left. + \nabla_\theta Q_{h+1}^\theta(s_{h+1},a_{h+1}) | s_h = s, a_h = a\right].
\end{aligned}$$

Here we use the convention that the gradient of $\nabla_\theta Q_h^\theta$ or $\nabla_\theta \pi_{\theta,h}$ is a function from $\mathcal{S}\times\mathcal{A}$ to a row vector in $\mathbb{R}^{1\times m}$. Thus, we get the *Policy Gradient Bellman equation*, given by

$$\begin{aligned}
Q_h^\theta &= r_h + \mathcal{P}_{\theta,h} Q_{h+1}^\theta \\
\nabla_\theta Q_h^\theta &= \mathcal{P}_{\theta,h}\left(\left(\nabla_\theta \log \Pi_{\theta,h+1}\right) Q_{h+1}^\theta + \nabla_\theta Q_{h+1}^\theta\right),
\end{aligned} \tag{1}$$

where we define the operator $\nabla_\theta \log \Pi_{\theta,h}$ by

$$\left(\left(\nabla_\theta \log \Pi_{\theta,h}\right) f\right)(s,a) = \left(\nabla_\theta \log \pi_{\theta,h}(a|s)\right) f(s,a).$$

Once we get the estimations of Q_1^θ and $\nabla_\theta Q_1^\theta$, we can calculate the policy gradient $\nabla_\theta v_\theta$ using the formula: $\nabla_\theta v_\theta = \int_{\mathcal{S}\times\mathcal{A}} \xi(s)\pi_{\theta,1}(a|s)\big(\nabla_\theta Q_1^\theta(s,a) + \left(\nabla_\theta \log \pi_{\theta,1}(a|s)\right) Q_1^\theta(s,a)\big)\mathrm{d}s\mathrm{d}a$.

5.2. Double Fitted Policy Gradient Iteration

In a similar spirit to Fitted Q Iteration (FQI), we develop our PG estimator based on the gradient Bellman equations (1). We derive our estimator by applying regression iteratively: Let $\widehat{Q}_{H+1}^{\theta,\text{FPG}} = \widehat{\nabla_\theta^j Q_{H+1}^{\theta,\text{FPG}}} = 0, \forall j \in [m]$. For $h = H, H-1, \ldots, 1$ and $j \in [m]$, let

$$\begin{aligned}
\widehat{Q}_h^{\theta,\text{FPG}} = \arg\min_{f\in\mathcal{F}} \Bigg[\lambda\rho(f) + \sum_{k=1}^K \Bigg(f\left(s_h^{(k)}, a_h^{(k)}\right) - r_h^{(k)} \\
- \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \,\middle|\, s_{h+1}^{(k)}\right) \widehat{Q}_{h+1}^{\theta,\text{FPG}}\left(s_{h+1}^{(k)}, a'\right)\mathrm{d}a'\Bigg)^2\Bigg]
\end{aligned} \tag{2}$$

$$\begin{aligned}
\widehat{\nabla_\theta^j Q_h^{\theta,\text{FPG}}} = \arg\min_{f\in\mathcal{F}} \Bigg[\lambda\rho(f) + \sum_{k=1}^K \Bigg(f\left(s_h^{(k)}, a_h^{(k)}\right) \\
- \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \,\middle|\, s_{h+1}^{(k)}\right)\Bigg(\widehat{\nabla_\theta^j Q_{h+1}^{\theta,\text{FPG}}}\left(s_{h+1}^{(k)}, a'\right) \\
+ \widehat{Q}_{h+1}^{\theta,\text{FPG}}\left(s_{h+1}^{(k)}, a'\right)\nabla_\theta^j \log \pi_{\theta,h+1}\left(a' \,\middle|\, s_{h+1}^{(k)}\right)\Bigg)\mathrm{d}a'\Bigg)^2\Bigg]
\end{aligned} \tag{3}$$

After the computation of $\widehat{Q}_h^{\theta,\text{FPG}}, \widehat{\nabla_\theta Q_h^{\theta,\text{FPG}}}$, the policy gradient can be estimated straightforwardly. The full algorithm is summarized in Algorithm 1.

5.3. Equivalence to a Model-based Plug-in Estimator

Next we show that the FPG estimator is equivalent to a model-based plugin estimator. Define the model-based re-

Algorithm 1 Fitted PG Algorithm

1: **Input:** Dataset $\mathcal{D}$, target policy π_θ, initial state distribution ξ.
2: **Initialize** $\widehat{Q_{H+1}^{\theta,\mathrm{FPG}}} = 0$ and $\nabla_\theta^j Q_{H+1}^{\theta,\mathrm{FPG}} = 0,\ \forall j \in [m]$.
3: **for** $h = H, H-1, \ldots, 1$ **do**
4: Calculate $\widehat{Q}_h^{\theta,\mathrm{FPG}}, \widehat{\nabla_\theta Q_h^{\theta,\mathrm{FPG}}}$ by solving (2) and (3).
5: **end for**
6: **Return**

$$\widehat{\nabla_\theta v_\theta^{\mathrm{FPG}}} = \int_{\mathcal{S}\times\mathcal{A}} \xi(s)\pi_{\theta,1}(a|s)\Big(\widehat{\nabla_\theta Q_1^{\theta,\mathrm{FPG}}}(s,a) + \widehat{Q}_1^{\theta,\mathrm{FPG}}(s,a)\nabla_\theta \log \pi_{\theta,1}(a|s)\Big)\mathrm{d}s\mathrm{d}a.$$

ward estimate $\widehat{r}$ and transition operator estimate $\widehat{\mathcal{P}}_\theta$ as followings: for any (possibly vector-valued) function f on $\mathcal{S}\times\mathcal{A}$ and $h \in [H]$,

$$\widehat{r}_h := \arg\min_{f'\in\mathcal{F}} \left[\sum_{k=1}^K \left(f'\left(s_h^{(k)}, a_h^{(k)}\right) - r_h^{(k)}\right)^2 + \lambda\rho(f')\right],$$
$$\widehat{\mathcal{P}}_{\theta,h} f := \arg\min_{f'\in\mathcal{F}} \left[\sum_{k=1}^K \left(f'\left(s_h^{(k)}, a_h^{(k)}\right) - \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \,\middle|\, s_h^{(k)}\right) f\left(s_{h+1}^{(k)}, a'\right)\mathrm{d}a'\right)^2 + \lambda\rho(f')\right].$$

Plugging $\widehat{\mathcal{P}}_\theta$ and $\widehat{r}$ into (1), we may calculate the policy gradient associated with the estimated model. Let $\widehat{Q}_{H+1}^{\theta,\mathrm{MB}} = \widehat{\nabla_\theta^j Q_{H+1}^{\theta,\mathrm{MB}}} = 0, j \in [m]$. For $h = H, H-1, \ldots, 1$,

$$\widehat{Q}_h^{\theta,\mathrm{MB}} = \widehat{r}_h + \widehat{\mathcal{P}}_{\theta,h}\widehat{Q}_{h+1}^{\theta,\mathrm{MB}},$$
$$\widehat{\nabla_\theta^j Q_h^{\theta,\mathrm{MB}}} = \widehat{\mathcal{P}}_{\theta,h}\left(\left(\nabla_\theta^j \log \Pi_{\theta,h+1}\right)\widehat{Q}_{h+1}^{\theta,\mathrm{MB}} + \widehat{\nabla_\theta^j Q_{h+1}^{\theta,\mathrm{MB}}}\right).$$

Then the model-based gradient estimator is

$$\widehat{\nabla_\theta v_\theta^{\mathrm{MB}}} = \int_{\mathcal{S}\times\mathcal{A}} \xi(s)\pi_{\theta,1}(a|s)\Big(\widehat{\nabla_\theta Q_1^{\theta,\mathrm{MB}}}(s,a) + \widehat{Q}_1^{\theta,\mathrm{MB}}(s,a)\nabla_\theta \log \pi_{\theta,1}(a|s)\Big)\mathrm{d}s\mathrm{d}a.$$

Note that the model-based plug-in approach makes intuitive sense, but is intractable to implement.

Remarkably, we show that the model-based plug-in estimator $\widehat{\nabla_\theta v_\theta^{\mathrm{MB}}}$ is essentially equivalent to the fitted PG estimator, when $\mathcal{F}$ is the class of linear functions.

Proposition 5.1. *When $\mathcal{F} = \{\phi(\cdot,\cdot)^\top w | w \in \mathbb{R}^d\}$ and the regulator ρ is chosen to be $\rho(\phi^\top w) = \|w\|^2$, we have*

- $\widehat{Q}_h^{\theta,\mathrm{FPG}} = \widehat{Q}_h^{\theta,\mathrm{MB}}, \widehat{\nabla_\theta Q_h^{\theta,\mathrm{FPG}}} = \widehat{\nabla_\theta Q_h^{\theta,\mathrm{MB}}}, \forall h \in [H]$;
- $\widehat{\nabla_\theta v_\theta^{\mathrm{FPG}}} = \widehat{\nabla_\theta v_\theta^{\mathrm{MB}}}$.

In the remainder, we focus on linear $\mathcal{F}$ and let $\rho(\phi^\top w) = \|w\|^2$. We will omit the superscript FPG and MB , and simply denote $\widehat{Q}_h^\theta, \widehat{\nabla_\theta Q_h^\theta}, \widehat{\nabla_\theta v_\theta}$ as our estimators.

5.4. FPG with Linear Function Approximation

Define the empirical covariance matrix: For $h \in [H]$,

$$\widehat{\Sigma}_h = \frac{1}{K}\left(\lambda I_d + \sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\phi\left(s_h^{(k)}, a_h^{(k)}\right)^\top\right),$$

where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix. In the case of linear function class, one could write down the expression of $\widehat{r}$ and $\widehat{\mathcal{P}}_\theta$ explicitly:

$$\widehat{r}_h(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right) r_h^{(k)} =: \phi(\cdot,\cdot)^\top \widehat{w}_{r,h}, \tag{4}$$

$$\left(\widehat{\mathcal{P}}_{\theta,h} f\right)(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K \left[\phi\left(s_h^{(k)}, a_h^{(k)}\right) \cdot \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \,\middle|\, s_{h+1}^{(k)}\right) f\left(s_{h+1}^{(k)}, a'\right)\mathrm{d}a'\right].$$

For $f(\cdot,\cdot) = \phi(\cdot,\cdot)^\top w \in \mathcal{F}$, the above become concise closed forms:

$$\left(\widehat{\mathcal{P}}_{\theta,h} f\right)(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{M}_{\theta,h} w,$$
$$\left(\widehat{\mathcal{P}}_{\theta,h}\left(\nabla_\theta \log \Pi_{\theta,h+1}\right) f\right)(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{\nabla_\theta M_{\theta,h}}\left(I_m \otimes w\right).$$

where the notation $\otimes$ is used to denote the Kronecker product between two matrices, $\widehat{M_{\theta,h}} \in \mathbb{R}^{d\times d}, \widehat{\nabla_\theta M_{\theta,h}} \in \mathbb{R}^{d\times md}$ are defined by

$$\widehat{M}_{\theta,h} := \widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K \left[\phi\left(s_h^{(k)}, a_h^{(k)}\right) \cdot \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \,\middle|\, s_{h+1}^{(k)}\right)\phi\left(s_{h+1}^{(k)}, a'\right)^\top \mathrm{d}a'\right], \tag{5}$$

$$\widehat{\nabla_\theta M_{\theta,h}} := \nabla_\theta \widehat{M}_{\theta,h} = \widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K \left[\phi\left(s_h^{(k)}, a_h^{(k)}\right) \cdot \int_{\mathcal{A}} \phi\left(s_{h+1}^{(k)}, a'\right)^\top \left(\nabla_\theta \pi_{\theta,h+1}\left(a' \,\middle|\, s_{h+1}^{(k)}\right) \otimes I_d\right)\mathrm{d}a'\right]. \tag{6}$$

In this way, one can easily compute $\widehat{Q}_h^\theta$ and $\widehat{\nabla_\theta Q_h^\theta}$ in a matrix recursive form, which we illustrate in Algorithm 2.

Runtime Complexity Algorithm 2 is computationally very efficient. Suppose that caculating integral against action distribution takes time $O(1)$. In Algorithm 2, the calculation of $\widehat{w}_{r,h}, \widehat{M}_{\theta,h}$ and $\widehat{\nabla_\theta M_{\theta,h}}$ require at most $O(KHmd^2)$ numeric operations. The recursive function fitting steps at line 3-5 require at most $O(Hmd^2)$ numeric operations. Thus the total runtime is only $O(KHmd^2)$.

Algorithm 2 FPG Estimation with Linear Approximation

Input: Dataset $\mathcal{D}$, target policy π_θ, initial state distribution ξ.
Calculate $\widehat{w}_{r,h}, \widehat{M}_{\theta,h}, \widehat{\nabla_\theta M_{\theta,h}}, h \in [H]$ according to (4), (5), (6)
Let $\widehat{w}^\theta_{H+1} = \mathbf{0}_d$ and $\widehat{W}^\theta_{H+1} = \mathbf{0}_{d\times m}$.
for $h = H, H-1, \ldots, 1$ **do**
 Set $\widehat{w}^\theta_h = \widehat{w}_{r,h} + \widehat{M}_{\theta,h}\widehat{w}^\theta_{h+1}$, $\widehat{W}^\theta_h = \widehat{\nabla_\theta M_{\theta,h}}(I_m \otimes \widehat{w}^\theta_{h+1}) + \widehat{M_{\theta,h}}\widehat{W}^\theta_{h+1}$.
end for
Return $\widehat{\nabla_\theta v_\theta} = \int_{\mathcal{S}\times\mathcal{A}} \xi(s)\pi_{\theta,1}(a|s)\phi(s,a)^\top \big(\widehat{W}^\theta_1 + \widehat{w}^\theta_1 \nabla_\theta \log \pi_{\theta,1}(a|s)\big)\mathrm{d}s\mathrm{d}a$.

6. Main Results

In this section we study the statistical properties of the FPG estimator with linear function approximation. Define the population covariance matrix as $\Sigma_h := \mathbb{E}\left[\phi\left(s_h^{(1)}, a_h^{(1)}\right)\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top\right]$, $h \in [H]$, where $\mathbb{E}$ represents the expectation over the data generating distribution by the behavior policy.

Assumption 6.1 (Boundedness Conditions)**.** Assume for any $h \in [H]$, Σ_h is invertible. There exist absolute constants C_1, G such that for any $h \in [H]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}, j \in [m]$, we have $\phi(s,a)^\top \Sigma_h^{-1}\phi(s,a) \le C_1 d$ and $\left|\nabla^j_\theta \log \pi_{\theta,h}(a|s)\right| \le G$.

Assumption 6.1 requires the data generating distribution to have a full-rank covariance matrix, effectively covering all d directions in the feature space. Note that this is a much weaker condition compared to the uniform coverage condition ($\max_{s,a,h} \frac{\mu_h^\theta(s,a)}{\bar{\mu}_h(s,a)}) < \infty$) made in prior works (Kallus & Uehara, 2020), which requires coverage on all (s,a) pairs. Define $\nu_h^\theta := \mathbb{E}^{\pi_\theta}[\phi(s_h,a_h)|s_1 \sim \xi]$ and $\Sigma_{\theta,h} := \mathbb{E}^{\pi_\theta}[\phi(s_h,a_h)\phi(s_h,a_h)^\top|s_1 \sim \xi]$.

6.1. Finite-Sample Variance-Aware Error Bound

Let us first consider finite-sample analysis of our estimator. We present a variance-aware error bound. Denote $\phi_{\theta,h}(s) := \mathbb{E}^{\pi_\theta}[\phi(s',a')|s' = s]$, $\varepsilon^\theta_{h,k} := Q_h^\theta\left(s_h^{(k)}, a_h^{(k)}\right) - r_h^{(k)} - \mathbb{E}^{\pi_\theta}[Q^\theta_{h+1}(s',a')\,|s' = s^{(k)}_{h+1}]$, and $\Lambda_\theta = \sum_{h=1}^H \mathrm{Cov}\left[\nabla_\theta\left(\varepsilon^\theta_{h,1}\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1}\nu_h^\theta\right)\right]$.

Theorem 6.2 (Finite Sample Guarantee)**.** *For any $t \in \mathbb{R}^m$, when $K \ge 36\kappa_1(4+\kappa_2+\kappa_3)^2 C_1 d H^2 \log\frac{8dmH}{\delta}$ and $\lambda \le C_1 d \min_{h\in[H]} \sigma_{\min}(\Sigma_h)\log\frac{8dmH}{\delta}$, with probability $1-\delta$, we have,*

$$|\langle t, \widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta\rangle| \le \sqrt{\frac{2t^\top\Lambda_\theta t}{K}\cdot\log\frac{8}{\delta}} + \frac{C_\theta\|t\|\log\frac{72mdH}{\delta}}{K},$$

where $C_\theta = 240C_1 d\sqrt{m}H^3\kappa_1(5+\kappa_2+\kappa_3)(\max_{j\in[m]}\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla^j_\theta\nu_1^\theta\| + HG\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\|)$ *and*

$$\kappa_1 = \max_{h\in[H]}\frac{\sigma_{\max}\left(\Sigma_h^{-\frac{1}{2}}\Sigma_{\theta,h}\Sigma_h^{-\frac{1}{2}}\right)}{\sigma_{\min}\left(\Sigma_{h+1}^{-\frac{1}{2}}\Sigma_{\theta,h+1}\Sigma_{h+1}^{-\frac{1}{2}}\right)\wedge 1},$$

$$\kappa_2 = \max_{h\in[H]}\left\|\Sigma_{h+1}^{-\frac{1}{2}}\mathbb{E}\left[\phi_{\theta,h+1}\left(s^{(1)}_{h+1}\right)\phi_{\theta,h+1}\left(s^{(1)}_{h+1}\right)^\top\right]\Sigma_{h+1}^{-\frac{1}{2}}\right\|^{\frac{1}{2}},$$

$$\kappa_3 = \frac{1}{G}\max_{j\in[m],h\in[H]}$$
$$\left\|\Sigma_{h+1}^{-\frac{1}{2}}\mathbb{E}\left[\left(\nabla^j_\theta\phi_{\theta,h+1}\left(s^{(1)}_{h+1}\right)\right)\left(\nabla^j_\theta\phi_{\theta,h+1}\left(s^{(1)}_{h+1}\right)\right)^\top\right]\Sigma_{h+1}^{-\frac{1}{2}}\right\|^{\frac{1}{2}}.$$

Theorem 6.2 shows that the finite-sample FPG error is largely determined by $\sqrt{\frac{t^T\Lambda_\theta t}{K}}$. Here Λ_θ gives a precise characterization of the error's covariance.

6.2. Worst-Case Error Bound and Distribution Shift

Next we derive a worst-case error bound that depends only on the distribution shift but not on reward/variance properties. The following theorem provides a worst-case guarantee under arbitrary choice of the reward function.

Theorem 6.3 (Finite Sample Guarantee - Reward Free)**.** *Let the conditions in Theorem 6.2 hold, with probability $1-\delta$, we have $\forall j \in [m]$,*

$$\left|\widehat{\nabla^j_\theta v_\theta} - \nabla^j_\theta v_\theta\right| \le 4b_\theta\sqrt{\frac{\min\{C_1 d, H\}\log\frac{8m}{\delta}}{K}} + \frac{C_\theta\log\frac{72mdH}{\delta}}{K},$$

where $b_\theta = H^2 G\max_{h\in[H]}\|\Sigma_h^{-\frac{1}{2}}\nu_h^\theta\| + H\max_{h\in[H]}\|\Sigma_h^{-\frac{1}{2}}\nabla^j_\theta\nu_h^\theta\|$ *and C_θ is the same as that in Theorem 6.2. If we in addition have $\phi(s',a')^\top\Sigma_h^{-1}\phi(s,a) \ge 0, \forall (s,a),(s',a') \in \mathcal{S}\times\mathcal{A}, h \in [H]$, we have*

$$\begin{aligned}&\left|\widehat{\nabla^j_\theta v_\theta} - \nabla^j_\theta v_\theta\right| \\ \le\,& 4H^2 G\sqrt{\frac{\min\{C_1 d, H\}\log\frac{8m}{\delta}}{K}}\max_{h\in[H]}\left\|\Sigma_h^{-\frac{1}{2}}\nu_h^\theta\right\| \\ &+ \frac{2C_\theta\log\frac{72mdH}{\delta}}{K}, \quad \forall j \in [m].\end{aligned}$$

The complete proofs of Theorem 6.2 and Theorem 6.3 are deferred to Appendix B.1 and B.2. To further simplify the expression in Theorem 6.3, we define a variant of χ^2-divergence restricted to the family $\mathcal{F}$: for any two groups of probability distributions $p_1 = \{p_{1,h}\}_{h=1}^H, p_2 = \{p_{2,h}\}_{h=1}^H$, define

$$\begin{aligned}\chi^2_{\mathcal{F}}(p_{1,h}, p_2) :&= \max_{h\in[H]} \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_{p_{1,h}}[f(x)]^2}{\mathbb{E}_{p_{2,h}}[f(x)^2]} - 1\\ &= \max_{h\in[H]} \nu_{p_{1,h}}^\top \Sigma_{p_{2,h}}^{-1} \nu_{p_{1,h}} - 1,\end{aligned}$$

where $\nu_p = \mathbb{E}_p[\phi(s,a)],\ \Sigma_p = \mathbb{E}_p[\phi(s,a)\phi(s,a)^\top]$. Let $\bar{\mu} = \{\bar{\mu}_h\}_{h=1}^H$ be the occupancy distribution of observation $(s_h^{(1)}, a_h^{(1)})$. Let $\mu^\theta = \{\mu_h^\theta\}_{h=1}^H$ be the occupancy distribution of (s_h, a_h) under policy π_θ. When we have $\phi(s',a')^\top \Sigma_h^{-1}\phi(s,a) \geq 0,\ \forall(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}, h \in [H]$, the result of Theorem 6.3 implies $\forall j \in [m]$,

$$\begin{aligned}&|\widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta|\\ \leq& 4H^2G\sqrt{\frac{\min\{C_1 d, H\}\log\frac{8m}{\delta}}{K}(1+\chi^2_{\mathcal{F}}(\mu^\theta,\bar{\mu}))} + \tilde{O}\left(\frac{1}{K}\right).\end{aligned}$$

The result of Theorem 6.3 matches the asymptotic bound provided in (Kallus & Uehara, 2020), but holds in finite sample regime and requires less stringent conditions.

The case of tabular MDP. In the tabular case, the condition $\phi(s,a)^\top\Sigma_h^{-1}\phi(s',a') \geq 0$ automatically holds. Furthermore, we have the following simplified guarantee:

Theorem 6.4 (Upper bound in tabular case). *In the tabular case with $\mathcal{F} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, if K is sufficiently large and $\lambda = 0$, then with probability at least $1-\delta$, $\forall j \in [m]$*

$$\begin{aligned}&|\widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta| \leq \tilde{O}\left(\frac{1}{K}\right) + 4H^2G\sqrt{\frac{\log\frac{8m}{\delta}}{K}}\\ &\cdot\sqrt{\min\left\{\max_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}\frac{\mu_h^\theta(s,a)}{\bar{\mu}_h(s,a)}, C_1 d\max_{h\in[H]}\mathbb{E}^{\pi_\theta}\left[\frac{\mu_h^\theta(s_h,a_h)}{\bar{\mu}_h(s_h,a_h)}\right]\right\}}\end{aligned}$$

6.3. Asymptotic Normality and Cramer-Rao Lower Bound

Next we show that FPG is an asymptotically normal and efficient estimator.

Theorem 6.5 (Asymptotic Normality). *The FPG estimator given by Algorithm 2 is asymptotically normal:*

$$\sqrt{K}\left(\widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta\right) \xrightarrow{d} \mathcal{N}(0, \Lambda_\theta).$$

An obvious corollary of Theorem 6.5 is that for any $t \in \mathbb{R}^m$,

$$\sqrt{K}\left\langle t, \widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta\right\rangle \xrightarrow{d} \mathcal{N}\left(0, t^\top\Lambda_\theta t\right).$$

An asymptotically efficient estimator has the minimal variance among all the unbiased estimators. The next theorem states the Cramer-Rao lower bound for PG estimation.

Theorem 6.6 (Cramer-Rao Lower Bound). *Let Assumption 4.1 hold. For any vector $t \in \mathbb{R}^m$, the variance of any unbiased estimator for $\langle t, \nabla_\theta v_\theta\rangle$ is lower bounded by $\frac{1}{K}t^\top\Lambda_\theta t$.*

The proofs of Theorem 6.5 and 6.6 are deferred to Appendix B.4, B.5. Theorem 6.6 along with Theorem 6.5 show that the FQI estimator is statistically optimal.

6.4. FPG for Policy Optimization

Lastly we briefly consider the use of FPG for off-policy policy optimization. Assume in the ideal setting we can reliably estimate the PG for all policies, obtaining $\widehat{\nabla_\theta v_\theta}$ for all $\theta \in \Theta$. Then we can simply set $\widehat{\nabla_\theta v_\theta} = 0$, identify all the stationary solutions, and pick the best one. For MDP with Lipschitz continuous policy gradients, we show that a policy with $\widehat{\nabla_\theta v_\theta} = 0$ would be nearly stationary/optimal.

Assumption 6.7. Suppose the parameter space Θ is bounded and the policy gradient is L-Lipschitz continuous and $\chi^2_{\mathcal{F}}$ is L'-Lipschitz continuous, i.e.,

$$\begin{aligned}\|\nabla_{\theta_1}v_{\theta_1} - \nabla_{\theta_2}v_{\theta_2}\| \leq L\|\theta_1-\theta_2\|, \quad \forall\theta_1,\theta_2\in\Theta,\\ \left|\chi^2_{\mathcal{F}}\left(\mu^{\theta_1},\bar{\mu}\right) - \chi^2_{\mathcal{F}}\left(\mu^{\theta_2},\bar{\mu}\right)\right| \leq L'\|\theta_1-\theta_2\|, \quad \forall\theta_1,\theta_2\in\Theta.\end{aligned}$$

Proposition 6.8. *Suppose assumption 6.7 and the condition of Theorem 6.3 hold. When K is sufficiently large, we have with probability at least $1-\delta$,*

$$\begin{aligned}\|\nabla_\theta v_\theta - \widehat{\nabla_\theta v_\theta}\| \leq 64H^2Gm\sqrt{\min\{C_1 d, H\}}\\ \cdot\sqrt{1+\chi^2_{\mathcal{F}}(\mu^\theta,\bar{\mu})}\sqrt{\frac{\log\frac{24DKLL'}{\delta HG}}{K}}, \quad \forall\theta\in\Theta.\end{aligned}$$

where D is the diameter of Θ. In addition, if the Polyak-Łojasiewicz condition holds, i.e., there exists a constant $c > 0$ such that for any $\theta\in\Theta$, $\frac{1}{2}\|\nabla_\theta v_\theta\|^2 \geq c(v_{\theta^*} - v_\theta)$, then for any $\widehat{\theta}$ such that $\widehat{\nabla_\theta v_{\widehat{\theta}}} = 0$, we have $v_{\theta^*} - v_{\widehat{\theta}} \leq \tilde{O}\left(\frac{m^2H^4\min\{C_1 d, H\}G^2}{K}\right)$.

In general, Proposition 6.8 implies a $O(1/\varepsilon^2)$ sample complexity for finding ε-stationary policies. This off-policy sample efficiency is remarkably better than the best know $O(1/\varepsilon^3)$ on-policy sample efficiency obtained by variance-reduced PG algorithm (Zhang et al., 2021a), as long as distribution shift is uniformly bounded. This improvement is due to that FPG makes full usage of data to evaluate PG at every θ. We remark that the discussion in this section is more of a stylish observation than a practically sound algorithm. How to incorporate FPG into policy gradient algorithms is an important future direction.

7. Experiments

We empirically evaluate the performance of FPG using the OpenAI gym FrozenLake and CliffWalking environment.

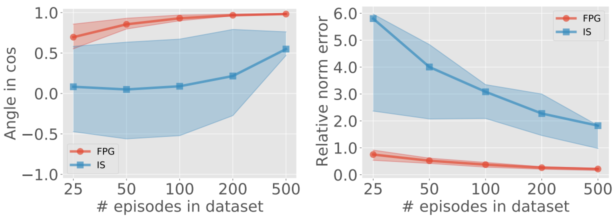

Figure 1. **Sample efficiency of FPG on off-policy data.** The off-policy PG estimation accuracy is evaluated using two metrics: $\cos\angle(\widehat{\nabla_\theta v_\theta}, \nabla_\theta v_\theta)$ and the relative error norm $\frac{\|\widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta\|}{\|\nabla_\theta v_\theta\|}$.

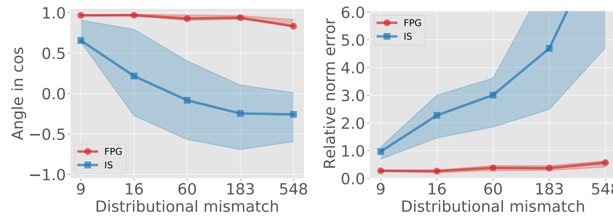

Figure 2. **Tolerance to off-policy distribution shift.** The distributional mismatch is measured by $\mathrm{cond}(\bar{\Sigma}^{\frac{1}{2}}\Sigma^{-1}\bar{\Sigma}^{\frac{1}{2}})$, where Σ is the data covariance and $\bar{\Sigma}$ is the target policy's occupancy measure.

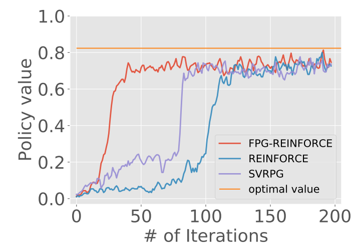

Figure 3. **FPG for policy optimization.** FPG is used as a module for PG estimates in REINFORCE, compared with other baselines.

For FrozenLake, we use softmax tabular policy parameterization and $H = 100$. For CliffWalking, we use softmax on top of a two-layer ReLU network for policy parameterization. We pick the target policy to be a fixed near-optimal policy, and test using dataset generated from different behavior policies. For comparison, we compute the true gradient using the policy gradient theorem and on-policy Monte Carlo simulation.

FPG's data efficiency Choosing the behavior policy to be the ε-greedy modification of the target policy for $\varepsilon = 0.1$, we generate datasets with varying sizes and evaluate the FPG's estimation error on two metrics: the cosine angle between the true policy gradient and the FPG estimator, and the relative estimation error in ℓ_2-norm. The closer the cosine is to 1 and the smaller the relative norm error is, the better the estimated policy gradient is. **Figure 1** shows that FPG gives good estimate even when the data is rather small. The FPG estimate converges to the true gradient with rather moderate variance. In comparison, importance sampling (IS) converges much slower and incurs substantially larger variance.

The effect of distributional mismatch Next we investigate the effect of distribution shift on off-policy PG estimation. We consider 5 choices of behavior policies: the target policy, the ε-greedy policies of the target policy, with $\varepsilon = 0.1, 0.3, 0.5$ and 0.7. We generate a dataset containing 200 episodes with each of these behavior policies, run FPG and IS, and evaluate their estimation errors. **Figure 2** shows that larger distribution mismatch leads to larger estimation error in both methods. However, when compared to IS, FPG is significantly more robust to off-policy distribution shift. The accuracy of FPG only degrades slightly with larger distribution mismatch, while IS suffers from exponentially blowing-up error and stops generating reasonable estimates.

FPG for policy optimization We further showcase FPG's applicability to policy optimization. In particular, we test FPG as a gradient estimation module in policy gradient optimization methods. We conduct an experiment using FPG in on-policy REINFORCE and compare it with the vanilla REINFORCE and SVRPG (Papini et al., 2018). All methods are configured to sample 100 on-policy episodes per iteration. When implementing the FPG-REINFORCE, we take advantage of FPG's off-policy capability and use data from the recent 5 iterations to improve the gradient estimation accuracy. **Figure 3** shows that such design indeed allows FPG-REINFORCE to converge significantly faster than the two baselines.

Next we test the use of FPG for offline policy optimization. Let the target policy be the optimal policy of the problem, and let the behavior policy be a 0.3-greedy variant of the target policy. We generate a dataset consisting of $K = 500$ episodes by simulating the behavior policy. For PG estimation, we only use the offline dataset and do not sample for fresh data. Thus, we replace the online policy gradient estimator in REINFORCE with an off-policy one using FPG, and for comparison we also test REINFORCE with an IS estimator. **Figure 4** shows that FPG-REINFORCE converges reasonably fast and approaches the optimal value. However, IS-REINFORCE appears to converge to a highly biased solution, due to that all PGs are estimated using the same small batch dataset and suffer from bias due to distribution shift.

FPG with deep neural network policy parameterization Further, we evaluate the efficiency of FPG when using a deep policy network for policy learning in the CliffWalking environment, where $H = 100$. Specifically, the environ-

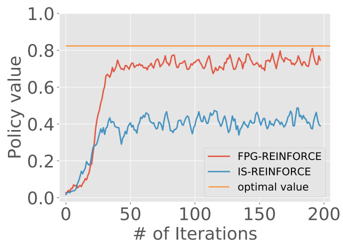

Figure 4. **FPG for offline policy optimization.** FPG is compared with IS as the PG estimator module in offline REINFORCE.

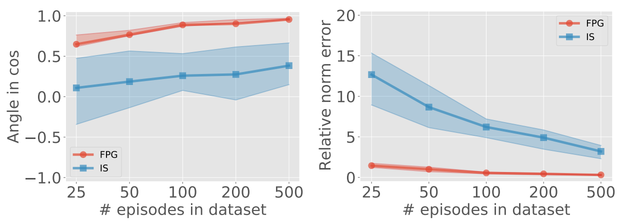

Figure 5. **Sample efficiency of FPG with deep policy networks on off-policy data.** The off-policy PG estimation accuracy is evaluated using two metrics: $\cos\angle(\widehat{\nabla v_\theta}, \nabla v_\theta)$ and the relative error norm $\frac{\|\widehat{\nabla v_\theta} - \nabla v_\theta\|}{\|\nabla v_\theta\|}$.

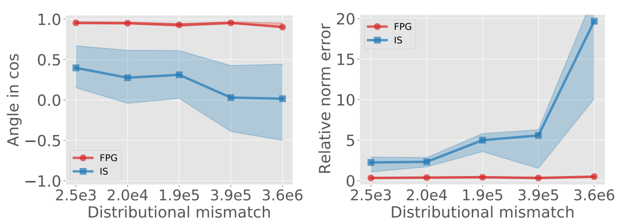

Figure 6. **Tolerance to off-policy distribution shift with deep policy networks.** The distributional mismatch is measured by $\mathrm{cond}(\Sigma^{1/2}\overline{\Sigma}^{-1}\Sigma^{1/2})$, where $\overline{\Sigma}$ is the data covariance and Σ is the target policy's occupancy measure.

ment is modified by adding artificial randomness for stochastic transitions, that is, at each transition, a random action is taken with probability 0.1. The policy is parameterized with a neural network with one ReLU hidden layer and a softmax layer.

As before, we test the performance of FPG against the size of off-policy data and the degree of distribution shift, using the same cosine and relative norm error metrics. We test FPG varying the size of the dataset. **Figure 5** shows that FPG still gives accurate estimates with moderate variance, in contrast to IS's inaccuracy and high variance. Both methods become more accurate asymptotically as the dataset size increases.

We also test FPG on datasets with different amount of mismatch from the target policy. **Figure 6** shows that FPG's estimation error is much lower and less affected by enlarging distributional mismatch than IS's. The distribution mismatch is large in the CliffWalking experiments because some state-action pairs are almost never visited by the target policy and seldom visited by the behavior policy. Such state-action pairs cause $\Sigma^{1/2}\overline{\Sigma}^{-1}\Sigma^{1/2}$ to be nearly singular, but they are irrelevant to our estimation. The general trends in these CliffWalking experiments with deep neural network policy are consistent with our theoretical results and FrozenLake experiments.

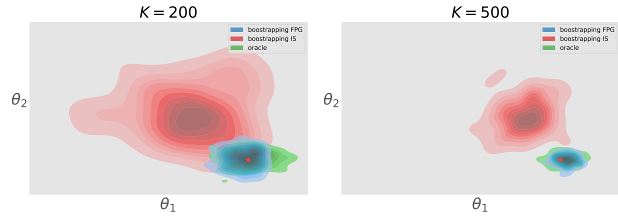

Figure 7. **Bootstrapped confidence region for FPG estimator, with sample size $K = 200, 500$.** The red dot marks the true gradient. Blue and red areas are confidence sets obtained by bootstrapping FPG and IS respectively. The green oracle gives the empirical confidence region for FPG estimates.

Bootstrap inference for FPG Finally we apply bootstrap inference to construct confidence regions of FPG estimates by subsampling episodes and estimating the bootstrapped probability distribution. We plot contours of bootstrapped confidence regions via quantile KDE. **Figure 7** visualizes the bootstrapped confidence regions in 2D, compared with the confidence region for IS and the ground-truth confidence set. Across all experiments, we observe that the contours of bootstrapping FPG are much smaller and more accurate than the ones of bootstrapping IS. As the K increases, the bootstrapped confidence regions become more concentrated, confirming our theoretical results.

8. Conclusion

We propose double Fitted Policy Gradient iteration (FPG) for off-policy PG estimation. FPG theoretically achieves near-optimal rate that matches the Cramar-Rao lower bound and empirically outperforms classic methods on a variety of tasks. Future work includes extension to non-linear function approximation and evaluation on more complex domains.

Acknowledgement This work is supported by NSF grants IIS-2107304, CMMI-1653435, AFOSR grant and ONR grant 1006977.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Chang, J. D., Uehara, M., Sreenivas, D., Kidambi, R., and Sun, W. Mitigating covariate shift in imitation learning via offline data without great coverage. 2021.

Degris, T., White, M., and Sutton, R. S. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Duan, Y., Wang, M., and Wainwright, M. J. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*, 2021.

Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R. E., and Levine, S. Q-prop: Sample-efficient policy gradient with an off-policy critic. *arXiv preprint arXiv:1611.02247*, 2016.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvari, C., and Wang, M. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pp. 4074–4084. PMLR, 2021.

Ji, X., Chen, M., Wang, M., and Zhao, T. Sample complexity of nonparametric off-policy evaluation on low-dimensional manifolds using deep networks. *arXiv preprint arXiv:2206.02887*, 2022.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Jie, T. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. *Advances in Neural Information Processing Systems*, 23: 1000–1008, 2010.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Kallus, N. and Uehara, M. Statistically efficient off-policy policy gradients. In *International Conference on Machine Learning*, pp. 5089–5100. PMLR, 2020.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Morimura, T., Uchibe, E., Yoshimoto, J., Peters, J., and Doya, K. Derivatives of logarithmic stationary distributions for policy gradient reinforcement learning. *Neural computation*, 22(2):342–376, 2010.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pp. 4026–4035. PMLR, 2018.

Peters, J. and Schaal, S. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Polyak, B. T. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Shelton, C. R. Policy improvement for pomdps using normalized importance sampling. *arXiv preprint arXiv:1301.2310*, 2013.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

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

Tokdar, S. T. and Kass, R. E. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

Tosatto, S., Carvalho, J., Abdulsamad, H., and Peters, J. A nonparametric off-policy policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 167–177. PMLR, 2020.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Wang, R., Wu, Y., Salakhutdinov, R., and Kakade, S. M. Instabilities of offline rl with pre-trained neural representation. *arXiv preprint arXiv:2103.04947*, 2021.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *arXiv preprint arXiv:1906.03393*, 2019.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

Xu, P., Gao, F., and Gu, Q. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pp. 541–551. PMLR, 2020.

Xu, T., Yang, Z., Wang, Z., and Liang, Y. Doubly robust off-policy actor-critic: Convergence and optimality. *arXiv preprint arXiv:2102.11866*, 2021.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*, 2021.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

Zhang, J., Ni, C., Szepesvari, C., Wang, M., et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021a.

Zhang, R., Zhang, X., Ni, C., and Wang, M. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. *arXiv preprint arXiv:2202.04970*, 2022.

Zhang, X., Chen, Y., Zhu, J., and Sun, W. Corruption-robust offline reinforcement learning. *arXiv preprint arXiv:2106.06630*, 2021b.

A. Technical Lemmas

Let $U_h^\theta = \mathcal{P}_{\theta,h}\left(\nabla_\theta \log \Pi_{\theta,h+1}\right) Q_{h+1}^\theta,\ \widehat{U}_h^\theta = \widehat{\mathcal{P}}_{\theta,h}\left(\nabla_\theta \log \Pi_{\theta,h+1}\right) Q_{h+1}^\theta,\ \tilde{U}_h^\theta = \widehat{\mathcal{P}}_{\theta,h}\left(\nabla_\theta \log \Pi_{\theta,h+1}\right) \widehat{Q}_{h+1}^\theta,$

Lemma A.1. *We have*

$$Q_h^\theta = \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\mathcal{P}_{\theta,h''}\right) r_{h'}, \quad \widehat{Q_h^\theta} = \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right) \widehat{r}_{h'},$$
$$\nabla_\theta Q_h^\theta = \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\mathcal{P}_{\theta,h''}\right) U_{h'}^\theta, \quad \widehat{\nabla_\theta Q_h^\theta} = \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right) \tilde{U}_{h'}^\theta.$$

Proof. By Bellman's equation, we have $Q_h^\theta = r_h + \mathcal{P}_{\theta,h} Q_{h+1}^\theta$. Therefore, by induction and use the fact that $Q_{H+1}^\theta = 0$, we have proved the first equation. By the policy gradient Bellman's equation, we have

$$\nabla_\theta Q_h^\theta(s,a) = \mathbb{E}^{\pi_\theta}\left[\left(\nabla_\theta \log \pi_{\theta,h+1}(a'|s')\right) Q_{h+1}^\theta(s',a')|s,a,h\right] + \mathbb{E}^{\pi_\theta}\left[\nabla_\theta Q_{h+1}^\theta(s',a')|s,a,h\right],$$

i.e., $\nabla_\theta Q_h^\theta = U_h^\theta + \mathcal{P}_{\theta,h}\nabla_\theta Q_{h+1}^\theta$. By induction, we have proved the third equation. The expressions of $\widehat{Q}_h^\theta$ and $\widehat{\nabla_\theta Q_h^\theta}$ can be derived directly from their definitions and induction. □

The decomposition leads to the following boundedness result:

Lemma A.2. *We have* $|Q_h^\theta(s,a)| \le H - h + 1,\ \|\nabla_\theta Q_h^\theta(s,a)\|_\infty \le G(H-h)^2,\ \forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H].$

Now we consider the decomposition of $Q_h^\theta - \widehat{Q}_h^\theta$:

Lemma A.3. *We have*

$$Q_h^\theta - \widehat{Q}_h^\theta = \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\left(Q_{h'}^\theta - \widehat{r}_{h'} - \widehat{\mathcal{P}}_{\theta,h'} Q_{h'+1}^\theta\right), \quad \forall h \in [H].$$

Proof. Simply note that

$$\begin{aligned}
Q_h^\theta - \widehat{Q}_h^\theta &= \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\mathcal{P}_{\theta,h''}\right) r_{h'} - \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\widehat{r}_{h'} \\
&= \sum_{h'=h}^{H}\left(\left(\prod_{h''=h}^{h'-1}\mathcal{P}_{\theta,h''}\right) - \left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\right) r_{h'} + \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)(r_{h'} - \widehat{r}_{h'}) \\
&= \sum_{h'=h}^{H}\sum_{h''=h}^{h'-1}\left(\prod_{h'''=h}^{h''-1}\widehat{\mathcal{P}}_{\theta,h'''}\right)\left(\mathcal{P}_{\theta,h''} - \widehat{\mathcal{P}}_{\theta,h''}\right)\left(\prod_{h'''=h''+1}^{h'-1}\mathcal{P}_{\theta,h'''}\right) r_{h'} + \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)(r_{h'} - \widehat{r}_{h'}) \\
&= \sum_{h''=h}^{H}\left(\prod_{h'''=h}^{h''-1}\widehat{\mathcal{P}}_{\theta,h'''}\right)\left(\mathcal{P}_{\theta,h''} - \widehat{\mathcal{P}}_{\theta,h''}\right)\sum_{h'=h''+1}^{H}\left(\prod_{h'''=h''+1}^{h'-1}\mathcal{P}_{\theta,h'''}\right) r_{h'} + \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)(r_{h'} - \widehat{r}_{h'}) \\
&= \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\left(\mathcal{P}_{\theta,h'} - \widehat{\mathcal{P}}_{\theta,h'}\right) Q_{h'+1}^\theta + \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)(r_{h'} - \widehat{r}_{h'}) \\
&= \sum_{h'=h}^{H}\left(\prod_{h''=h}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\left(Q_{h'}^\theta - \widehat{r}_{h'} - \widehat{\mathcal{P}}_{\theta,h'} Q_{h'+1}^\theta\right),
\end{aligned}$$

which is the desired result. □

The following lemma provides an upper bound of matrix production, which will be used when bounding the higher order terms of the finite sample bound.

Lemma A.4. *For any series of matrices $A_1, A_2, \ldots, A_n$ and $\Delta A_1, \Delta A_2, \ldots \Delta A_n$, we have*

$$\left\|\prod_{i=1}^{n}(A_i + \Delta A_i) - \prod_{i=1}^{n} A_i\right\| \le \prod_{i=1}^{n}(\|A_i\| + \|\Delta A_i\|) - \prod_{i=1}^{n}\|A_i\|.$$

Proof. We have

$$\left\|\prod_{i=1}^{n}(A_i + \Delta A_i) - \prod_{i=1}^{n} A_i\right\| = \left\|\sum_{\delta\in\{0,1\}^n\setminus\{(1,1,\ldots,1)\}}\prod_{i=1}^{n} A_i^{\delta_i}(\Delta A_i)^{1-\delta_i}\right\| \le \sum_{\delta\in\{0,1\}^n\setminus\{(1,1,\ldots,1)\}}\prod_{i=1}^{n}\|A_i\|^{\delta_i}\|\Delta A_i\|^{1-\delta_i}$$
$$= \prod_{i=1}^{n}(\|A_i\| + \|\Delta A_i\|) - \prod_{i=1}^{n}\|A_i\|.$$

□

When $\mathcal{F}$ is the class of the linear functions, there exists matrix $M_{\theta,h}$ such that the transition probability satisfies

$$\mathbb{E}^{\pi_\theta}\left[\phi(s', a')^\top | s, a, h\right] = \phi(s, a)^\top M_{\theta,h}.$$

The following lemma gives an upper bound on the 2-norm of $M_{\theta,h}$ and its derivatives.

Lemma A.5. *We have* $\left\|\Sigma_{\theta,h}^{\frac{1}{2}} M_{\theta,h}\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\| \le 1$ *and* $\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\left(\nabla_\theta^j M_{\theta,h}\right)\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\| \le G,\ \forall j \in [m], h \in [H].$

Proof. Note that for any $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R},\ f(s,a) := \mu^\top\phi(s,a)$, we have

$$\begin{aligned}\mathbb{E}^{\pi_\theta}\left[f^2(s_{h+1}, a_{h+1})|s_1 \sim \xi\right] &= \mathbb{E}^{\pi_\theta}\left[\mathbb{E}^{\pi_\theta}\left[f^2(s_{h+1}, a_{h+1})|s_h, a_h\right]|s_1 \sim \xi\right]\\ &\ge \mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[f(s_{h+1}, a_{h+1})|s_h, a_h]^2|s_1 \sim \xi]\end{aligned}$$

The LHS satisfies

$$\mathbb{E}^{\pi_\theta}[f^2(s_{h+1}, a_{h+1})|s_1 \sim \xi] = \mu^\top\Sigma_{\theta,h+1}\mu$$

and the RHS satisfies

$$\mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[f(s_{h+1}, a_{h+1})|s_h, a_h]^2|s_1 \sim \xi] = \mathbb{E}^{\pi_\theta}[\mu^\top M_{\theta,h}^\top\phi(s_h, a_h)\phi(s_h, a_h)^\top M_{\theta,h}\mu|s_1 \sim \xi] = \mu^\top M_{\theta,h}^\top\Sigma_{\theta,h}M_{\theta,h}\mu$$

Therefore, we have $\mu^\top\Sigma_{\theta,h+1}\mu \ge \mu^\top M_{\theta,h}^\top\Sigma_{\theta,h}M_{\theta,h}\mu,\ \forall\mu$, which implies $\|\Sigma_{\theta,h}^{\frac{1}{2}}M_{\theta,h}\Sigma_{\theta,h+1}^{-\frac{1}{2}}\| \le 1$. Similarly, let $g : \mathcal{S}\times\mathcal{A} \to \mathbb{R},\ g(s,a) := \left(\nabla_\theta^j \log\pi_{\theta,h+1}(s,a)\right)\mu^\top\phi(s,a)$, we have

$$\begin{aligned}\mathbb{E}^{\pi_\theta}[g^2(s_{h+1}, a_{h+1})|s_1 \sim \xi] =&\mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[g^2(s_{h+1}, a_{h+1})|s_h, a_h]|s_1 \sim \xi]\\ \ge&\mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[g(s_{h+1}, a_{h+1})|s_h, a_h]^2|s_1 \sim \xi]\end{aligned}$$

The LHS satisfies

$$\begin{aligned}\mathbb{E}^{\pi_\theta}[g^2(s_{h+1}, a_{h+1})|s_1 \sim \xi] &= \mu^\top\mathbb{E}^{\pi_\theta}\left[\left(\nabla_\theta^j\log\pi_{\theta,h+1}(a|s)\right)^2\phi(s_{h+1}, a_{h+1})\phi(s_{h+1}, a_{h+1})^\top|s_1 \sim \xi\right]\mu\\ &\le G^2\mu^\top\mathbb{E}^{\pi_\theta}\left[\phi(s_{h+1}, a_{h+1})\phi(s_{h+1}, a_{h+1})^\top|s_1 \sim \xi\right]\mu = G^2\mu^\top\Sigma_{\theta,h+1}\mu\end{aligned}$$

and the RHS satisfies

$$\begin{aligned}&\mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[g(s_{h+1}, a_{h+1})|s_h, a_h]^2|s_1 \sim \xi]\\ =&\mathbb{E}^{\pi_\theta}[\mu^\top\left(\nabla_\theta^j M_{\theta,h}\right)^\top\phi(s_h, a_h)\phi(s_h, a_h)^\top\left(\nabla_\theta^j M_{\theta,h}\right)\mu|s_1 \sim \xi]\\ =&\mu^\top\left(\nabla_\theta^j M_{\theta,h}\right)^\top\Sigma_{\theta,h}\left(\nabla_\theta^j M_{\theta,h}\right)\mu.\end{aligned}$$

Therefore, we get $G^2\mu^\top \Sigma_{\theta,h+1}\mu \geq \mu^\top \left(\nabla_\theta^j M_{\theta,h}\right)^\top \Sigma_{\theta,h}\left(\nabla_\theta^j M_{\theta,h}\right)\mu,\ \forall \mu$, which implies $\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\left(\nabla_\theta^j M_\theta\right)\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\| \leq G$. □

Lemma A.6. *We have with probability at least $1-\delta$,*

$$\left\|\Sigma_h^{-\frac{1}{2}}\left(\frac{1}{K}\sum_{k=1}^K \phi(s_h^{(k)},a_h^{(k)})\phi(s_h^{(k)},a_h^{(k)})^\top\right)\Sigma_h^{-\frac{1}{2}} - I_d\right\| \leq \sqrt{\frac{2C_1 d\log\frac{2dH}{\delta}}{K}} + \frac{2C_1 d\log\frac{2dH}{\delta}}{3K}.$$

Proof. Define

$$X_h^{(k)} = \Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top \Sigma_h^{-\frac{1}{2}} \in \mathbb{R}^{d\times d}.$$

It's easy to see that $X_h^{(1)}, X_h^{(2)}, \ldots, X_h^{(K)}$ are independent and $\mathbb{E}\left[X_h^{(k)}\right] = I_d$. In the remaining part of the proof, we will apply the matrix Bernstein's inequality to analyze the concentration of $\frac{1}{K}\sum_{k=1}^K X_h^{(k)}$. We first consider the matrix-valued variance $\mathrm{Var}\left(X_h^{(k)}\right) = \mathbb{E}\left[\left(X_h^{(k)} - I_d\right)^2\right] = \mathbb{E}\left[\left(X_h^{(k)}\right)^2\right] - I_d$. For any vector $\mu\in\mathbb{R}^d$,

$$\begin{aligned}
\mu^\top \mathbb{E}\left[\left(X_h^{(k)}\right)^2\right]\mu =& \mathbb{E}\left[\left\|X_h^{(k)}\mu\right\|^2\right] = \mathbb{E}\left[\left\|\Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right] \\
\leq& \mathbb{E}\left[\left\|\Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\right\|^2\left\|\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right] \leq C_1 d\mathbb{E}\left[\left\|\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right] \\
=& C_1 d\mu^\top\mathbb{E}\left[X_h^{(k)}\right]\mu = C_1 d\|\mu\|^2,
\end{aligned}$$

where we used the identity $\left\|\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-\frac{1}{2}}\mu\right\|^2 = \mu^\top X_h^{(k)}\mu$ and $\mathbb{E}\left[X_h^{(k)}\right] = I_d$. We have

$$\mathrm{Var}(X_h^{(k)}) \preceq \mathbb{E}\left[\left(X_h^{(k)}\right)^2\right] \preceq C_1 d I_d.$$

Additionally,

$$-I_d \preceq X_h^{(k)} - I_d = \Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-\frac{1}{2}} - I_d \preceq C_1 d I_d - I_d.$$

Therefore, $\|X_h^{(k)} - I_d\| \leq C_1 d$. Since $X_h^{(1)}, X_h^{(2)}, \ldots, X_h^{(K)}$ are *i.i.d.*, by the matrix-form Bernstein inequality, we have

$$\mathbb{P}\left(\left\|\sum_{k=1}^K X_h^{(k)} - I_d\right\| \geq \varepsilon\right) \leq 2d\cdot\exp\left(-\frac{\varepsilon^2/2}{C_1 dK + C_1 d\varepsilon/3}\right), \quad \forall\varepsilon > 0.$$

With probability at least $1-\delta$,

$$\left\|\frac{1}{K}\sum_{k=1}^K\left(X_h^{(k)} - I_d\right)\right\| \leq \sqrt{\frac{2C_1 d\log\frac{2d}{\delta}}{K}} + \frac{2C_1 d\log\frac{2d}{\delta}}{3K},$$

Taking a union bound over $h\in[H]$, we derive the desired result. □

Let $\Delta\Sigma_h^{-1} = \widehat{\Sigma}_h^{-1} - \Sigma_h$,

Lemma A.7. *If* $\left\|\Sigma_h^{-\frac{1}{2}}\widehat{\Sigma}_h\Sigma_h^{-\frac{1}{2}} - I_d\right\| \leq \frac{1}{2}$, *then* $\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\| \leq 2\left\|\Sigma_h^{-\frac{1}{2}}\widehat{\Sigma}_h\Sigma_h^{-\frac{1}{2}} - I_d\right\|$.

Proof. Note that

$$\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\| = \left\|\Sigma_h^{\frac{1}{2}}\left(\widehat{\Sigma}_h^{-1}-\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\| \leq \left\|\Sigma_h^{\frac{1}{2}}\widehat{\Sigma}_h^{-1}\Sigma_h^{\frac{1}{2}}\right\|\left\|\Sigma_h^{-\frac{1}{2}}\widehat{\Sigma}_h\Sigma_h^{-\frac{1}{2}}-I_d\right\|. \tag{7}$$

Because we have $\left\|\Sigma_h^{-\frac{1}{2}}\widehat{\Sigma}_h\Sigma_h^{-\frac{1}{2}}-I_d\right\| \leq \frac{1}{2}$, we get $\sigma_{\min}\left(\Sigma_h^{-\frac{1}{2}}\widehat{\Sigma}_h\Sigma_h^{-\frac{1}{2}}\right) \geq \frac{1}{2}$, which implies $\left\|\Sigma_h^{\frac{1}{2}}\widehat{\Sigma}_h^{-1}\Sigma_h^{\frac{1}{2}}\right\| \leq 2$. Combining this result with (7) finishes the proof. ☐

Let $\Delta Y_{\theta,h} = \frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)^{\top}-\Sigma_h M_{\theta,h}$,

Lemma A.8. *With probability at least $1-\delta$, the following inequalities hold simultaneously:*

$$\left\|\Sigma_h^{-\frac{1}{2}}\left(\Delta Y_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\| \leq (\kappa_2\vee 1)\sqrt{\frac{2C_1 d\log\frac{4dH}{\delta}}{K}}+\frac{4C_1 d\log\frac{4dH}{\delta}}{3K}, \tag{8}$$

$$\left\|\Sigma_h^{-\frac{1}{2}}\left(\nabla_\theta^j\left(\Delta Y_{\theta,h}\right)\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\| \leq (\kappa_3\vee 1)\,G\sqrt{\frac{2C_1 d\log\frac{4mdH}{\delta}}{K}}+\frac{4C_1 dG\log\frac{4mdH}{\delta}}{3K}, \quad \forall j\in[m], \tag{9}$$

where κ_2,κ_3 are defined in Theorem 6.2.

Proof. Take

$$Y_{\theta,h}^{(k)} := \Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)^{\top}\Sigma_{h+1}^{-\frac{1}{2}}, \quad \forall k\in[K].$$

Then, $\Sigma_h^{-\frac{1}{2}}\left(\Delta Y_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}} = \frac{1}{K}\sum_{k=1}^{K}\left(Y_{\theta,h}^{(k)}-\Sigma_h^{\frac{1}{2}}M_{\theta,h}\Sigma_{h+1}^{-\frac{1}{2}}\right)$. Note that

$$\begin{aligned}\mathbb{E}\left[Y_{\theta,h}^{(k)}\right] =&\mathbb{E}\left[\Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)^{\top}\Sigma_{h+1}^{-\frac{1}{2}}\right]\\ =&\mathbb{E}\left[\Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\mathbb{E}^{\pi_\theta}\left[\phi\left(s',a'\right)^{\top}|s_h^{(k)},a_h^{(k)},h\right]\Sigma_{h+1}^{-\frac{1}{2}}\right]\\ =&\mathbb{E}\left[\Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^{\top}M_{\theta,h}\Sigma_{h+1}^{-\frac{1}{2}}\right] = \Sigma_h^{\frac{1}{2}}M_{\theta,h}\Sigma_{h+1}^{-\frac{1}{2}},\end{aligned} \tag{10}$$

To this end, $\Sigma_h^{-\frac{1}{2}}\left(\Delta Y_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}} = \frac{1}{K}\sum_{k=1}^{K}\left(Y_{\theta,h}^{(k)}-\mathbb{E}\left[Y_{\theta,h}^{(k)}\right]\right)$. Since the trajectories are i.i.d., we use the matrix-form Bernstein inequality to estimate $\left\|\Sigma_h^{-\frac{1}{2}}(\Delta Y_{\theta,h})\Sigma_{h+1}^{-\frac{1}{2}}\right\|$. For any $\mu\in\mathbb{R}^d$, we have

$$\begin{aligned}\mu^{\top}\mathbb{E}\left[Y_{\theta,h}^{(k)}\left(Y_{\theta,h}^{(k)}\right)^{\top}\right]\mu =&\mathbb{E}\left[\left\|\left(Y_{\theta,h}^{(k)}\right)^{\top}\mu\right\|^2\right] = \mathbb{E}\left[\left\|\Sigma_{h+1}^{-\frac{1}{2}}\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^{\top}\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right]\\ \leq&\mathbb{E}\left[\left\|\Sigma_{h+1}^{-\frac{1}{2}}\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\right\|^2\left\|\phi\left(s_h^{(k)},a_h^{(k)}\right)^{\top}\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right].\end{aligned}$$

Parallel to the proof of Lemma A.6, it holds that $\left\|\Sigma_{h+1}^{-\frac{1}{2}}\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\right\|^2 \leq C_1 d$. Therefore,

$$\begin{aligned}\mu^{\top}\mathbb{E}\left[Y_{\theta,h}^{(k)}\left(Y_{\theta,h}^{(k)}\right)^{\top}\right]\mu &\leq \mathbb{E}\left[C_1 d\left\|\phi\left(s_h^{(k)},a_h^{(k)}\right)^{\top}\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right]\\ &= C_1 d\mu^{\top}\Sigma_h^{-\frac{1}{2}}\mathbb{E}\left[\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^{\top}\right]\Sigma_h^{-\frac{1}{2}}\mu\\ &= C_1 d\|\mu\|^2,\end{aligned}$$

where we have used the fact $\Sigma_h = \mathbb{E}\left[\phi\left(s_h^{(k)}, a_h^{(k)}\right)\phi\left(s_h^{(k)}, a_h^{(k)}\right)^\top\right]$. It follows that

$$\mathrm{Var}_1\left(Y_{\theta,h}^{(k)}\right) := \mathbb{E}\left[\left(Y_{\theta,h}^{(k)} - \mathbb{E}\left[Y_{\theta,h}^{(k)}\right]\right)\left(Y_{\theta,h}^{(k)} - \mathbb{E}\left[Y_{\theta,h}^{(k)}\right]\right)^\top\right] \preceq \mathbb{E}\left[Y_{\theta,h}^{(k)}\left(Y_{\theta,h}^{(k)}\right)^\top\right] \preceq C_1 d I_d.$$

Analogously,

$$\begin{aligned}\mathrm{Var}_2\left(Y_{\theta,h}^{(k)}\right) :=& \mathbb{E}\left[\left(Y_{\theta,h}^{(k)} - \mathbb{E}\left[Y_{\theta,h}^{(k)}\right]\right)^\top\left(Y_{\theta,h}^{(k)} - \mathbb{E}\left[Y_{\theta,h}^{(k)}\right]\right)\right] \preceq \mathbb{E}\left[\left(Y_{\theta,h}^{(k)}\right)^\top Y_{\theta,h}^{(k)}\right] \\ \preceq& C_1 d \Sigma_{h+1}^{-\frac{1}{2}} \mathbb{E}\left[\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)^\top\right]\Sigma_{h+1}^{-\frac{1}{2}}.\end{aligned}$$

Therefore, $\max\left\{\left\|\mathrm{Var}_1\left(Y_{\theta,h}^{(k)}\right)\right\|, \left\|\mathrm{Var}_2\left(Y_{\theta,h}^{(k)}\right)\right\|\right\} \le C_1 d\left(\kappa_2^2 \vee 1\right)$. It also holds that $\|Y_{\theta,h}^{(k)}\| \le C_1 d$. Hence,

$$\left\|Y_{\theta,h}^{(k)} - \Sigma_h^{\frac{1}{2}} M_{\theta,h}\Sigma_{h+1}^{-\frac{1}{2}}\right\| \le 2C_1 d.$$

Applying Matrix Bernstein's inequality, we derive for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^K\left(Y_{\theta,h}^{(k)} - \Sigma_h^{\frac{1}{2}} M_{\theta,h}\Sigma_{h+1}^{-\frac{1}{2}}\right)\right\| > \varepsilon\right) \le 2d\exp\left(-\frac{\varepsilon^2/2}{C_1 dK\left(\kappa_2^2 \vee 1\right) + 2C_1 d\varepsilon/3}\right),$$

which implies (8) holds with probability $1 - \frac{\delta}{2}$. For (9), notice that for any $j \in [m]$, we have $\Sigma_h^{-\frac{1}{2}}(\nabla_\theta^j(\Delta Y_{\theta,h}))\Sigma_{h+1}^{-\frac{1}{2}} = \frac{1}{K}\sum_{k=1}^K\left(\nabla_\theta^j Y_{\theta,h}^{(k)} - \mathbb{E}\left[\nabla_\theta^j Y_{\theta,h}^{(k)}\right]\right)$, and $\nabla_\theta^j Y_{\theta,h}^{(k)} = \Sigma_h^{-\frac{1}{2}}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\left(\nabla_\theta^j\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\right)^\top\Sigma_{h+1}^{-\frac{1}{2}}$. For any $\mu \in \mathbb{R}^d$, we have

$$\begin{aligned}\mu^\top\mathbb{E}\left[\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)^\top\right]\mu =& \mathbb{E}\left[\left\|\Sigma_{h+1}^{-\frac{1}{2}}\left(\nabla_\theta^j\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\right)\phi\left(s_h^{(k)}, a_h^{(k)}\right)^\top\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right] \\ \le& \mathbb{E}\left[\left\|\Sigma_{h+1}^{-\frac{1}{2}}\nabla_\theta^j\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\right\|^2\left\|\phi\left(s_h^{(k)}, a_h^{(k)}\right)^\top\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right].\end{aligned}$$

Since we have

$$\begin{aligned}&\left(\nabla_\theta^j\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\right)^\top\Sigma_{h+1}^{-1}\nabla_\theta^j\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right) \\ =& \int_{\mathcal{A}\times\mathcal{A}}\pi_{\theta,h+1}\left(a\,\middle|\,s_{h+1}^{(k)}\right)\pi_{\theta,h+1}\left(a'\,\middle|\,s_{h+1}^{(k)}\right)\left(\nabla_\theta^j\log\pi_{\theta,h+1}\left(a\,\middle|\,s_{h+1}^{(k)}\right)\right)\left(\nabla_\theta^j\log\pi_{\theta,h+1}\left(a'\,\middle|\,s_{h+1}^{(k)}\right)\right) \\ &\cdot\phi\left(s_{h+1}^{(k)}, a\right)^\top\Sigma_{h+1}^{-1}\phi\left(s_{h+1}^{(k)}, a'\right)\mathrm{d}a\mathrm{d}a' \\ \le& G^2\int_{\mathcal{A}\times\mathcal{A}}\pi_{\theta,h+1}\left(a|s_{h+1}^{(k)}\right)\pi_{\theta,h+1}\left(a'|s_{h+1}^{(k)}\right)\left\|\Sigma_{h+1}^{-\frac{1}{2}}\phi\left(s_{h+1}^{(k)}, a\right)\right\|\left\|\Sigma_{h+1}^{-\frac{1}{2}}\phi\left(s_{h+1}^{(k)}, a'\right)\right\|\mathrm{d}a\mathrm{d}a' \le G^2 C_1 d,\end{aligned}$$

which implies

$$\mu^\top\mathbb{E}\left[\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)^\top\right]\mu \le G^2 C_1 d\mathbb{E}\left[\left\|\phi\left(s_h^{(k)}, a_h^{(k)}\right)^\top\Sigma_h^{-\frac{1}{2}}\mu\right\|^2\right] = G^2 C_1 d\|\mu\|^2.$$

Therefore,

$$\mathrm{Var}_1\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right) := \mathbb{E}\left[\left(\nabla_\theta^j Y_{\theta,h}^{(k)} - \mathbb{E}\left[\nabla_\theta^j Y_{\theta,h}^{(k)}\right]\right)\left(\nabla_\theta^j Y_{\theta,h}^{(k)} - \mathbb{E}\left[\nabla_\theta^j Y_{\theta,h}^{(k)}\right]\right)^\top\right] \preceq \mathbb{E}\left[\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)^\top\right] \preceq G^2 C_1 d I_d.$$

Meanwhile, we have

$$\mathrm{Var}_2\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right) \preceq \mathbb{E}\left[\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)^\top\nabla_\theta^j Y_{\theta,h}^{(k)}\right] \preceq C_1 d\Sigma_h^{-\frac{1}{2}}\mathbb{E}\left[\left(\nabla_\theta^j\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\right)\left(\nabla_\theta^j\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)\right)^\top\right]\Sigma_h^{-\frac{1}{2}}.$$

In conclusion, we get

$$\max\left\{\left\|\mathrm{Var}_1\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)\right\|,\left\|\mathrm{Var}_2\left(\nabla_\theta^j Y_{\theta,h}^{(k)}\right)\right\|\right\}\le G^2C_1d\left(\kappa_3^2\vee 1\right),$$

Note that $\left\|\nabla_\theta^j Y_{\theta,h}^{(k)}\right\|\le C_1dG$, we know $\left\|\nabla_\theta^j Y_{\theta,h}^{(k)}-\mathbb{E}\left[\nabla_\theta^j Y_{\theta,h}^{(k)}\right]\right\|\le 2C_1dG$. By Matrix Bernstein's inequality, we get for any $\varepsilon>0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^K\left(\nabla_\theta^j Y_{k,h}^\theta-\Sigma_h^{\frac12}\left(\nabla_\theta^j M_{\theta,h}\right)\Sigma_{h+1}^{-\frac12}\right)\right\|\ge\varepsilon\right)\le 2d\exp\left(-\frac{\varepsilon^2/2}{G^2C_1dK\left(\kappa_3^2\vee1\right)+2C_1dG\varepsilon/3}\right),$$

taking a union bound over all $j\in[m]$ and $h\in[H]$ proves that (9) holds with probability $1-\frac{\delta}{2}$. Using a union bound argument again, we know with probability $1-\delta$, (8) and (9) hold simultaneously, which has finished the proof. ☐

Lemma A.9. *For $h\in[H]$, with probability at least $1-\delta$, the following inequalities hold simultaneously: $\forall j\in[m]$,*

$$\left\|\Sigma_h^{-\frac12}\frac1K\sum_{k=1}^K\phi\left(s_h^{(k)},a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right\|\le\sqrt{d}(H-h+1)\left(\sqrt{\frac{2\log\frac{8dH}{\delta}}{K}}+\frac{2\sqrt{C_1d}\log\frac{8dH}{\delta}}{K}+\frac{2C_1d\left(\log\frac{8dH}{\delta}\right)^{\frac32}}{3K^{\frac32}}\right) \tag{11}$$

$$\left\|\Sigma_h^{-\frac12}\frac1K\sum_{k=1}^K\phi\left(s_h^{(k)},a_h^{(k)}\right)\nabla_\theta^j\varepsilon_{h,k}^\theta\right\|\le2\sqrt{d}G(H-h)^2\left(\sqrt{\frac{2\log\frac{8mdH}{\delta}}{K}}+\frac{2\sqrt{C_1d}\log\frac{8mdH}{\delta}}{K}+\frac{2C_1d\left(\log\frac{8mdH}{\delta}\right)^{\frac32}}{3K^{\frac32}}\right). \tag{12}$$

Proof. Let $X_{\theta,h}^{(k)}:=\Sigma_h^{-\frac12}\phi\left(s_h^{(k)},a_h^{(k)}\right)\varepsilon_{h,k}^\theta\in\mathbb{R}^d$ and let $\mathcal{F}_{h,k}$ be σ-algebra generated by the history up to step h at episode k, we have $\mathbb{E}\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]=0$. We apply matrix-form Freedman's inequality to analyze the concentration property. Consider conditional variances $\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]:=\mathbb{E}\left[X_{\theta,h}^{(k)}\left(X_{\theta,h}^{(k)}\right)^\top\middle|\mathcal{F}_{h,k}\right]\in\mathbb{R}^{d\times d}$ and $\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]:=\mathbb{E}\left[\left(X_{\theta,h}^{(k)}\right)^\top X_{\theta,h}^{(k)}|\mathcal{F}_{h,k}\right]\in\mathbb{R}$. It holds that

$$\begin{aligned}\left\|\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|=&\left\|\mathbb{E}\left[X_{\theta,h}^{(k)}\left(X_{\theta,h}^{(k)}\right)^\top\middle|\mathcal{F}_{h,k}\right]\right\|\le\mathbb{E}\left[\left\|X_{\theta,h}^{(k)}\left(X_{\theta,h}^{(k)}\right)^\top\right\|\middle|\mathcal{F}_{h,k}\right]\\=&\mathbb{E}\left[\left\|X_{\theta,h}^{(k)}\right\|^2\middle|\mathcal{F}_{h,k}\right]=\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]=&\mathbb{E}\left[\left\|X_{\theta,h}^{(k)}\right\|^2\middle|\mathcal{F}_{h,k}\right]=\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right)\mathrm{Var}\left[\varepsilon_{h,k}^\theta\middle|s_h^{(k)},a_h^{(k)},h\right]\\\le&(H-h+1)^2\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right),\end{aligned}$$

where we have used $\mathrm{Var}\left[\varepsilon_{h,k}^\theta\middle|s_h^{(k)},a_h^{(k)},h\right]\le(H-h+1)^2$. Note that

$$\begin{aligned}\sum_{k=1}^K\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right)=&Kd+K\mathrm{Tr}\left(\Sigma_h^{-\frac12}\left(\frac1K\sum_{k=1}^K\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\right)\Sigma_h^{-\frac12}-I_d\right)\\\le&Kd+Kd\left\|\Sigma_h^{-\frac12}\left(\frac1K\sum_{k=1}^K\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\right)\Sigma_h^{-\frac12}-I_d\right\|.\end{aligned}$$

We take

$$\sigma^2 := Kd(H-h+1)^2\left(1+\sqrt{\frac{2C_1 d\log\frac{8dH}{\delta}}{K}}+\frac{2C_1 d\log\frac{8dH}{\delta}}{3K}\right). \tag{13}$$

According to Lemma A.6, it holds that

$$\mathbb{P}\left(\left\|\sum_{k=1}^{K}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|\leq\sum_{k=1}^{K}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\leq\sigma^2\right)\geq 1-\frac{\delta}{4}. \tag{14}$$

Additionally, we have $\left\|X_{\theta,h}^{(k)}\right\|\leq (H-h+1)\sqrt{C_1 d}$. The Freedman's inequality therefore implies that for any $\varepsilon>0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{k=1}^{K}X_{\theta,h}^{(k)}\right|\geq\varepsilon,\ \max\left\{\left\|\sum_{k=1}^{K}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|,\sum_{k=1}^{K}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\}\leq\sigma^2\right)\\ \leq&2d\exp\left(-\frac{\varepsilon^2/2}{\sigma^2+(H-h+1)\sqrt{C_1 d}\varepsilon/3}\right),\end{aligned} \tag{15}$$

where σ^2 is defined in (13). We take

$$\varepsilon := \sigma\sqrt{2\log\frac{8d}{\delta}}+\frac{2(H-h+1)\sqrt{C_1 d}}{3}\log\frac{8d}{\delta}.$$

Then we get

$$\mathbb{P}\left(\left|\sum_{k=1}^{K}X_{\theta,h}^{(k)}\right|\geq\varepsilon,\ \max\left\{\left\|\sum_{k=1}^{K}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|,\sum_{k=1}^{K}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\}\leq\sigma^2\right)\leq\frac{\delta}{4},$$

which implies

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{k=1}^{K}X_{\theta,h}^{(k)}\right|\geq\varepsilon\right)\leq&\mathbb{P}\left(\left|\sum_{k=1}^{K}X_{\theta,h}^{(k)}\right|\geq\varepsilon,\ \max\left\{\left\|\sum_{k=1}^{K}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|,\sum_{k=1}^{K}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\}\leq\sigma^2\right)\\ &+\mathbb{P}\left(\max\left\{\left\|\sum_{k=1}^{K}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|,\sum_{k=1}^{K}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\}>\sigma^2\right)\leq\frac{\delta}{2}.\end{aligned}$$

which, combined with a union bound over $h\in[H]$, has proved (11). We use Freedman's inequality again to prove (12). For a fixed $j\in[m]$, we have

$$\begin{aligned}\left\|\mathrm{Var}_1\left[\nabla_\theta^j X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|=&\left\|\mathbb{E}\left[\left(\nabla_\theta^j X_{\theta,h}^{(k)}\right)\left(\nabla_\theta^j X_{\theta,h}^{(k)}\right)^\top\middle|\mathcal{F}_{h,k}\right]\right\|\leq\mathbb{E}\left[\left\|\left(\nabla_\theta^j X_{\theta,h}^{(k)}\right)\left(\nabla_\theta^j X_{\theta,h}^{(k)}\right)^\top\right\|\middle|\mathcal{F}_{h,k}\right]\\ =&\mathbb{E}\left[\left\|\nabla_\theta^j X_{\theta,h}^{(k)}\right\|^2\middle|\mathcal{F}_{h,k}\right]=\mathrm{Var}_2\left[\nabla_\theta^j X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}_2\left[\nabla_\theta^j X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]=&\mathbb{E}\left[\left\|\nabla_\theta^j X_{\theta,h}^{(k)}\right\|^2\middle|\mathcal{F}_{h,k}\right]=\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right)\mathrm{Var}\left[\nabla_\theta^j\varepsilon_{h,k}^\theta\middle|s_h^{(k)},a_h^{(k)}\right]\\ \leq&4G^2(H-h)^2\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma_h^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right),\end{aligned}$$

where we have used $\mathrm{Var}\left[\nabla_\theta^j\varepsilon_{h,k}^\theta\middle|s_h^{(k)},a_h^{(k)},h\right]\leq 4G^2(H-h)^2$. Furthermore, notice that $\left\|\nabla_\theta^j X_{\theta,h}^{(k)}\right\|\leq 2G(H-h)\sqrt{C_1 d}$, the remaining steps will be exactly the same as those in the proof of (11), combined with a union bound over $j\in[m]$. In this way, we have proved (12). Taking a union bound again finishes the proof. □

B. Proofs of Main Theorems

Define $\widehat{\nu}_h^\theta := \left(\prod_{h'=1}^{h-1} \widehat{M}_{\theta,h'}^\top\right) \nu_1^\theta$. We may prove the following decomposition of $\nabla_\theta v_\theta - \widehat{\nabla_\theta v_\theta}$:

Lemma B.1. *We have $\nabla_\theta v_\theta - \widehat{\nabla_\theta v_\theta} = E_1 + E_2 + E_3$, where*

$$
\begin{aligned}
E_1 =& \sum_{h=1}^{H} \nabla_\theta \left[\left(\nu_h^\theta\right)^\top \Sigma_h^{-1} \frac{1}{K} \sum_{k=1}^{K} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^\theta \right] \\
E_2 =& \sum_{h=1}^{H} \nabla_\theta \left[\left(\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} - \left(\nu_h^\theta\right)^\top \Sigma_h^{-1} \right) \frac{1}{K} \sum_{k=1}^{K} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^\theta \right] \\
E_3 =& \frac{\lambda}{K} \sum_{h=1}^{H} \nabla_\theta \left[\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} w_h^\theta \right].
\end{aligned}
$$

The proof of Lemma B.1 is deferred to appendix C. Based on this observation, here we show the proofs of our main theorems.

B.1. Proof of Theorem 6.2

Proof. We use Lemma B.1 to decompose $\langle \nabla_\theta v_\theta - \widehat{\nabla_\theta v_\theta}, t\rangle = \langle E_1, t\rangle + \langle E_2, t\rangle + \langle E_3, t\rangle$. To bound each term individually, we introduce the following lemmas, whose proofs are deferred to appendix C.

Lemma B.2. *For any $t \in \mathbb{R}^m$, with probability $1-\delta$, we have*

$$
|\langle E_1, t\rangle| \le \sqrt{\frac{2t^\top \Lambda_\theta t \log(2/\delta)}{K}} + \frac{2\log(2/\delta)\sqrt{C_1 md}\|t\|B}{3K}.
$$

where $B = \sum_{h=1}^{H}(H-h+1)\max_{j\in[m]} \sqrt{\left(\nabla_\theta^j \nu_h^\theta\right)^\top \Sigma_h^{-1} \nabla_\theta^j \nu_h^\theta} + 2G\sum_{h=1}^{H}(H-h)^2 \sqrt{\left(\nu_h^\theta\right)^\top \Sigma_h^{-1} \nu_h^\theta}$.

Lemma B.3. *Let E_2^j be the jth entry of E_2, suppose $K \ge 36\kappa_1(4+\kappa_2+\kappa_3)^2 C_1 dH^2 \log\frac{24dmH}{\delta}$ and $\lambda \le C_1 d \min_{h\in[H]} \sigma_{\min}(\Sigma_h) \log\frac{24dmH}{\delta}$, then with probability $1-\delta$,*

$$
|E_2^j| \le 240\sqrt{\kappa_1}(2+\kappa_2+\kappa_3)\sqrt{C_1} dH^3 \left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta\right\| + HG\left\|\Sigma_{\theta,1}^{-\frac{1}{2}} \nu_1^\theta\right\| \right) \max_{h\in[H]} \left\|\Sigma_{\theta,h}^{\frac{1}{2}} \Sigma_h^{-\frac{1}{2}}\right\| \frac{\log\frac{24dmH}{\delta}}{K}, \quad \forall j \in [m].
$$

Lemma B.4. *Let E_3^j be the jth entry of E_3, suppose $K \ge 36\kappa_1(4+\kappa_2+\kappa_3)^2 \log\frac{24dmH}{\delta} C_1 dH^3$ and $\lambda \le C_1 dH \min_{h\in[H]} \sigma_{\min}(\Sigma_h) \log\frac{24dmH}{\delta}$, with probability $1-\delta$,*

$$
|E_3^j| \le 6C_1 dH^2 \max_{h\in[H]} \left\|\Sigma_{\theta,h}^{\frac{1}{2}} \Sigma_h^{-\frac{1}{2}}\right\| \left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta\right\| + HG\left\|\Sigma_{\theta,1}^{-\frac{1}{2}} \nu_1^\theta\right\| \right) \frac{\log\frac{24dmH}{\delta}}{K}, \quad \forall j \in [m].
$$

Let $B_1^j = \sum_{h=1}^{H}(H-h+1)\sqrt{\left(\nabla_\theta^j \nu_h^\theta\right)^\top \Sigma_h^{-1} \nabla_\theta^j \nu_h^\theta}$, $B_2 = \sum_{h=1}^{H}(H-h)^2 G\sqrt{\left(\nu_h^\theta\right)^\top \Sigma_h^{-1} \nu_h^\theta}$, then we have the relation

$B = \max_{j\in[m]} B_1^j + 2B_2$. For any $j \in [m]$, note that

$$
\begin{aligned}
B_1^j = &\sum_{h=1}^{H}(H-h+1)\left\|\Sigma_h^{-\frac{1}{2}}\nabla_\theta^j \nu_h^\theta\right\| \le \sum_{h=1}^{H} H\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left\|\Sigma_{\theta,h}^{-\frac{1}{2}}\nabla_\theta^j\left(\left(\prod_{h'=1}^{h-1} M_{\theta,h'}\right)^\top \nu_1^\theta\right)\right\| \\
= &\sum_{h=1}^{H} H\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\left(\prod_{h'=1}^{h-1}\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}M_{\theta,h'}\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right\|\right)\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\right. \\
&\left.+\sum_{h'=1}^{h-1}\left(\prod_{h''\neq h'}\left\|\Sigma_{\theta,h''}^{\frac{1}{2}}M_{\theta,h''}\Sigma_{\theta,h''+1}^{-\frac{1}{2}}\right\|\right)\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}\left(\nabla_\theta^j M_{\theta,h'}\right)\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right\|\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\right) \\
\le &\sum_{h=1}^{H} H\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\| + HG\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\right) \le H^2\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\| + HG\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\right),
\end{aligned}
$$

where we use the result of Lemma A.5. Similarly,

$$
\begin{aligned}
B_2 = &\sum_{h=1}^{H}(H-h)^2 G\left\|\Sigma_h^{-\frac{1}{2}}\nu_h^\theta\right\| \le \sum_{h=1}^{H} H^2 G\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\prod_{h'=1}^{h-1}\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}M_{\theta,h'}\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right\|\right)\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\| \\
\le &\sum_{h=1}^{H} H^2 G\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\| \le H^3 G\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|.
\end{aligned}
$$

We conclude that when $K \ge 36C_1 dH^2\kappa_1(4+\kappa_2+\kappa_3)^2\log\frac{24dmH}{\delta}$, we have

$$
\begin{aligned}
&\left(\max_{j\in[m]} B_1^j + 2B_2\right)\frac{2\log\frac{2}{\delta}\sqrt{C_1 md}\|t\|}{K} \\
\le &H^2\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\max_{j\in[m]}\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\| + 2HG\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\right)\frac{2\log\frac{2}{\delta}\sqrt{C_1 md}\|t\|}{K},
\end{aligned}
$$

and therefore, with probability $1-3\delta$, we have

$$
\begin{aligned}
&|\langle E_1, t\rangle| + |\langle E_2, t\rangle| + |\langle E_3, t\rangle| \le \sqrt{\frac{2t^\top\Lambda_\theta t\log(2/\delta)}{K}} \\
&+\sqrt{\kappa_1}(5+\kappa_2+\kappa_3)\left(\max_{j\in[m]}\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\| + HG\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\right)\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\frac{240C_1 dH^3\log\frac{24dmH}{\delta}}{K} \\
\le&\sqrt{\frac{2t^\top\Lambda_\theta t\log(2/\delta)}{K}} + \kappa_1(5+\kappa_2+\kappa_3)\left(\max_{j\in[m]}\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\| + HG\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\right)\frac{240C_1 dH^3\sqrt{m}\|t\|\log\frac{24dmH}{\delta}}{K}.
\end{aligned}
$$

replacing δ by $\frac{\delta}{3}$, we have finished the proof. ☐

B.2. Proof of Theorem 6.3

Proof. According to the result of Theorem 6.2, we know

$$
|\langle t, \widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta\rangle| \le \sqrt{\frac{2t^\top\Lambda_\theta t}{K}\cdot\log\frac{8}{\delta}} + \frac{C_\theta\|t\|\log\frac{72mdH}{\delta}}{K},
$$

Pick $t = e_j, j \in [m]$, we have

$$\begin{aligned}
t^\top \Lambda_\theta t =& \mathbb{E}\left[\sum_{h=1}^H \left(\nabla_\theta^j \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta\right)\right)^2\right] \\
\leq& 2\mathbb{E}\left[\sum_{h=1}^H \left(\nabla_\theta^j \varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta\right)^2\right] + 2\mathbb{E}\left[\sum_{h=1}^H \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nabla_\theta^j \nu_h^\theta\right)^2\right] \\
\leq& 2\mathbb{E}\left[\sum_{h=1}^H G^2 (H-h)^4 \left(\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta\right)^2\right] + 2\mathbb{E}\left[\sum_{h=1}^H (H-h+1)^2 \left(\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nabla_\theta^j \nu_h^\theta\right)^2\right] \\
\leq& 2\sum_{h=1}^H \left((H-h)^4 G^2 \left\|\Sigma_h^{-\frac{1}{2}} \nu_h^\theta\right\|^2 + (H-h+1)^2 \left\|\Sigma_h^{-\frac{1}{2}} \nabla_\theta^j \nu_h^\theta\right\|^2\right).
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
t^\top \Lambda_\theta t =& \mathbb{E}\left[\sum_{h=1}^H \left(\nabla_\theta^j \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta\right)\right)^2\right] \\
\leq& 2\mathbb{E}\left[\sum_{h=1}^H \left(\nabla_\theta^j \varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta\right)^2\right] + 2\mathbb{E}\left[\sum_{h=1}^H \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nabla_\theta^j \nu_h^\theta\right)^2\right] \\
\leq& 2C_1 d\mathbb{E}\left[\sum_{h=1}^H \left(\nabla_\theta^j \varepsilon_{h,1}^\theta\right)^2\right] \max_{h\in[H]} \left\|\Sigma_h^{-\frac{1}{2}} \nu_h^\theta\right\|^2 + 2C_1 d\mathbb{E}\left[\sum_{h=1}^H \left(\varepsilon_{h,1}^\theta\right)^2\right] \max_{h\in[H]} \left\|\Sigma_h^{-\frac{1}{2}} \nabla_\theta^j \nu_h^\theta\right\|^2.
\end{aligned}$$

Define

$$\begin{aligned}
\zeta_h &= \nabla_\theta^j Q_h^\theta\left(s_h^{(1)}, a_h^{(1)}\right) - \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a \middle| s_{h+1}^{(1)}\right)\left(\nabla_\theta^j Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a\right) + Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a\right)\nabla_\theta^j \log \pi_{\theta,h+1}\left(a \middle| s_{h+1}^{(1)}\right)\right) \mathrm{d}a \\
\eta_h &= \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a \middle| s_{h+1}^{(1)}\right)\left(\nabla_\theta^j Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a\right) + Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a\right)\nabla_\theta^j \log \pi_{\theta,h+1}\left(a \middle| s_{h+1}^{(1)}\right)\right) \mathrm{d}a \\
&\quad - \nabla_\theta^j Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a_{h+1}^{(1)}\right) - Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a_{h+1}^{(1)}\right)\nabla_\theta^j \log \pi_{\theta,h+1}\left(a_{h+1}^{(1)} \middle| s_{h+1}^{(1)}\right).
\end{aligned}$$

Note that the sequence $\zeta_1, \eta_1, \zeta_2, \eta_2 \ldots, \zeta_H, \eta_H$ forms a martingale difference sequence, therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{h=1}^H \left(\nabla_\theta^j \varepsilon_{h,1}^\theta\right)^2\right] =& \mathbb{E}\left[\sum_{h=1}^H \zeta_h^2\right] \leq \mathbb{E}\left[\sum_{h=1}^H (\zeta_h^2 + \eta_h^2)\right] = \mathbb{E}\left[\left(\sum_{h=1}^H (\zeta_h + \eta_h)\right)^2\right] \\
=& \mathbb{E}\left[\left(\nabla_\theta Q_1^\theta\left(s_1^{(1)}, a_1^{(1)}\right) - \sum_{h=1}^H Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a_{h+1}^{(1)}\right)\nabla_\theta \log \pi_{\theta,h+1}\left(a_{h+1}^{(1)} \middle| s_{h+1}^{(1)}\right)\right)^2\right] \\
\leq& 4H^4 G^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[\sum_{h=1}^H \left(\varepsilon_{h,1}^\theta\right)^2\right] =& \mathbb{E}\left[\sum_{h=1}^H \left(Q_h^\theta\left(s_h^{(1)}, a_h^{(1)}\right) - r_h^{(1)} - \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a \middle| s_{h+1}^{(1)}\right) Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a\right) \mathrm{d}a\right)^2\right] \\
&+ \mathbb{E}\left[\sum_{h=1}^H \left(\int_{\mathcal{A}} \pi_{\theta,h+1}\left(a \middle| s_{h+1}^{(1)}\right) Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a\right) \mathrm{d}a - Q_{h+1}^\theta\left(s_{h+1}^{(1)}, a_{h+1}^{(1)}\right)\right)^2\right] \\
=& \mathbb{E}\left[\left(Q_1^\theta\left(s_1^{(1)}, a_1^{(1)}\right) - \sum_{h=1}^H r_h^{(1)}\right)^2\right] \leq H^2.
\end{aligned}$$

Therefore,

$$t^\top \Lambda_\theta t \leq 8C_1 dH^4 G^2 \max_{h\in[H]} \left\| \Sigma_h^{-\frac{1}{2}} \nu_h^\theta \right\|^2 + 2C_1 dH^2 \max_{h\in[H]} \left\| \Sigma_h^{-\frac{1}{2}} \nabla_\theta^j \nu_h^\theta \right\|^2.$$

Therefore, taking a union bound over j, we get

$$\left| \widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta \right| \leq 4b_\theta \sqrt{\frac{\min\{C_1 d, H\} \log \frac{8m}{\delta}}{K}} + \frac{2C_\theta \sqrt{m} \log \frac{72mdH}{\delta}}{K}, \quad \forall j \in [m],$$

where

$$b_\theta = H^2 G \max_{h\in[H]} \left\| \Sigma_h^{-\frac{1}{2}} \nu_h^\theta \right\| + H \max_{h\in[H]} \left\| \Sigma_h^{-\frac{1}{2}} \nabla_\theta^j \nu_h^\theta \right\|.$$

When we in addition have $\phi(s', a')^\top \Sigma_h^{-1} \phi(s,a) \geq 0, \forall h \in [H], (s,a), (s',a') \in \mathcal{S}\times\mathcal{A}$, we have for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
\left| \left(\nabla_\theta^j \nu_h^\theta\right)^\top \Sigma_h^{-1} \phi(s,a) \right| =& \left| \mathbb{E}^{\pi_\theta} \left[\phi(s_h, a_h) \Sigma_h^{-1} \phi(s,a) \sum_{h'=1}^{h} \nabla_\theta^j \log \pi_{\theta,h'}(a_{h'}|s_{h'}) \right] \right| \\
\leq& \mathbb{E}^{\pi_\theta} \left[\phi(s_h, a_h) \Sigma_h^{-1} \phi(s,a) \sum_{h'=1}^{h} \left| \nabla_\theta^j \log \pi_{\theta,h'}(a_{h'}|s_{h'}) \right| \right] \\
\leq& Gh \mathbb{E}^{\pi_\theta} \left[\phi(s_h, a_h) \Sigma_h^{-1} \phi(s,a) \right] \\
=& Gh \left(\nu_h^\theta\right)^\top \Sigma_h^{-1} \phi(s,a),
\end{aligned}$$

which implies

$$\left(\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nabla_\theta^j \nu_h^\theta \right)^2 \leq G^2 h^2 \left(\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta \right)^2.$$

Therefore, we get

$$\begin{aligned}
t^\top \Lambda_\theta t =& \mathbb{E}\left[\sum_{h=1}^{H} \left(\nabla_\theta^j \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta \right) \right)^2 \right] \\
\leq& 2\mathbb{E}\left[\sum_{h=1}^{H} \left(\nabla_\theta^j \varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta \right)^2 \right] + 2\mathbb{E}\left[\sum_{h=1}^{H} \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nabla_\theta^j \nu_h^\theta \right)^2 \right] \\
\leq& 2\mathbb{E}\left[\sum_{h=1}^{H} G^2 (H-h)^4 \left(\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta \right)^2 \right] + 2\mathbb{E}\left[\sum_{h=1}^{H} G^2 h^2 (H-h+1)^2 \left(\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta \right)^2 \right] \\
\leq& 2H^2 G^2 \sum_{h=1}^{H} (H-h+1)^2 \left\| \Sigma_h^{-\frac{1}{2}} \nu_h^\theta \right\|^2,
\end{aligned}$$

and

$$\begin{aligned}
t^\top \Lambda_\theta t \leq& 2\mathbb{E}\left[\sum_{h=1}^{H} \left(\nabla_\theta^j \varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nu_h^\theta \right)^2 \right] + 2\mathbb{E}\left[\sum_{h=1}^{H} \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top \Sigma_h^{-1} \nabla_\theta^j \nu_h^\theta \right)^2 \right] \\
\leq& 2C_1 d \mathbb{E}\left[\sum_{h=1}^{H} \left(\nabla_\theta^j \varepsilon_{h,1}^\theta \right)^2 \right] \max_{h\in[H]} \left\| \Sigma_h^{-\frac{1}{2}} \nu_h^\theta \right\|^2 + 2C_1 d G^2 H^2 \mathbb{E}\left[\sum_{h=1}^{H} \left(\varepsilon_{h,1}^\theta \right)^2 \right] \max_{h\in[H]} \left\| \Sigma_h^{-\frac{1}{2}} \nu_h^\theta \right\|^2
\end{aligned}$$

Repeating the steps that we bound $\mathbb{E}\left[\sum_{h=1}^{H} \left(\varepsilon_{h,1}^\theta \right)^2 \right]$ and $\mathbb{E}\left[\sum_{h=1}^{H} \left(\nabla_\theta^j \varepsilon_{h,1}^\theta \right)^2 \right]$, we get

$$\left| \widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta \right| \leq 4H^2 G \sqrt{\frac{\min\{C_1 d, H\} \log \frac{8m}{\delta}}{K}} \max_{h\in[H]} \left\| \Sigma_h^{-\frac{1}{2}} \nu_h^\theta \right\| + \frac{2C_\theta \log \frac{72mdH}{\delta}}{K}, \quad \forall j \in [m].$$

□

B.3. Proof of Theorem 6.4

Proof. When the MDP is tabular and ϕ is the one-hot vector, we have

$$\nu_{h,s,a}^{\theta} = \mu_{\theta,h}(s,a), \quad \Sigma_h = \text{diag}(\bar{\mu}_h(s,a))$$

which implies

$$\begin{aligned}
t^\top \Lambda_\theta t =& \mathbb{E}\left[\sum_{h=1}^{H}\left(\nabla_\theta^j\left(\varepsilon_{h,1}^{\theta}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top \Sigma_h^{-1}\nu_h^{\theta}\right)\right)^2\right] \\
\leq & 2\mathbb{E}\left[\sum_{h=1}^{H}\left(\nabla_\theta^j\varepsilon_{h,1}^{\theta}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top \Sigma_h^{-1}\nu_h^{\theta}\right)^2\right] + 2\mathbb{E}\left[\sum_{h=1}^{H}\left(\varepsilon_{h,1}^{\theta}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top \Sigma_h^{-1}\nabla_\theta^j\nu_h^{\theta}\right)^2\right] \\
=& 2\mathbb{E}\left[\sum_{h=1}^{H}\left(\nabla_\theta^j\varepsilon_{h,1}^{\theta}\right)^2\left(\frac{\mu_{\theta,h}(s,a)}{\bar{\mu}_h(s,a)}\right)^2\right] + 2G^2H^2\mathbb{E}\left[\sum_{h=1}^{H}\left(\varepsilon_{h,1}^{\theta}\right)^2\left(\frac{\mu_{\theta,h}(s,a)}{\bar{\mu}_h(s,a)}\right)^2\right] \\
=& 2\mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{H}\left(\nabla_\theta^j\varepsilon_{h,1}^{\theta}\right)^2\frac{\mu_{\theta,h}(s,a)}{\bar{\mu}_h(s,a)}\right] + 2G^2H^2\mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{H}\left(\varepsilon_{h,1}^{\theta}\right)^2\frac{\mu_{\theta,h}(s,a)}{\bar{\mu}_h(s,a)}\right] \\
\leq & 2\mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{H}\left(\nabla_\theta^j\varepsilon_{h,1}^{\theta}\right)^2\right]\max_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}\frac{\mu_{\theta,h}(s,a)}{\bar{\mu}_h(s,a)} + 2G^2H^2\mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{H}\left(\varepsilon_{h,1}^{\theta}\right)^2\right]\max_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}\frac{\mu_{\theta,h}(s,a)}{\bar{\mu}_h(s,a)}.
\end{aligned}$$

Following the same argument above, we can derive

$$t^\top \Lambda_\theta t \leq 4H^4G^2\max_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}\frac{\mu_h^{\theta}(s,a)}{\bar{\mu}_h(s,a)},$$

i.e.,

$$\left|\widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta\right| \leq 4H^2G\sqrt{\frac{\log\frac{8m}{\delta}}{K}}\max_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}\frac{\mu_h^{\theta}(s,a)}{\bar{\mu}_h(s,a)} + \frac{2C_\theta\log\frac{72mdH}{\delta}}{K}, \quad \forall j\in[m].$$

On the other hand, the result of Theorem 6.3 implies

$$\left|\widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta\right| \leq 4H^2G\sqrt{\frac{\min\{C_1d,H\}\log\frac{8m}{\delta}}{K}}\max_{h\in[H]}\left\|\Sigma_h^{-\frac{1}{2}}\nu_h^{\theta}\right\| + \frac{2C_\theta\log\frac{72mdH}{\delta}}{K}, \quad \forall j\in[m].$$

Using the relation $\left\|\Sigma_h^{-\frac{1}{2}}\nu_h^{\theta}\right\| = \sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\left(\mu_h^{\theta}(s,a)\right)^2}{\bar{\mu}_h(s,a)}} = \sqrt{\mathbb{E}^{\pi_\theta}\frac{\mu_h^{\theta}(s_h,a_h)}{\bar{\mu}(s_h,a_h)}}$, and taking minimum over the above two inequalities, we have finished the proof. □

B.4. Proof of Theorem 6.5

Proof. We use the same decomposition as in Theorem 6.2. Define a martingale difference sequence $\{e_k^{\theta}\}_{k=1}^K$ by

$$\begin{aligned}
e_k^{\theta} =& \frac{1}{\sqrt{K}}\sum_{h=1}^{H}\nabla_\theta\left(\left(\nu_h^{\theta}\right)^\top\Sigma_h^{-1}\phi(s_h^{(k)},a_h^{(k)})\varepsilon_{h,k}^{\theta}\right) \\
=& \frac{1}{\sqrt{K}}\sum_{h=1}^{H}\left(\nabla_\theta\nu_h^{\theta}\right)^\top\Sigma_h^{-1}\phi(s_h^{(k)},a_h^{(k)})\varepsilon_{h,k}^{\theta} + \frac{1}{\sqrt{K}}\sum_{h=1}^{H}\left(\nu_h^{\theta}\right)^\top\Sigma_h^{-1}\phi(s_h^{(k)},a_h^{(k)})\nabla_\theta\varepsilon_{h,k}^{\theta},
\end{aligned}$$

we have

$$\|e_k^{\theta}\|_\infty \leq \frac{1}{\sqrt{K}}\sum_{h=1}^{H}\max_{j\in[m]}\|\Sigma_h^{-\frac{1}{2}}\nabla_\theta^j\nu_h^{\theta}\|\sqrt{C_1d}(H-h+1) + \frac{2}{\sqrt{K}}\sum_{h=1}^{H}\|\Sigma_h^{-\frac{1}{2}}\nu_h^{\theta}\|\sqrt{C_1d}(H-h)^2G \to 0,$$

where we use the result of Lemma A.2. Furthermore,

$$\mathbb{E}\left[e_k^{\theta}\left(e_k^{\theta}\right)^{\top}\right]_{ij} = \frac{1}{K}\mathbb{E}\left[\sum_{h=1}^{H}\left[\nabla_{\theta_1}^{i}\left(\left(\nu_h^{\theta_1}\right)^{\top}\Sigma_h^{-1}\phi(s_h^{(k)},a_h^{(k)})\varepsilon_{h,k}^{\theta_1}\right)\right]\left[\nabla_{\theta_2}^{j}\left(\left(\nu_h^{\theta_2}\right)^{\top}\Sigma_h^{-1}\phi(s_h^{(k)},a_h^{(k)})\varepsilon_{h,k}^{\theta_2}\right)\right]^{\top}\right]\Bigg|_{\theta_1=\theta_2=\theta}$$
$$= \frac{[\Lambda_\theta]_{ij}}{K}.$$

Therefore,by WLLN, we have

$$\sum_{k=1}^{K}\left[e_k^{\theta}\left(e_k^{\theta}\right)^{\top}\right]_{ij} \to_p \sum_{k=1}^{K}\mathbb{E}\left[e_k^{\theta}\left(e_k^{\theta}\right)^{\top}\right]_{ij} = [\Lambda_\theta]_{ij},$$

To finish the rest of the proof, we introduce the following lemmas,

Lemma B.5 (Martingale CLT, Corollary 2.8 in (McLeish et al., 1974))**.** *Let $\{X_{mn}, n = 1, \dots, k_m\}$ be a martingale difference array (row-wise) on the probability triple $(\Omega, \mathcal{F}, P)$.Suppose X_{mn} satisfy the following two conditions:*

$$\max_{1\le n\le k_m}|X_{mn}| \xrightarrow{p} 0, \text{ and } \sum_{n=1}^{k_m} X_{mn}^2 \xrightarrow{p} \sigma^2$$

for $k_m \to \infty$. Then $\sum_{n=1}^{k_m} X_{mn} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$.

Lemma B.6 (Cramér–Wold Theorem)**.** *Let $X_n = (X_n^1, X_n^2, \dots, X_n^k)^{\top}$ be a k-dimensional random vector series and $X = (X^1, X^2, \dots, X^k)^{\top}$ be a random vector of same dimension. Then X_n converges in distribution to X if and only if for any constant vector $t = (t_1, t_2, \dots, t_k)^{\top}$, $t^{\top}X_n$ converges to $t^{\top}X$ in distribution.*

Lemma B.5 implies $\sum_{k=1}^{K} t^{\top}e_k \to_d \mathcal{N}(0, t^{\top}\Lambda_\theta t)$ for any t, and Lemma B.6 implies

$$\sqrt{K}E_1 = \sum_{k=1}^{K} e_k \to_p \mathcal{N}(0, \Lambda_\theta).$$

Furthermore, notice that the results of Lemma B.3 and Lemma B.4 imply $\sqrt{K}E_2 \to_p 0, \sqrt{K}E_3 \to_p 0$. Combining the above results, we have finished the proof. □

B.5. Proof of Theorem 6.6

Proof. Our proof is similar to that of (Hao et al., 2021). We first derive the influence function of policy gradient estimator for sake of completeness. We denote each of the K sampled trajectories as

$$\boldsymbol{\tau} := (s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_H, a_H, r_H, s_{H+1})$$

We denote $\bar{\pi}(a \mid s)$ as the behavior policy. The distribution of trajectory is then given by

$$\mathcal{P}(d\boldsymbol{\tau}) = \bar{\xi}\left(ds_1, da_1\right)p_1\left(ds_2 \mid s_1, a_1\right)\bar{\pi}_2\left(da_2 \mid s_2\right)\dots\bar{\pi}_H\left(da_H \mid s_H\right)p_H\left(ds_{H+1} \mid s_H, a_H\right)$$

Define $p_\eta = p + \eta\Delta p$ as a new transition probability function and $\mathcal{P}_\eta := \mathcal{P} + \eta\Delta\mathcal{P}$ where $\Delta\mathcal{P}$ satisfies

$$(\Delta\mathcal{P}_h)\mathcal{F} \subseteq \mathcal{F}, \forall h \in [H].$$

Define $g_{\eta,h}\left(s' \mid s, a\right) := \frac{\partial}{\partial\eta}\log p_{\eta,h}\left(s' \mid s, a\right)$ and the score function as

$$g_\eta(\boldsymbol{\tau}) := \frac{\partial}{\partial\eta}\log\mathcal{P}_\eta(d\boldsymbol{\tau}) = \sum_{h=1}^{H} g_{\eta,h}\left(s_{h+1} \mid s_h, a_h\right).$$

Without loss of generality, we assume p_η is continuously derivative with respect to η. This guarantees that we can change the order of taking derivatives with respect to η and θ. When the subscript η vanishes, it means $\eta = 0$ and the underlying transition probability is $p(s'|s,a)$, i.e. $p_0(s'|s,a) = p(s'|s,a)$. Then we denote $g_h(s'|s,a) := \frac{\partial}{\partial \eta} \log p_{\eta,h}(s'|s,a)\Big|_{\eta=0}$, and $g(\boldsymbol{\tau}) = \sum_{h=1}^{H} g_h(s_{h+1}|s_h,a_h)$. We define the policy value under new transition kernel is

$$v_{\theta,\eta} := \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{H} r_h(s_h,a_h)\middle| s_1 \sim \xi, \mathcal{P}_\eta\right]$$

Then, our objective function is

$$\psi_\eta := \nabla_\theta v_{\theta,\eta} = \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{H} \nabla_\theta \log \pi_{\theta,h}\left(a_h \mid s_h\right) \cdot \left(\sum_{h'=h}^{H} r_h\left(s_{h'},a_{h'}\right)\right)\middle| s_1 \sim \xi, \mathcal{P}_\eta\right].$$

We are going to compute the influence function with respect to the above objective function. We denote this influence function as $\mathcal{I}(\boldsymbol{\tau})$. By definition, it satisfies that

$$\frac{\partial}{\partial \eta}\psi_\eta\Big|_{\eta=0} = \mathbb{E}\left[g(\boldsymbol{\tau})\mathcal{I}(\boldsymbol{\tau})\right].$$

By exchanging the order of derivatives, we find that

$$\frac{\partial}{\partial \eta}\psi_\eta\Big|_{\eta=0} = \nabla_\theta\left[\frac{\partial}{\partial \eta}v_{\theta,\eta}\Big|_{\eta=0}\right].$$

Therefore, we calculate the derivatives.

$$\begin{aligned}
\frac{\partial}{\partial \eta}v_{\theta,\eta} &= \frac{\partial}{\partial \eta}\left[\sum_{h=1}^{H}\int_{(\mathcal{S}\times\mathcal{A})^h} r_h\left(s_h,a_h\right)\xi(s_1)\prod_{j=1}^{h-1} p_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)\prod_{j=1}^{h}\pi_{\theta,j}\left(a_j \mid s_j\right)d\boldsymbol{\tau}_h\right] \\
&= \sum_{h=1}^{H}\int_{(\mathcal{S}\times\mathcal{A})^h} r_h\left(s_h,a_h\right)\left(\sum_{j=1}^{h-1} g_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)\right)\xi(s_1)\prod_{j=1}^{h-1} p_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)\prod_{j=1}^{h}\pi_{\theta,j}\left(a_j \mid s_j\right)d\boldsymbol{\tau}_h \\
&= \int_{(\mathcal{S}\times\mathcal{A})^H}\sum_{h=1}^{H} r_h\left(s_h,a_h\right)\left(\sum_{j=1}^{h-1} g_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)\right)\left[\xi(s_1)\prod_{j=1}^{H} p_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)\prod_{j=1}^{H}\pi_{\theta,j}\left(a_j \mid s_j\right)\right]d\boldsymbol{\tau}.
\end{aligned}$$

We denote $Q_{h,\eta}^{\theta}$ and $\nabla_\theta Q_{h,\eta}^{\theta}$ as the state-action function and its gradient with underlying transition probability being p_η. For sake of simplicity, we define the state value function as

$$V_h^\theta(s) := \mathbb{E}^{\pi_\theta}\left[\sum_{h'=h}^{H} r_h(s_h,a_h)\middle| s_h = s, \mathcal{P}\right].$$

We denote $V_{h,\eta}^{\theta}(s)$ as the same function except for transition probability substituted by p_η. Therefore,

$$\begin{aligned}
\frac{\partial}{\partial \eta}v_{\theta,\eta} &= \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{H} r_h\left(s_h,a_h\right)\left(\sum_{j=1}^{h-1} g_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)\right)\middle| s_1 \sim \xi, \mathcal{P}_\eta\right] \\
&= \mathbb{E}^{\pi_\theta}\left[\sum_{j=1}^{H} g_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)\sum_{h=j+1}^{H} r_h\left(s_h,a_h\right)\middle| s_1 \sim \xi, \mathcal{P}_\eta\right] \\
&= \mathbb{E}^{\pi_\theta}\left[\sum_{j=1}^{H} g_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)\cdot\mathbb{E}^{\pi_\theta}\left[\sum_{h=j+1}^{H} r_h\left(s_h,a_h\right)\middle| s_{j+1}\right]\middle| s_1 \sim \xi, \mathcal{P}_\eta\right] \\
&= \mathbb{E}^{\pi_\theta}\left[\sum_{j=1}^{H}\mathbb{E}\left[g_{\eta,j}\left(s_{j+1} \mid s_j,a_j\right)V_{j+1,\eta}^{\theta}(s_{j+1})\middle| s_j,a_j\right]\middle| s_1 \sim \xi, \mathcal{P}_\eta\right].
\end{aligned}$$

Therefore,

$$\frac{\partial}{\partial \eta} v_{\theta,\eta}\bigg|_{\eta=0} = \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{H} \mathbb{E}\left[g_h\left(s_{h+1} \mid s_h, a_h\right) V_{h+1}^{\theta}(s_{h+1})\big| s_h, a_h\right]\bigg| s_1 \sim \xi, \mathcal{P}_\eta\right]. \tag{16}$$

We notice that $\Sigma_h = \mathbb{E}\left[\phi(s_h^{(1)}, a_h^{(1)})\phi(s_h^{(1)}, a_h^{(1)})^\top\right]$. We denote $w_h(s,a) := \phi^\top(s,a)\Sigma_h^{-1}\nu_h^\theta = \phi^\top(s,a)\Sigma_h^{-1}\mathbb{E}^{\pi_\theta}\left[\phi(s_h,a_h) \mid s_1 \sim \xi\right]$. We leverage the following fact to rewrite (16): for any $f(s,a) = w_f^\top \phi(s,a) \in \mathcal{F}$ where $w_f \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbb{E}^{\pi_\theta}\left[f(s_h,a_h)\right] &= \mathbb{E}^{\pi_\theta}\left[w_f^\top \phi(s_h,a_h)\right] \\
&= \mathbb{E}^{\pi_\theta}\left[w_f^\top \mathbb{E}\left[\phi(s_h^{(1)}, a_h^{(1)})\phi^\top(s_h^{(1)}, a_h^{(1)})\right]\Sigma_h^{-1}\phi(s_h,a_h)\right] \\
&= \mathbb{E}\left[w_f^\top \phi(s_h^{(1)}, a_h^{(1)})\phi^\top(s_h^{(1)}, a_h^{(1)})\Sigma_h^{-1}\mathbb{E}^{\pi_\theta}\left[\phi(s_h,a_h)\right]\right] \\
&= \mathbb{E}\left[f(s_h^{(1)}, a_h^{(1)}) w_h(s_h^{(1)}, a_h^{(1)})\right]
\end{aligned}$$

Since

$$\mathbb{E}\left[g_h\left(s' \mid s,a\right) V_{h+1}^\theta(s') | s,a\right] = \frac{\partial}{\partial \eta}\left(Q_{h,\eta}^\theta(s,a) - r_\eta(s,a)\right)\bigg|_{\eta=0} \in \mathcal{F},$$

we have

$$\begin{aligned}
\frac{\partial}{\partial \eta} v_{\theta,\eta}\bigg|_{\eta=0} &= \mathbb{E}\left[\sum_{h=1}^{H} w_h(s_h^{(1)}, a_h^{(1)})\mathbb{E}\left[g_h\left(s' \mid s_h^{(1)}, a_h^{(1)}\right)\cdot V_{h+1}^\theta\left(s'\right) \mid s_h^{(1)}, a_h^{(1)}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_{s' \sim p(\cdot|s_h^{(1)}, a_h^{(1)})}\left[\sum_{h=1}^{H} w_h(s_h^{(1)}, a_h^{(1)}) g_h\left(s' \mid s_h^{(1)}, a_h^{(1)}\right)\cdot V_{h+1}^\theta\left(s'\right)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_{s' \sim p(\cdot|s_h^{(1)}, a_h^{(1)})}\left[\sum_{h=1}^{H} w_h(s_h^{(1)}, a_h^{(1)}) g_h\left(s' \mid s_h^{(1)}, a_h^{(1)}\right)\left(V_{h+1}^\theta\left(s'\right) - \mathbb{E}\left[V_{h+1}^\theta\left(s'\right) \mid s_h^{(1)}, a_h^{(1)}\right]\right)\right]\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H} w_h(s_h^{(1)}, a_h^{(1)}) g_h\left(s_{h+1}^{(1)} \mid s_h^{(1)}, a_h^{(1)}\right)\left(V_{h+1}^\theta\left(s_{h+1}^{(1)}\right) - \mathbb{E}\left[V_{h+1}^\theta\left(s_{h+1}^{(1)}\right) \mid s_h^{(1)}, a_h^{(1)}\right]\right)\right] \\
&= \mathbb{E}\left[g\left(\boldsymbol{\tau}\right)\sum_{h=1}^{H} w_h(s_h^{(1)}, a_h^{(1)})\left(V_{h+1}^\theta\left(s_{h+1}^{(1)}\right) - \mathbb{E}\left[V_{h+1}^\theta\left(s_{h+1}^{(1)}\right) \mid s_h^{(1)}, a_h^{(1)}\right]\right)\right].
\end{aligned}$$

Taking gradient in both sides and we have

$$\nabla_\theta\left(\frac{\partial}{\partial \eta} v_{\theta,\eta}\bigg|_{\eta=0}\right) = \mathbb{E}\left\{g\left(\boldsymbol{\tau}\right)\cdot \nabla_\theta\left[\sum_{h=1}^{H} w_h(s_h^{(1)}, a_h^{(1)})\left(V_{h+1}^\theta\left(s_{h+1}^{(1)}\right) - \mathbb{E}\left[V_{h+1}^\theta\left(s_{h+1}^{(1)}\right) \mid s_h^{(1)}, a_h^{(1)}\right]\right)\right]\right\}.$$

The implies that the influence function we want is

$$\mathcal{I}(\boldsymbol{\tau}) = \nabla_\theta\left[\sum_{h=1}^{H} w_h(s_h^{(1)}, a_h^{(1)})\left(V_{h+1}^\theta\left(s_{h+1}^{(1)}\right) - \mathbb{E}\left[V_{h+1}^\theta\left(s_{h+1}^{(1)}\right) \mid s_h^{(1)}, a_h^{(1)}\right]\right)\right].$$

Insert the expression of $w_h(s,a)$ and exploit $\varepsilon_{h,k}^\theta = Q_h^\theta(s_h^{(k)}, a_h^{(k)}) - r_h^{(k)} - \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \mid s_{h+1}^{(k)}\right) Q_{h+1}^\theta\left(s_{h+1}^{(k)}, a'\right) \mathrm{d}a'$, we can rewrite the influence function as

$$\mathcal{I}(\boldsymbol{\tau}) = -\nabla_\theta\left[\sum_{h=1}^{H} \phi(s_h^{(1)}, a_h^{(1)})^\top \Sigma_h^{-1}\varepsilon_{h,1}^\theta \nu_h^\theta\right]$$

Therefore, since the cross terms vanish by taking conditional expectation, we have

$$\mathbb{E}\left[\mathcal{I}(\boldsymbol{\tau})^\top \mathcal{I}(\boldsymbol{\tau})\right] = \mathbb{E}\left[\sum_{h=1}^{H}\left(\nabla_\theta\left(\varepsilon_{h,1}^\theta \phi(s_h^{(1)}, a_h^{(1)})^\top \Sigma_h^{-1}\nu_h^\theta\right)\right)^\top \nabla_\theta\left(\varepsilon_{h,1}^\theta \phi(s_h^{(1)}, a_h^{(1)})^\top \Sigma_h^{-1}\nu_h^\theta\right)\right] = \Lambda_\theta.$$

For any vector $t \in \mathbb{R}^m$, when it comes to $\langle t, \psi_\eta \rangle$, by linearity we have

$$\frac{\partial}{\partial \eta} \langle t, \psi_\eta \rangle \Big|_{\eta=0} = \mathbb{E}[g(\boldsymbol{\tau}) \langle t, \mathcal{I}(\boldsymbol{\tau}) \rangle].$$

Then the influence function of $\langle t, \nabla_\theta v_\theta \rangle$ is $\langle t, \mathcal{I}(\boldsymbol{\tau}) \rangle$. The Cramer-Rao lower bound for $\langle t, \nabla_\theta v_\theta \rangle$ is

$$\mathbb{E}\left[\langle t, \mathcal{I}(\boldsymbol{\tau}) \rangle^2 \right] = t^\top \mathbb{E}\left[\mathcal{I}(\boldsymbol{\tau})^\top \mathcal{I}(\boldsymbol{\tau}) \right] t = t^\top \Lambda_\theta t.$$

By continuous mapping theorem, a trivial corollary of Theorem 6.6 is that for any $t \in \mathbb{R}^m$,

$$\sqrt{K} \left(\left\langle t, \widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta \right\rangle \right) \xrightarrow{d} \mathcal{N}\left(0, t^\top \Lambda_\theta t\right).$$

This implies that the variance of any unbiased estimator for $\langle t, \nabla_\theta v_\theta \rangle \in \mathbb{R}$ is lower bounded by $\frac{1}{\sqrt{K}} t^\top \Lambda_\theta t$. □

C. Missing Proofs

C.1. Proof of Proposition 4.3

Proof. The differentiability of w_h^θ comes from the differentiability of Q_h^θ. And simply taking derivatives w.r.t. θ on both sides of $Q_h^\theta = \phi^\top w_h^\theta$, we get the desired result. □

C.2. Proof of Proposition 5.1

Proof. To prove the equality, it suffices to prove that given the same input $\widehat{Q}_{h+1}^\theta, \nabla_\theta^j \widehat{Q}_{h+1}^\theta$, we have

$$\begin{aligned}
&\arg\min_{f\in\mathcal{F}} \left(\sum_{k=1}^K \left(f(s_h^{(k)}, a_h^{(k)}) - r_h^{(k)} - \int_{\mathcal{A}} \pi_{\theta,h+1}(a'|s_{h+1}^{(k)}) \widehat{Q}_{h+1}^\theta(s_{h+1}^{(k)}, a') \mathrm{d}a' \right)^2 + \lambda\rho(f) \right) = \widehat{r}_h + \widehat{\mathcal{P}}_{\theta,h} \widehat{Q}_{h+1}^\theta \\
&\arg\min_{f\in\mathcal{F}} \left(\sum_{k=1}^K \left(f(s_h^{(k)}, a_h^{(k)}) - \int_{\mathcal{A}} \pi_{\theta,h+1}(a'|s_{h+1}^{(k)}) \left(\left(\nabla_\theta^j \log \pi_{\theta,h+1}(a'|s_{h+1}^{(k)}) \right) \widehat{Q}_{h+1}^\theta(s_{h+1}^{(k)}, a') + \widehat{\nabla_\theta^j Q_{h+1}^\theta}(s_{h+1}^{(k)}, a') \right) \mathrm{d}a' \right)^2 \right. \\
&\left. + \lambda\rho(f) \right) \\
=&\widehat{\mathcal{P}}_{\theta,h} \left(\left(\nabla_\theta^j \log \Pi_\theta \right) \widehat{Q}_{h+1}^\theta + \widehat{\nabla_\theta^j Q_{h+1}^\theta} \right).
\end{aligned}$$

The second equation holds due to the definition of $\widehat{\mathcal{P}}_{\theta,h}$. For the first equation, note that when $\mathcal{F}$ is the class of linear functions and $\rho(\phi^\top w) = \|w\|^2$, the LHS has a closed form solution:

$$\begin{aligned}
&\arg\min_{f\in\mathcal{F}} \left(\sum_{k=1}^K \left(f(s_h^{(k)}, a_h^{(k)}) - r_h^{(k)} - \int_{\mathcal{A}} \pi_{\theta,h+1}(a'|s_{h+1}^{(k)}) \widehat{Q}_{h+1}^\theta(s_{h+1}^{(k)}, a') \mathrm{d}a' \right)^2 + \lambda\rho(f) \right) \\
=&\phi^\top \widehat{\Sigma}_h^{-1} \frac{1}{K} \sum_{k=1}^K \left(r_h^{(k)} + \int_{\mathcal{A}} \pi_{\theta,h+1}(a'|s_{h+1}^{(k)}) \widehat{Q}_{h+1}^\theta(s_{h+1}^{(k)}, a') \mathrm{d}a' \right) \\
=&\phi^\top \widehat{\Sigma}_h^{-1} \frac{1}{K} \sum_{k=1}^K r_h^{(k)} + \phi^\top \widehat{\Sigma}_h^{-1} \frac{1}{K} \sum_{k=1}^K \int_{\mathcal{A}} \pi_\theta(a'|s_{h+1}^{(k)}) \widehat{Q}_{h+1}^\theta(s_{h+1}^{(k)}, a') \mathrm{d}a' \\
=&\widehat{r}_h + \widehat{\mathcal{P}}_{\theta,h} \widehat{Q}_{h+1}^\theta.
\end{aligned}$$

Therefore, we have finished the proof. □

C.3. Proof of Proposition 6.8

Proof. The result of Theorem 6.3 implies for any fixed θ, when we choose $\lambda \leq \log\frac{8dmH}{\delta}C_1 d\min_{h\in[H]}\sigma_{\min}(\Sigma_h)$, and $K \geq 36\kappa_1(4+\kappa_2+\kappa_3)^2\log\frac{8dmH}{\delta}C_1dH^2$ sufficiently large such that

$$
\begin{aligned}
&4H^2G\sqrt{\min\{C_1d,H\}}\sqrt{1+\chi^2_{\mathcal{F}}(\mu^\theta,\bar{\mu})}\sqrt{\frac{2\log\frac{24m}{\delta}}{K}}\\
\geq &480C_1dm^{0.5}H^{3.5}\kappa_1(5+\kappa_2+\kappa_3)(\max_{j\in[m]}\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\|+HG\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\|)\frac{\log\frac{72mdH}{\delta}}{K},
\end{aligned}
$$

then we have

$$
\|\widehat{\nabla_\theta v_\theta}-\nabla_\theta v_\theta\| \leq 8H^2G\sqrt{m\min\{C_1d,H\}}\sqrt{1+\chi^2_{\mathcal{F}}(\mu^\theta,\bar{\mu})}\sqrt{\frac{2\log\frac{24m}{\delta}}{K}}.
$$

Note that when the diameter of Θ is bounded by D, for any $\varepsilon>0$, it's always possible to find an ε-net $\mathcal{N}_\varepsilon$ such that $|\mathcal{N}_\varepsilon| \leq \left(\frac{mD}{\varepsilon}\right)^m$. Taking a union bound over $\mathcal{N}_\varepsilon$, we get with probability $1-\delta$,

$$
\|\widehat{\nabla_\theta v_\theta}-\nabla_\theta v_\theta\| \leq 16H^2Gm\sqrt{\min\{C_1d,H\}}\sqrt{1+\chi^2_{\mathcal{F}}(\mu^\theta,\bar{\mu})}\sqrt{\frac{\log\frac{24mD}{\delta\varepsilon}}{K}}, \quad \forall\theta\in\mathcal{N}_\varepsilon.
$$

Therefore, for any $\theta\in\Theta$, pick $\theta'\in\mathcal{N}_\varepsilon$ such that $\|\theta-\theta'\|\leq\varepsilon$, we have

$$
\begin{aligned}
\|\widehat{\nabla_\theta v_\theta}-\nabla_\theta v_\theta\| &\leq 2L\varepsilon + \|\widehat{\nabla_\theta v_{\theta'}}-\nabla_\theta v_{\theta'}\|\\
&\leq 2L\varepsilon + 16H^2Gm\sqrt{\min\{C_1d,H\}}\sqrt{1+\chi^2_{\mathcal{F}}(\mu^{\theta'},\bar{\mu})}\sqrt{\frac{\log\frac{24mD}{\delta\varepsilon}}{K}}.
\end{aligned}
$$

Because $\chi^2_{\mathcal{F}}(\mu^{\theta'},\bar{\mu})$ is L'-Lipschitz in θ, we have

$$
\|\widehat{\nabla_\theta v_\theta}-\nabla_\theta v_\theta\| \leq 2L\varepsilon + 16H^2Gm\sqrt{\min\{C_1d,H\}}\sqrt{1+2L'\varepsilon+\chi^2_{\mathcal{F}}(\mu^{\theta},\bar{\mu})}\sqrt{\frac{\log\frac{24mD}{\delta\varepsilon}}{K}}.
$$

In particular, pick

$$
\varepsilon = \min\left\{\frac{1}{L'}, \frac{16H^2Gm}{L}\sqrt{\frac{\min\{C_1d,H\}}{K}}\right\},
$$

we get

$$
\|\widehat{\nabla_\theta v_\theta}-\nabla_\theta v_\theta\| \leq 64H^2Gm\sqrt{\min\{C_1d,H\}}\sqrt{1+\chi^2_{\mathcal{F}}(\mu^\theta,\bar{\mu})}\sqrt{\frac{\log\frac{24DKLL'}{\delta HG}}{K}}, \quad \forall\theta\in\Theta.
$$

□

C.4. Proof of Lemma B.1

Proof. Note that

$$
\begin{aligned}
\nabla_\theta Q_1^\theta - \widehat{\nabla_\theta Q_1^\theta} &= \sum_{h=1}^H\left(\prod_{h'=1}^{h-1}\mathcal{P}_{\theta,h'}\right)U_h^\theta - \sum_{h=1}^H\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\tilde{U}_h^\theta\\
&= \sum_{h=1}^H\left(\prod_{h'=1}^{h-1}\mathcal{P}_{\theta,h'}\right)U_h^\theta - \sum_{h=1}^H\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\widehat{U}_h^\theta + \sum_{h=1}^H\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\widehat{U}_h^\theta-\tilde{U}_h^\theta\right)\\
&= \sum_{h=1}^H\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\nabla_\theta Q_h^\theta - \widehat{U}_h^\theta - \widehat{\mathcal{P}}_{\theta,h}\nabla_\theta Q_{h+1}^\theta\right) + \sum_{h=1}^H\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\widehat{U}_h^\theta-\tilde{U}_h^\theta\right).
\end{aligned}
$$

For the first term, we have

$$\begin{aligned}
&\sum_{h=1}^{H}\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\nabla_\theta Q_h^\theta-\widehat{U}_h^\theta-\widehat{\mathcal{P}}_{\theta,h}\nabla_\theta Q_{h+1}^\theta\right)\\
=&\sum_{h=1}^{H}\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\phi^\top\widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)},a_h^{(k)}\right)\\
&\cdot\left(\nabla_\theta Q_h^\theta\left(s_h^{(k)},a_h^{(k)}\right)-\int_{\mathcal{A}}\left(\left(\nabla_\theta\pi_{\theta,h+1}\left(a'\,\middle|\,s_{h+1}^{(k)}\right)\right)Q_{h+1}^\theta\left(s_{h+1}^{(k)},a'\right)+\pi_{\theta,h+1}\left(a'\,\middle|\,s_{h+1}^{(k)}\right)\nabla_\theta Q_{h+1}^\theta\left(s_{h+1}^{(k)},a'\right)\right)\mathrm{d}a'\right)\\
&+\frac{\lambda}{K}\sum_{h=1}^{H}\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\phi^\top\widehat{\Sigma}_h^{-1}\nabla_\theta w_h^\theta\\
=&\sum_{h=1}^{H}\phi^\top\left(\prod_{h'=1}^{h-1}\widehat{M}_{\theta,h'}\right)\widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)},a_h^{(k)}\right)\nabla_\theta\varepsilon_{h,k}^\theta+\frac{\lambda}{K}\sum_{h=1}^{H}\phi^\top\left(\prod_{h'=1}^{h-1}\widehat{M}_{\theta,h'}\right)\widehat{\Sigma}_h^{-1}\nabla_\theta w_h^\theta.
\end{aligned}$$

Using the definition of $\widehat{\nu}_h^\theta$, we get

$$\begin{aligned}
&\int_{\mathcal{S}\times\mathcal{A}}\xi(s)\pi_{\theta,1}(a|s)\left(\sum_{h=1}^{H}\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\nabla_\theta Q_h^\theta-\widehat{U_h^\theta}-\widehat{\mathcal{P}}_{\theta,h}\nabla_\theta Q_{h+1}^\theta\right)\right)(s,a)\mathrm{d}s\mathrm{d}a\\
=&\sum_{h=1}^{H}\left(\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)},a_h^{(k)}\right)\nabla_\theta\varepsilon_{h,k}^\theta+\frac{\lambda}{K}\sum_{h=1}^{H}\left(\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}_h^{-1}\nabla_\theta w_h^\theta.
\end{aligned}\tag{17}$$

For the second term, by Lemma A.3, we have

$$\begin{aligned}
\sum_{h=1}^{H}\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\widehat{U}_h^\theta-\tilde{U}_h^\theta\right)&=\sum_{h=1}^{H}\left(\prod_{h'=1}^{h}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\nabla_\theta\log\Pi_{\theta,h+1}\right)\left(Q_{h+1}^\theta-\widehat{Q}_{h+1}^\theta\right)\\
&=\sum_{h=1}^{H}\left(\prod_{h'=1}^{h}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\nabla_\theta\log\Pi_{\theta,h+1}\right)\sum_{h'=h+1}^{H}\left(\prod_{h''=h+1}^{h'-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\left(Q_{h'}^\theta-\widehat{r}_{h'}-\widehat{\mathcal{P}}_{\theta,h'}Q_{h'+1}^\theta\right)\\
&=\sum_{h=1}^{H}\sum_{h'=1}^{h-1}\left(\prod_{h=1}^{h'}\widehat{\mathcal{P}}_{\theta,h}\right)\left(\nabla_\theta\log\Pi_{\theta,h'+1}\right)\left(\prod_{h''=h'+1}^{h-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\left(Q_h^\theta-\widehat{r}_h-\widehat{\mathcal{P}}_{\theta,h}Q_{h+1}^\theta\right).
\end{aligned}$$

Meanwhile, again by Lemma A.3, we have

$$\left(\nabla_\theta\log\Pi_{\theta,1}\right)\left(Q_1^\theta-\widehat{Q}_1^\theta\right)=\left(\nabla_\theta\log\Pi_{\theta,1}\right)\sum_{h=1}^{H}\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(Q_h^\theta-\widehat{r}_h-\widehat{\mathcal{P}}_{\theta,h}Q_{h+1}^\theta\right),$$

which implies

$$
\begin{aligned}
&\sum_{h=1}^{H}\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\widehat{U}_h^\theta-\tilde{U}_h^\theta\right)+\left(\nabla_\theta\log\Pi_{\theta,1}\right)\left(Q_1^\theta-\widehat{Q}_1^\theta\right)\\
=&\sum_{h=1}^{H}\left(\left(\nabla_\theta\log\Pi_{\theta,1}\right)\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)+\sum_{h'=1}^{h-1}\left(\prod_{h=1}^{h'}\widehat{\mathcal{P}}_{\theta,h}\right)\left(\nabla_\theta\log\Pi_{\theta,h'+1}\right)\left(\prod_{h''=h'+1}^{h-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\right)\left(Q_h^\theta-\widehat{r}_h-\widehat{\mathcal{P}}_{\theta,h}Q_{h+1}^\theta\right)\\
=&\sum_{h=1}^{H}\sum_{h'=0}^{h-1}\left(\prod_{h=1}^{h'}\widehat{\mathcal{P}}_{\theta,h}\right)\left(\nabla_\theta\log\Pi_{\theta,h'+1}\right)\left(\prod_{h''=h'+1}^{h-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\left(Q_h^\theta-\widehat{r}_h-\widehat{\mathcal{P}}_{\theta,h}Q_{h+1}^\theta\right)\\
=&\sum_{h=1}^{H}\sum_{h'=0}^{h-1}\left(\prod_{h=1}^{h'}\widehat{\mathcal{P}}_{\theta,h}\right)\left(\nabla_\theta\log\Pi_{\theta,h'+1}\right)\left(\prod_{h''=h'+1}^{h-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\phi^\top\widehat{\Sigma}_h^{-1}\\
&\cdot\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)},a_h^{(k)}\right)\left(Q_h^\theta\left(s_h^{(k)},a_h^{(k)}\right)-r_h^{(k)}-\int_{\mathcal{A}}\pi_{\theta,h+1}\left(a'\,\middle|\,s_{h+1}^{(k)}\right)Q_{h+1}^\theta\left(s_{h+1}^{(k)},a'\right)\mathrm{d}a'\right)\\
&+\frac{\lambda}{K}\sum_{h=1}^{H}\sum_{h'=0}^{h-1}\left(\prod_{h''=1}^{h'}\widehat{\mathcal{P}}_{\theta,h''}\right)\left(\nabla_\theta\log\Pi_{\theta,h'+1}\right)\left(\prod_{h''=h'+1}^{h-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\phi^\top\widehat{\Sigma}_h^{-1}w_h^\theta.
\end{aligned}
$$

For each $j\in[m]$, notice the relation

$$
\begin{aligned}
\left(\nabla_\theta^j\widehat{\nu}_h^\theta\right)^\top&=\left(\nabla_\theta^j\nu_1^\theta\right)^\top\left(\prod_{h'=1}^{h-1}\widehat{M}_{\theta,h'}\right)+\sum_{h'=1}^{h-1}\left(\nu_1^\theta\right)^\top\left(\prod_{h''=1}^{h'-1}\widehat{M}_{\theta,h''}\right)\left(\widehat{\nabla_\theta^j M_{\theta,h'}}\right)\left(\prod_{h''=h'+1}^{h-1}\widehat{M}_{\theta,h''}\right)\\
&=\int_{\mathcal{S}\times\mathcal{A}}\xi(s)\pi_{\theta,1}(a|s)\left(\sum_{h'=0}^{h-1}\left(\prod_{h''=1}^{h'}\widehat{\mathcal{P}}_{\theta,h''}\right)\left(\nabla_\theta\log\Pi_{\theta,h'+1}\right)\left(\prod_{h''=h'+1}^{h-1}\widehat{\mathcal{P}}_{\theta,h''}\right)\phi^\top\right)(s,a)\mathrm{d}s\mathrm{d}a.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\left[\int\xi(s)\pi_{\theta,1}(a|s)\left(\sum_{h=1}^{H}\left(\prod_{h'=1}^{h-1}\widehat{\mathcal{P}}_{\theta,h'}\right)\left(\widehat{U}_h^\theta-\tilde{U}_h^\theta\right)+\left(\nabla_\theta\log\Pi_{\theta,1}\right)\left(Q_1^\theta-\widehat{Q_1^\theta}\right)\right)(s,a)\mathrm{d}s\mathrm{d}a\right]_j\\
=&\sum_{h=1}^{H}\left(\nabla_\theta^j\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)},a_h^{(k)}\right)\left(Q_h^\theta\left(s_h^{(k)},a_h^{(k)}\right)-r_h^{(k)}-\int_{\mathcal{A}}\pi_{\theta,h+1}\left(a'\,\middle|\,s_{h+1}^{(k)}\right)Q_{h+1}^\theta\left(s_{h+1}^{(k)},a'\right)\mathrm{d}a'\right)\\
&+\frac{\lambda}{K}\sum_{h=1}^{H}\left(\nabla_\theta^j\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}_h^{-1}w_h^\theta\\
=&\sum_{h=1}^{H}\left(\nabla_\theta^j\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)},a_h^{(k)}\right)\varepsilon_{h,k}^\theta+\frac{\lambda}{K}\sum_{h=1}^{H}\left(\nabla_\theta^j\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}_h^{-1}w_h^\theta.
\end{aligned}
\tag{18}
$$

Combing the results of (17) and (18), we get for each $j \in [m]$,

$$
\begin{aligned}
\nabla_\theta^j v_\theta - \widehat{\nabla_\theta^j v_\theta} &= \int_{\mathcal{S}\times\mathcal{A}} \xi(s)\pi_{\theta,1}(a|s)\left(\nabla_\theta^j Q_1^\theta - \widehat{\nabla_\theta^j Q_1^\theta} + \left(\nabla_\theta^j \log \Pi_{\theta,1}\right)\left(Q_1^\theta - \widehat{Q_1^\theta}\right)\right)(s,a)\mathrm{d}s\mathrm{d}a \\
&= \sum_{h=1}^H \Bigg(\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} \frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\nabla_\theta^j \varepsilon_{h,k}^\theta + \frac{\lambda}{K}\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1}\nabla_\theta^j w_h^\theta + \left(\nabla_\theta^j \nu_h^\theta\right)^\top \widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \\
&\quad + \frac{\lambda}{K}\left(\nabla_\theta^j \widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} w_h^\theta \Bigg) \\
&= \sum_{h=1}^H \nabla_\theta^j \left(\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta + \frac{\lambda}{K}\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} w_h^\theta \right) \\
&= \sum_{h=1}^H \nabla_\theta^j \Bigg(\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta + \frac{\lambda}{K}\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} w_h^\theta \\
&\quad + \left(\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} - \left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\right)\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \Bigg).
\end{aligned}
$$

Rewriting the above decomposition in a vector form, we get

$$
\begin{aligned}
\nabla_\theta v_\theta - \widehat{\nabla_\theta v_\theta} = \sum_{h=1}^H \nabla_\theta \Bigg(&\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \\
&+ \frac{\lambda}{K}\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} w_h^\theta + \left(\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} - \left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\right)\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \Bigg),
\end{aligned}
$$

which is the desired result. □

C.5. Proof of Lemma B.2

Proof. Note that,

$$
\begin{aligned}
\langle E_1, t\rangle &= \sum_{h=1}^H \left\langle \nabla_\theta \left(\left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \right), t \right\rangle \\
&= \sum_{h=1}^H \left\langle \left(\nabla_\theta \nu_h^\theta\right)^\top \Sigma_h^{-1}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta, t \right\rangle + \sum_{h=1}^H \left\langle \left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\nabla_\theta \varepsilon_{h,k}^\theta, t \right\rangle.
\end{aligned}
$$

Let $e_k = \sum_{h=1}^H \left\langle \nabla_\theta \left(\left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \right), t \right\rangle$, we have

$$
\begin{aligned}
|e_k| \leq& \sqrt{C_1 dm}\|t\| \sum_{h=1}^H (H-h+1)\max_{j\in[m]} \sqrt{\left(\nabla_\theta^j \nu_h^\theta\right)^\top \Sigma_h^{-1}\nabla_\theta^j \nu_h^\theta} + 2G\sqrt{C_1 dm}\|t\| \sum_{h=1}^H (H-h)^2 \sqrt{\left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\nu_h^\theta} \\
=& B\sqrt{C_1 dm}\|t\|.
\end{aligned}
$$

We have

$$
\sum_{k=1}^K \mathrm{Var}[e_k] = \sum_{k=1}^K \mathbb{E}\left[\left\langle \sum_{h=1}^H \nabla_\theta \left(\left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \right), t \right\rangle^2 \right] = K t^\top \Lambda_\theta t.
$$

We pick $\sigma^2 = K t^\top \Lambda_\theta t$, the Bernstein's inequality implies that for any $\varepsilon \in \mathbb{R}$,

$$
\mathbb{P}\left(\left| \sum_{k=1}^K e_k \right| \geq \varepsilon \right) \leq 2\exp\left(-\frac{\varepsilon^2/2}{\sigma^2 + \sqrt{C_1 dm}\|t\| B\varepsilon/3} \right).
$$

Therefore, if we pick $\varepsilon = \sigma\sqrt{2\log(2/\delta)} + 2\log(2/\delta)\sqrt{C_1 dm}\|t\|B/3$, we get

$$\mathbb{P}\left(\left|\sum_{k=1}^{K} e_k\right| \geq \varepsilon\right) \leq \delta,$$

i.e., we have with probability $1-\delta$,

$$\left|\frac{1}{K}\sum_{k=1}^{K} e_k\right| \leq \sqrt{\frac{2t^\top \Lambda_\theta t \log(2/\delta)}{K}} + \frac{2\log(2/\delta)\sqrt{C_1 dm}\|t\|B}{3K}$$

□

C.6. Proof of Lemma B.3

Proof. For an arbitrarily given θ_0, let $\Sigma_{\theta_0,h} = \mathbb{E}^{\pi_{\theta_0}}[\phi(s_h,a_h)\phi(s_h,a_h)^\top]$, we have

$$\begin{aligned}
&\left(\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} - \left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\right)\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \\
&= \left(\nu_1^\theta\right)^\top \left(\left(\prod_{h'=1}^{h-1}\widehat{M}_{\theta,h'}\right)\widehat{\Sigma}_h^{-1} - \left(\prod_{h'=1}^{h-1} M_{\theta,h'}\right)\Sigma_h^{-1}\right)\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \\
&= \left(\Sigma_{\theta_0,1}^{-\frac{1}{2}}\nu_1^\theta\right)^\top \left(\left(\prod_{h'=1}^{h-1}\Sigma_{\theta_0,h'}^{\frac{1}{2}}\widehat{M}_{\theta,h'}\Sigma_{\theta_0,h'+1}^{-\frac{1}{2}}\right)\Sigma_{\theta_0,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\Sigma_h^{\frac{1}{2}}\widehat{\Sigma}_h^{-1}\Sigma_h^{\frac{1}{2}} - \left(\prod_{h'=1}^{h-1}\Sigma_{\theta_0,h'}^{\frac{1}{2}} M_{\theta,h'}\Sigma_{\theta_0,h'+1}^{-\frac{1}{2}}\right)\Sigma_{\theta_0,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right) \\
&\quad \cdot \Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta.
\end{aligned}$$

Taking derivatives on both sides, and let $\theta_0 = \theta$, we get

$$\nabla_\theta^j E_2 = \nabla_\theta^j\left(\sum_{h=1}^{H}\left(\left(\widehat{\nu}_h^\theta\right)^\top \widehat{\Sigma}_h^{-1} - \left(\nu_h^\theta\right)^\top \Sigma_h^{-1}\right)\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right) = E_{21}^j + E_{22}^j + E_{23}^j,$$

where

$$\begin{aligned}
E_{21}^j =& \sum_{h=1}^{H}\left(\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right)^\top \left(\left(\prod_{h'=1}^{h-1}\Sigma_{\theta,h'}^{\frac{1}{2}}\widehat{M}_{\theta,h'}\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right)\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\Sigma_h^{\frac{1}{2}}\widehat{\Sigma}_h^{-1}\Sigma_h^{\frac{1}{2}} - \left(\prod_{h'=1}^{h-1}\Sigma_{\theta,h'}^{\frac{1}{2}} M_{\theta,h'}\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right)\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right) \\
&\cdot \Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\nabla_\theta^j\varepsilon_{h,k}^\theta \\
E_{22}^j =& \sum_{h=1}^{H}\left(\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right)^\top \left(\left(\prod_{h'=1}^{h-1}\Sigma_{\theta,h'}^{\frac{1}{2}}\widehat{M}_{\theta,h'}\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right)\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\Sigma_h^{\frac{1}{2}}\widehat{\Sigma}_h^{-1}\Sigma_h^{\frac{1}{2}} - \left(\prod_{h'=1}^{h-1}\Sigma_{\theta,h'}^{\frac{1}{2}} M_{\theta,h'}\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right)\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right) \\
&\cdot \Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \\
E_{23}^j =& \sum_{h=1}^{H}\left(\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right)^\top \\
&\cdot \left(\nabla_\theta^j\left(\prod_{h'=1}^{h-1}\Sigma_{\theta_0,h'}^{\frac{1}{2}}\widehat{M}_{\theta,h'}\Sigma_{\theta_0,h'+1}^{-\frac{1}{2}}\right)\Bigg|_{\theta_0=\theta}\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\Sigma_h^{\frac{1}{2}}\widehat{\Sigma}_h^{-1}\Sigma_h^{\frac{1}{2}} - \nabla_\theta^j\left(\prod_{h'=1}^{h-1}\Sigma_{\theta_0,h'}^{\frac{1}{2}} M_{\theta,h'}\Sigma_{\theta_0,h'+1}^{-\frac{1}{2}}\right)\Bigg|_{\theta_0=\theta}\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right) \\
&\cdot \Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^{K}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta.
\end{aligned}$$

Therefore, using the result of Lemma A.4, we get

$$|E_{21}^j| \le \sum_{h=1}^H \left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\| \left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\| \left(\left(\prod_{h'=1}^{h-1}\left(1+\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}\left(\Delta M_{\theta,h'}\right)\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right\|\right)\right)\left(1+\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\|\right)-1\right)$$
$$\cdot \left\|\Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)},a_h^{(k)}\right)\nabla_\theta^j \varepsilon_{h,k}^\theta\right\|$$
$$|E_{22}^j| \le \sum_{h=1}^H \left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\| \left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\| \left(\left(\prod_{h'=1}^{h-1}\left(1+\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}\left(\Delta M_{\theta,h'}\right)\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right\|\right)\right)\left(1+\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\|\right)-1\right)$$
$$\cdot \left\|\Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)},a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right\|$$
$$|E_{23}^j| \le \sum_{h=1}^H\sum_{h'=1}^{h-1} G\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\| \left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|$$
$$\cdot\left(\left(\prod_{h''\neq h'}\left(1+\left\|\Sigma_{\theta,h''}^{\frac{1}{2}}\left(\Delta M_{\theta,h''}\right)\Sigma_{\theta,h''}^{-\frac{1}{2}}\right\|\right)\right)\left(1+\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_{\theta,h'}\right)}{G}\right)\Sigma_{\theta,h'}^{-\frac{1}{2}}\right\|\right)\left(1+\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\|\right)-1\right)$$
$$\cdot \left\|\Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)},a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right\|,$$

where $\Delta\Sigma_h^{-1}=\widehat{\Sigma}_h^{-1}-\Sigma_h$ and we use the fact $\left\|\Sigma_{\theta,h}^{\frac{1}{2}}M_{\theta,h}\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\|\le 1$ and $\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\left(\nabla_\theta^j M_{\theta,h}\right)\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\|\le G$ from Lemma A.5. Furthermore, we have

$$\begin{aligned}&\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\left(\Delta M_{\theta,h}\right)\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\|\\ \le&\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left\|\Sigma_{h+1}^{\frac{1}{2}}\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\|\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta M_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\|=\sqrt{\kappa_1}\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta M_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\|\\ =&\sqrt{\kappa_1}\left\|\Sigma_h^{\frac{1}{2}}\left(\widehat{\Sigma}_h^{-1}\frac{1}{K}\sum_{k=1}^K\phi\left(s_h^{(k)},a_h^{(k)}\right)\int_{\mathcal{A}}\phi\left(s_{h+1}^{(k)},a'\right)\pi_{\theta,h+1}\left(a'\left|s_{h+1}^{(k)}\right.\right)\mathrm{d}a'-M_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\|\\ \le&\sqrt{\kappa_1}\left(\left(1+\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\|\right)\left(1+\left\|\Sigma_h^{-\frac{1}{2}}\left(\Delta Y_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\|\right)-1\right),\end{aligned}$$

where $\Delta Y_{\theta,h}=\frac{1}{K}\sum_{k=1}^K\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi_{\theta,h+1}\left(s_{h+1}^{(k)}\right)^\top-\Sigma_h M_{\theta,h}$ and the last inequality uses Lemma A.4 again. Similarly, we have

$$\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_{\theta,h}\right)}{G}\right)\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\|\le\sqrt{\kappa_1}\left(\left(1+\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h\right)\Sigma_h^{\frac{1}{2}}\right\|\right)\left(1+\left\|\Sigma_h^{-\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta Y_{\theta,h}\right)}{G}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\|\right)-1\right).$$

Now, define $\alpha=6\sqrt{\kappa_1}(4+\kappa_2+\kappa_3)\sqrt{\frac{C_1 d\log\frac{8dmH}{\delta}}{K}}$ and pick

$$K\ge 36\kappa_1(4+\kappa_2+\kappa_3)^2C_1dH^2\log\frac{8dmH}{\delta},\quad \lambda\le C_1d\min_{h\in[H]}\sigma_{\min}(\Sigma_h)\cdot\log\frac{8dmH}{\delta},$$

we get $\alpha\le\frac{1}{H}$. Using the results of Lemma A.6, Lemma A.7, Lemma A.8, we get with probability $1-2\delta$,

$$\begin{aligned}\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\|\le&2\left\|\Sigma_h^{-\frac{1}{2}}\widehat{\Sigma}_h\Sigma_h^{-\frac{1}{2}}-I_d\right\|\le 2\sqrt{\frac{2C_1d\log\frac{2dH}{\delta}}{K}}+\frac{4C_1d\log\frac{2dH}{\delta}}{3K}+\frac{2\lambda\|\Sigma_h^{-1}\|}{K}\\ \le&4\sqrt{\frac{C_1d\log\frac{8dmH}{\delta}}{K}}\le\alpha\le 1,\end{aligned}\tag{19}$$

and $\forall j \in [m]$,

$$\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta Y_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\| \leq 2(\kappa_2+1)\sqrt{\frac{C_1 d\log\frac{8dmH}{\delta}}{K}} \leq 1, \quad \left\|\Sigma_h^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta Y_{\theta,h}\right)}{G}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\| \leq 2(\kappa_3+1)\sqrt{\frac{C_1 d\log\frac{8dmH}{\delta}}{K}} \leq 1,$$

which implies

$$\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\left(\Delta M_{\theta,h}\right)\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\| \leq 2\sqrt{\kappa_1}\left(\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\| + \left\|\Sigma_h^{\frac{1}{2}}\left(\Delta Y_{\theta,h}\right)\Sigma_{h+1}^{-\frac{1}{2}}\right\|\right) \leq \alpha, \tag{20}$$

where we use the fact $(1+x_1)(1+x_2)-1 \leq 2(x_1+x_2)$ whenever $x_1, x_2 \in [0,1]$. Similarly, we get

$$\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_{\theta,h}\right)}{G}\right)\Sigma_{\theta,h+1}^{-\frac{1}{2}}\right\| \leq \alpha, \quad \forall j \in [m]. \tag{21}$$

Meanwhile, by Lemma A.9, we get with probability $1-\delta$,

$$\left\|\Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right\| \leq 4\sqrt{d}(H-h+1)\sqrt{\frac{\log\frac{8dmH}{\delta}}{K}} \tag{22}$$

$$\left\|\Sigma_h^{-\frac{1}{2}}\frac{1}{K}\sum_{k=1}^K \phi\left(s_h^{(k)}, a_h^{(k)}\right)\nabla_\theta^j\varepsilon_{h,k}^\theta\right\| \leq 8\sqrt{d}G(H-h)^2\sqrt{\frac{\log\frac{8dmH}{\delta}}{K}}. \tag{23}$$

Combining the results of (19), (20), (21), (22),(23) and use a union bound, we have with probability $1-3\delta$,

$$\begin{aligned}
|E_{21}^j| \leq & \sum_{h=1}^H 16h\alpha\sqrt{d}G(H-h)^2\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{8dmH}{\delta}}{K}} \\
\leq & 16\alpha\sqrt{d}H^4G\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{8dmH}{\delta}}{K}}, \\
|E_{22}^j| \leq & \sum_{h=1}^H 8h\alpha\sqrt{d}(H-h+1)\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{8dmH}{\delta}}{K}} \\
\leq & 8\alpha\sqrt{d}H^3\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{8dmH}{\delta}}{K}} \\
|E_{23}^j| \leq & \sum_{h=1}^H 16Gh\alpha\sqrt{d}(H-h+1)(h-1)\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{8dmH}{\delta}}{K}} \\
\leq & 16\alpha\sqrt{d}H^4G\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{8dmH}{\delta}}{K}},
\end{aligned}$$

where we use the fact $(1+\alpha)^h - 1 \leq 2h\alpha$ whenever $\alpha h \leq 1$. Summing up the above terms and using the definition of α, we get

$$|E_2^j| \leq 240\sqrt{\kappa_1}(4+\kappa_2+\kappa_3)\sqrt{C_1}dH^3\left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\| + HG\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\right)\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\frac{\log\frac{8dmH}{\delta}}{K}, \quad \forall j \in [m]$$

Replacing δ by $\frac{\delta}{3}$, we have finished the proof. $\square$

C.7. Proof of Lemma B.4

Proof. Similar to the decomposition in the proof of Lemma B.3, we have

$$
\begin{aligned}
&\left|E_3^j\right| = \frac{\lambda}{K}\left|\sum_{h=1}^{H}\nabla_\theta^j\left(\left(\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}_h^{-1}w_h^\theta\right)\right| \\
\le&\frac{\lambda}{K}\sum_{h=1}^{H}\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\prod_{h'=1}^{h-1}\left(1+\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}\left(\Delta M_{\theta,h'}\right)\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right\|\right)\right)\left(1+\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\|\right)\left\|\Sigma_h^{-1}\right\|\left\|\Sigma_h^{\frac{1}{2}}\nabla_\theta^j w_h^\theta\right\| \\
+&\frac{\lambda}{K}\sum_{h=1}^{H}\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\prod_{h'=1}^{h-1}\left(1+\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}\left(\Delta M_{\theta,h'}\right)\Sigma_{\theta,h'+1}^{-\frac{1}{2}}\right\|\right)\right)\left(1+\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\|\right)\left\|\Sigma_h^{-1}\right\|\left\|\Sigma_h^{\frac{1}{2}}w_h^\theta\right\| \\
+&\frac{\lambda}{K}\sum_{h=1}^{H}\sum_{h'=1}^{h-1}G\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\| \\
&\cdot\left(\prod_{h''\neq h'}\left(1+\left\|\Sigma_{\theta,h''}^{\frac{1}{2}}\left(\Delta M_{\theta,h''}\right)\Sigma_{\theta,h''}^{-\frac{1}{2}}\right\|\right)\right)\left(1+\left\|\Sigma_{\theta,h'}^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_{\theta,h'}\right)}{G}\right)\Sigma_{\theta,h'}^{-\frac{1}{2}}\right\|\right)\left(1+\left\|\Sigma_h^{\frac{1}{2}}\left(\Delta\Sigma_h^{-1}\right)\Sigma_h^{\frac{1}{2}}\right\|\right) \\
&\cdot\left\|\Sigma_h^{-1}\right\|\left\|\Sigma_h^{\frac{1}{2}}w_h^\theta\right\| \\
\le&\frac{\lambda}{K}\sum_{h=1}^{H}(1+\alpha)^h\left\|\Sigma_h^{-1}\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_h^{\frac{1}{2}}\nabla_\theta^j w_h^\theta\right\|+\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\left\|\Sigma_h^{\frac{1}{2}}w_h^\theta\right\|+G(h-1)\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_h^{\frac{1}{2}}w_h^\theta\right\|\right),
\end{aligned}
$$

where α is defined in the same way as that in the proof of Lemma B.3. Similarly, we have $\alpha \le \frac{1}{H}$ with probability $1-3\delta$ and we have

$$
\begin{aligned}
\left\|\Sigma_h^{\frac{1}{2}}\nabla_\theta^j w_h^\theta\right\|^2 &= \mathbb{E}\left[\left(\nabla_\theta^j Q_h^\theta\left(s_h^{(1)},a_h^{(1)}\right)\right)^2\right] \le G^2(H-h)^4 \\
\left\|\Sigma_h^{\frac{1}{2}}w_h^\theta\right\|^2 &= \mathbb{E}\left[\left(Q_h^\theta\left(s_h^{(1)},a_h^{(1)}\right)\right)^2\right] \le (H-h+1)^2.
\end{aligned}
$$

We conclude

$$
\begin{aligned}
&|E_3^j| \\
\le&3\frac{\lambda}{K}\sum_{h=1}^{H}\left\|\Sigma_h^{-1}\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|G(H-h)^2+\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|(H-h+1)+G(h-1)\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|(H-h+1)\right) \\
\le&6\frac{\lambda}{K}H^2\max_{h\in[H]}\left\|\Sigma_h^{-1}\right\|\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|GH+\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\right) \\
\le&6\frac{\log\frac{8dmH}{\delta}C_1dH^2}{K}\max_{h\in[H]}\left\|\Sigma_{\theta,h}^{\frac{1}{2}}\Sigma_h^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nu_1^\theta\right\|GH+\left\|\Sigma_{\theta,1}^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\right).
\end{aligned}
$$

Replacing δ by $\frac{\delta}{3}$, we have finished the proof. □

D. Extension to Time-homogeneous Discounted MDP

D.1. Approach

Our method can be easily extended to the case of time-homogeneous discounted MDP. Similar to the time-inhomogeneous case, under the setting of the time-homogeneous discounted MDP, an instance of MDP is defined by $(\mathcal{S},\mathcal{A},p,r,\xi,\gamma)$ where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $\gamma\in(\frac{1}{2},1)$ is the discount factor, $p:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}_+$ is the transition probability, $r:\mathcal{S}\times\mathcal{A}\to[0,1]$ is the reward function and $\xi:\mathcal{S}\to\mathbb{R}_+$ is the initial state distribution. Similarly, the policy $\pi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}_+$ is a distribution over the action space conditioned on an arbitrary given state s. We define the value function and Q function by

$$v^\theta = \mathbb{E}\left[\sum_{h=1}^{\infty}\gamma^{h-1}r(s_h,a_h)\right], \quad Q^\theta(s,a) = \mathbb{E}\left[\sum_{h=1}^{\infty}\gamma^{h-1}r(s_h,a_h)\middle| s_1=s, a_1=a\right]$$

Note that here the reward and Q function no longer contain the subscript h. We still consider the class of linear functions $\mathcal{F}$ with state-action feature ϕ, and denote $\mathcal{P}_\theta$ as the transition operator where $\theta \in \mathbb{R}^m$ is the parameter of the policy.

Assumption D.1. For any $f \in \mathcal{F}$, we have $\mathcal{P}_\theta f \in \mathcal{F}$, and we suppose $r \in \mathcal{F}$.

In addition, we assume that the constant function belongs to $\mathcal{F}$, i.e., there exists some w_0 such that $\phi(s,a)^\top w_0 = 1, \forall s \in \mathcal{S}, a \in \mathcal{A}$. Define the covariance matrix Σ and its empirical version $\widehat{\Sigma}$ by

$$\Sigma := \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H}\phi\left(s_h^{(1)},a_h^{(1)}\right)\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\right], \quad \widehat{\Sigma} := \frac{1}{HK}\left(\lambda I_d + \sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\right),$$

where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix.

Assumption D.2 (Boundedness Conditions)**.** Assume Σ is invertible. There exist absolute constants C_1, G such that for any $(s,a) \in \mathcal{S}\times\mathcal{A}, j \in [m]$, we have

$$\phi(s,a)^\top\Sigma^{-1}\phi(s,a) \le C_1 d, \quad \left|\nabla_\theta^j \log\pi_\theta(a|s)\right| \le G.$$

Define $\widehat{w}_r \in \mathbb{R}^d, \widehat{M_\theta} \in \mathbb{R}^{d\times d}, \widehat{\nabla_\theta M_\theta} \in \mathbb{R}^{d\times d}, j \in [m]$ by

$$\widehat{w}_r := \widehat{\Sigma}^{-1}\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)},a_h^{(k)}\right)r_h^{(k)},$$

$$\widehat{M_\theta} := \widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)},a_h^{(k)}\right)\int_{\mathcal{A}}\pi_\theta\left(a'\middle|s_{h+1}^{(k)}\right)\phi\left(s_{h+1}^{(k)},a'\right)^\top \mathrm{d}a',$$

$$\widehat{\nabla_\theta^j M_\theta} := \widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)},a_h^{(k)}\right)\int_{\mathcal{A}}\phi\left(s_{h+1}^{(k)},a'\right)^\top\nabla_\theta\pi_\theta\left(a'\middle|s_{h+1}^{(k)}\right)\mathrm{d}a', \quad j \in [m].$$

In this way, one can compute

$$\widehat{Q}^\theta(\cdot,\cdot) = \phi(\cdot,\cdot)^\top\widehat{w}^\theta, \quad \widehat{\nabla_\theta^j Q^\theta}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top\widehat{\nabla_\theta^j w^\theta},$$

where

$$\widehat{w}^\theta = \left(I_d - \gamma\widehat{M_\theta}\right)^{-1}\widehat{w}_r, \quad \widehat{\nabla_\theta^j w^\theta} = \left(I_d - \gamma\widehat{M_\theta}\right)^{-1}\widehat{\nabla_\theta^j M_\theta}\widehat{w}^\theta.$$

Then the estimator is derived from

$$\widehat{\nabla_\theta v_\theta} = \int_{\mathcal{S}\times\mathcal{A}}\xi(s)\pi_\theta(a|s)\left(\widehat{\nabla_\theta Q^\theta}(s,a) + (\nabla_\theta\log\pi_\theta(a|s))\,\widehat{Q}^\theta(s,a)\right)\mathrm{d}s\mathrm{d}a.$$

D.2. Results

Define $\nu_h^\theta := \mathbb{E}^{\pi_\theta}\left[\phi(s_h,a_h)|s_1\sim\xi\right], \nu^\theta = \sum_{h=1}^{\infty}\gamma^{h-1}\nu_h^\theta$ and $\Sigma_\theta := \mathbb{E}^{\pi_\theta}\left[\phi(s,a)\phi(s,a)^\top\middle| s\sim\xi_\theta, a\sim\pi_\theta(\cdot|s)\right]$ where ξ_θ is the stationary distribution under π_θ. Define

$$\phi_\theta(s) = \int_{\mathcal{A}}\pi_\theta(a'|s)\phi(s,a')\mathrm{d}a', \quad \varepsilon_{h,k}^\theta = Q^\theta\left(s_h^{(k)},a_h^{(k)}\right) - r_h^{(k)} - \gamma\int_{\mathcal{A}}\pi_\theta\left(a'\middle|s_{h+1}^{(k)}\right)Q^\theta\left(s_{h+1}^{(k)},a'\right)\mathrm{d}a',$$

and

$$\Lambda_\theta = \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H}\left(\nabla_\theta\left(\varepsilon_{h,1}^\theta\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nu^\theta\right)\right)^\top\nabla_\theta\left(\varepsilon_{h,1}^\theta\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nu^\theta\right)\right].$$

We first give the finite sample guarantee.

Theorem D.3 (Finite Sample Guarantee). *For any $t \in \mathbb{R}^m$, when $K \geq 36\kappa_1(4+\kappa_2+\kappa_3)^2 C_1 d(1-\gamma)^{-2}\log\frac{16dmH}{\delta}$ and $\lambda \leq \log\frac{8dmH}{\delta} C_1 d\sigma_{\min}(\Sigma)$, with probability $1-\delta$, we have,*

$$|\langle t, \widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta \rangle| \leq \sqrt{\frac{2t^\top \Lambda_\theta t}{HK} \cdot \log\frac{8}{\delta}} + \frac{C_\theta \|t\| \log\frac{32mdH}{\delta}}{HK},$$

where $C_\theta = 240 C_1 d m^{0.5}(1-\gamma)^{-3}\kappa_1(5+\kappa_2+\kappa_3)\left(\max_{j\in[m]}\left\|\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j \nu_1^\theta\right\| + \frac{G}{1-\gamma}\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\right)$ and

$$\kappa_1 = \frac{\sigma_{\max}\left(\Sigma^{-\frac{1}{2}}\Sigma_\theta\Sigma^{-\frac{1}{2}}\right)}{\sigma_{\min}\left(\Sigma^{-\frac{1}{2}}\Sigma_\theta\Sigma^{-\frac{1}{2}}\right)\wedge 1}, \quad \kappa_2 = \left\|\Sigma^{-\frac{1}{2}}\mathbb{E}\left[\phi_\theta\left(s_{h+1}^{(1)}\right)\phi_\theta\left(s_{h+1}^{(1)}\right)^\top\right]\Sigma^{-\frac{1}{2}}\right\|^{\frac{1}{2}},$$

$$\kappa_3 = \frac{1}{G}\max_{j\in[m]}\left\|\Sigma^{-\frac{1}{2}}\mathbb{E}\left[\left(\nabla_\theta^j\phi_\theta\left(s_{h+1}^{(1)}\right)\right)\left(\nabla_\theta^j\phi_\theta\left(s_{h+1}^{(1)}\right)\right)^\top\right]\Sigma^{-\frac{1}{2}}\right\|^{\frac{1}{2}}.$$

Theorem D.4 (Finite Sample Guarantee - Reward Free). *Let the conditions in Theorem D.3 hold, with probability $1-\delta$, we have for any reward function r,*

$$\left|\widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta\right| \leq 4b_\theta\sqrt{\frac{\log\frac{8m}{\delta}}{HK}} + \frac{2C_\theta \log\frac{32mdH}{\delta}}{HK}, \quad \forall j \in [m],$$

where $b_\theta = \frac{G}{(1-\gamma)^2}\left\|\Sigma^{-\frac{1}{2}}\nu^\theta\right\| + \frac{1}{1-\gamma}\left\|\Sigma^{-\frac{1}{2}}\nabla_\theta^j\nu^\theta\right\|$ and C_θ is the same as that in Theorem D.3. If we in addition have $\phi(s',a')^\top\Sigma^{-1}\phi(s,a) \geq 0, \forall (s,a),(s',a') \in \mathcal{S}\times\mathcal{A}$, we have

$$\left|\widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta\right| \leq \frac{16G}{(1-\gamma)^2}\sqrt{\frac{\log\frac{8m}{\delta}}{HK}}\left\|\Sigma^{-\frac{1}{2}}\nu^\theta\right\|\log(C_1 d) + \frac{2C_\theta\sqrt{m}\log\frac{32mdH}{\delta}}{HK}, \quad \forall j \in [m].$$

The complete proofs of Theorem D.3 and Theorem D.4 are deferred to Appendix D.6.1 and D.6.2. Next we show that FPG is an asymptotically normal and efficient estimator.

Theorem D.5 (Asymptotic Normality). *The FPG estimator is asymptotically normal:*

$$\sqrt{HK}\left(\widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta\right) \xrightarrow{d} \mathcal{N}(0, \Lambda_\theta).$$

The proof of Theorem D.5 is deferred to Appendix D.6.3. An obvious corollary of Theorem D.5 is that for any vector $t \in \mathbb{R}^m$,

$$\sqrt{HK}\left\langle t, \widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta\right\rangle \xrightarrow{d} \mathcal{N}\left(0, t^\top\Lambda_\theta t\right).$$

The following theorem states the Cramer Rao bound for FPG estimation.

Theorem D.6. *Let Assumption D.1 hold. For any vector $t \in \mathbb{R}^m$, the variance of any unbiased estimator for $t^\top\nabla_\theta v_\theta \in \mathbb{R}$ is lower bounded by $\frac{1}{\sqrt{HK}} t^\top\Lambda_\theta t$.*

The proof of Theorem D.6 is deferred to Appendix D.6.4.

D.3. Additional Notations

Define

$$\begin{aligned}
\widehat{r}(\cdot,\cdot) &:= \phi(\cdot,\cdot)^\top \widehat{w}_r, \\
U^\theta &:= \gamma\mathcal{P}_\theta\left(\nabla_\theta \log\Pi_\theta\right) Q^\theta, \\
\widehat{U}^\theta &:= \gamma\widehat{\mathcal{P}}_\theta\left(\nabla_\theta \log\Pi_\theta\right) Q^\theta, \\
\tilde{U}^\theta &:= \gamma\widehat{\mathcal{P}}_\theta\left(\nabla_\theta \log\Pi_\theta\right) \widehat{Q}^\theta \\
\Delta Y_\theta &:= \frac{1}{HK}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\phi_\theta\left(s_{h+1}^{(k)}\right)^\top - \Sigma M_\theta \\
\Delta\Sigma^{-1} &:= \widehat{\Sigma}^{-1} - \Sigma.
\end{aligned}$$

When $\mathcal{F}$ is the class of the linear functions, there exists matrix M_θ such that the transition probability satisfies

$$\mathbb{E}^{\pi_\theta}\left[\phi(s',a')^\top|s,a\right] = \phi(s,a)^\top M_\theta.$$

D.4. Technical Lemmas

Lemma D.7. *We have*

$$Q^\theta = \sum_{h=1}^{\infty} \gamma^{h-1} \left(\mathcal{P}_\theta\right)^{h-1} r, \quad \nabla_\theta Q^\theta = \sum_{h=1}^{\infty} \gamma^{h-1} \left(\mathcal{P}_\theta\right)^{h-1} U^\theta.$$

Proof. By Bellman's equation, we have $Q^\theta = r + \gamma \mathcal{P}_\theta Q^\theta$, which implies

$$Q^\theta = r + \gamma \mathcal{P}_\theta Q^\theta = r + \gamma \mathcal{P}_\theta r + \gamma^2 \left(\mathcal{P}_\theta\right)^2 Q^\theta = \ldots = \sum_{h=1}^{\infty} \gamma^{h-1} \left(\mathcal{P}_\theta\right)^{h-1} r,$$

which proves the first equation. Differentiating on both sides of the Bellman's equation w.r.t. θ, we have

$$\nabla_\theta Q^\theta(s,a) = \gamma \mathbb{E}^{\pi_\theta}\left[\left(\nabla_\theta \log \pi_\theta(a'|s')\right) Q^\theta(s',a')|s,a\right] + \gamma \mathbb{E}^{\pi_\theta}\left[\nabla_\theta Q^\theta(s',a')|s,a\right],$$

i.e., $\nabla_\theta Q^\theta = U^\theta + \gamma \mathcal{P}_\theta \nabla_\theta Q^\theta$. By induction, we have proved the second equation. □

The decomposition leads to the following boundedness result:

Lemma D.8. *We have* $|Q^\theta(s,a)| \le \frac{1}{1-\gamma}$, $\left\|\nabla_\theta Q^\theta(s,a)\right\|_\infty \le \frac{G}{(1-\gamma)^2}$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$.

Lemma D.9. *For any series of matrices* $A_1, A_2, \ldots, A_n$ *and* $\Delta A_1, \Delta A_2, \ldots \Delta A_n$*, we have*

$$\left\|\prod_{i=1}^n (A_i + \Delta A_i) - \prod_{i=1}^n A_i\right\| \le \prod_{i=1}^n \left(\|A_i\| + \|\Delta A_i\|\right) - \prod_{i=1}^n \|A_i\|.$$

Proof. We have

$$\begin{aligned} \left\|\prod_{i=1}^n (A_i + \Delta A_i) - \prod_{i=1}^n A_i\right\| &= \left\|\sum_{\delta \in \{0,1\}^n \setminus \{(1,1,\ldots,1)\}} \prod_{i=1}^n A_i^{\delta_i} (\Delta A_i)^{1-\delta_i}\right\| \le \sum_{\delta \in \{0,1\}^n \setminus \{(1,1,\ldots,1)\}} \prod_{i=1}^n \|A_i\|^{\delta_i} \|\Delta A_i\|^{1-\delta_i} \\ &= \prod_{i=1}^n \left(\|A_i\| + \|\Delta A_i\|\right) - \prod_{i=1}^n \|A_i\|. \end{aligned}$$

□

The following lemma gives an upper bound on the 2-norm of M_θ and its derivatives.

Lemma D.10. *We have* $\left\|\Sigma_\theta^{\frac{1}{2}} M_\theta \Sigma_\theta^{-\frac{1}{2}}\right\| \le 1$ *and* $\left\|\Sigma_\theta^{\frac{1}{2}} \left(\nabla_\theta^j M_\theta\right) \Sigma_\theta^{-\frac{1}{2}}\right\| \le G$, $\forall j \in [m]$.

Proof. Note that for any $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $f(s,a) := \mu^\top \phi(s,a)$, and any fixed $h \in \mathbb{N}_+$, we have

$$\begin{aligned} \mathbb{E}^{\pi_\theta}\left[f^2(s_{h+1},a_{h+1})|s_1 \sim \xi_\theta\right] &= \mathbb{E}^{\pi_\theta}\left[\mathbb{E}^{\pi_\theta}\left[f^2(s_{h+1},a_{h+1})|s_h,a_h\right]|s_1 \sim \xi_\theta\right] \\ &\ge \mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[f(s_{h+1},a_{h+1})|s_h,a_h]^2|s_1 \sim \xi_\theta]. \end{aligned}$$

The LHS satisfies

$$\mathbb{E}^{\pi_\theta}[f^2(s_{h+1},a_{h+1})|s_1 \sim \xi_\theta] = \mu^\top \Sigma_\theta \mu,$$

and the RHS satisfies

$$\mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[f(s_{h+1},a_{h+1})|s_h,a_h]^2|s_1 \sim \xi_\theta] = \mathbb{E}^{\pi_\theta}[\mu^\top M_\theta^\top \phi(s_h,a_h)\phi(s_h,a_h)^\top M_\theta \mu|s_1 \sim \xi_\theta] = \mu^\top M_\theta^\top \Sigma_\theta M_\theta \mu.$$

Therefore, we have $\mu^\top \Sigma_\theta \mu \geq \mu^\top M_\theta^\top \Sigma_\theta M_\theta \mu,\ \forall \mu$, which implies $\|\Sigma_\theta^{\frac{1}{2}} M_\theta \Sigma_\theta^{-\frac{1}{2}}\| \leq 1$. Similarly, let $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R},\ g(s,a) := \left(\nabla_\theta^j \log \pi_\theta(s,a)\right) \mu^\top \phi(s,a)$, we have

$$\begin{aligned}\mathbb{E}^{\pi_\theta}[g^2(s_{h+1},a_{h+1})|s_1 \sim \xi_\theta] =& \mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[g^2(s_{h+1},a_{h+1})|s_h,a_h]|s_1 \sim \xi_\theta] \\ \geq& \mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[g(s_{h+1},a_{h+1})|s_h,a_h]^2|s_1 \sim \xi_\theta].\end{aligned}$$

The LHS satisfies

$$\begin{aligned}\mathbb{E}^{\pi_\theta}[g^2(s_{h+1},a_{h+1})|s_1 \sim \xi_\theta] &= \mu^\top \mathbb{E}^{\pi_\theta}\left[\left(\nabla_\theta^j \log \pi_\theta(a|s)\right)^2 \phi(s_{h+1},a_{h+1})\phi(s_{h+1},a_{h+1})^\top |s_1 \sim \xi_\theta\right]\mu \\ &\leq G^2 \mu^\top \mathbb{E}^{\pi_\theta}\left[\phi(s_{h+1},a_{h+1})\phi(s_{h+1},a_{h+1})^\top |s_1 \sim \xi_\theta\right]\mu = G^2 \mu^\top \Sigma_\theta \mu,\end{aligned}$$

and the RHS satisfies

$$\begin{aligned}&\mathbb{E}^{\pi_\theta}[\mathbb{E}^{\pi_\theta}[g(s_{h+1},a_{h+1})|s_h,a_h]^2|s_1 \sim \xi_\theta] \\ =& \mathbb{E}^{\pi_\theta}[\mu^\top \left(\nabla_\theta^j M_\theta\right)^\top \phi(s_h,a_h)\phi(s_h,a_h)^\top \left(\nabla_\theta^j M_\theta\right)\mu | s_1 \sim \xi_\theta] \\ =& \mu^\top \left(\nabla_\theta^j M_\theta\right)^\top \Sigma_\theta \left(\nabla_\theta^j M_\theta\right)\mu.\end{aligned}$$

Therefore, we get $G^2 \mu^\top \Sigma_\theta \mu \geq \mu^\top \left(\nabla_\theta^j M_\theta\right)^\top \Sigma_\theta \left(\nabla_\theta^j M_\theta\right)\mu,\ \forall \mu$, which implies $\left\|\Sigma_\theta^{\frac{1}{2}} \left(\nabla_\theta^j M_\theta\right) \Sigma_\theta^{-\frac{1}{2}}\right\| \leq G$. □

D.5. Probabilistic Events

We define the following probabilistic events: For $j \in [m]$,

$$\begin{aligned}
\mathcal{E}_\Sigma &: \left\|\Sigma^{-\frac{1}{2}}\left(\frac{1}{HK}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\right)\Sigma^{-\frac{1}{2}} - I_d\right\| \leq \sqrt{\frac{2C_1 d \log\frac{8dH}{\delta}}{K}} + \frac{2C_1 d \log\frac{8dH}{\delta}}{3K}, \\
\mathcal{E}_{Y,0} &: \left\|\Sigma^{-\frac{1}{2}}\left(\Delta Y_\theta\right)\Sigma^{-\frac{1}{2}}\right\| \leq (\kappa_2 \vee 1)\sqrt{\frac{2C_1 d \log\frac{16dH}{\delta}}{K}} + \frac{4C_1 d \log\frac{16dH}{\delta}}{3K}, \\
\mathcal{E}_{Y,j} &: \left\|\Sigma^{-\frac{1}{2}}\left(\nabla_\theta^j\left(\Delta Y_\theta\right)\right)\Sigma^{-\frac{1}{2}}\right\| \leq (\kappa_3 \vee 1)\, G\sqrt{\frac{2C_1 d \log\frac{16mdH}{\delta}}{K}} + \frac{4C_1 dG \log\frac{16mdH}{\delta}}{3K}, \\
\mathcal{E}_Y &:= \bigcap_{j=0}^m \mathcal{E}_{Y,j} \\
\mathcal{E}_{\varepsilon,0} &: \left\|\Sigma^{-\frac{1}{2}}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)},a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right\| \leq \frac{\sqrt{d}}{1-\gamma}\left(\sqrt{\frac{2\log\frac{32dH}{\delta}}{KH}} + \frac{2\sqrt{C_1 d}\log\frac{32dH}{\delta}}{K\sqrt{H}} + \frac{2C_1 d\left(\log\frac{32dH}{\delta}\right)^{\frac{3}{2}}}{3K^{\frac{3}{2}}\sqrt{H}}\right) \\
\mathcal{E}_{\varepsilon,j} &: \left\|\Sigma^{-\frac{1}{2}}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)},a_h^{(k)}\right)\nabla_\theta^j \varepsilon_{h,k}^\theta\right\| \leq \frac{2\sqrt{d}G}{(1-\gamma)^2}\left(\sqrt{\frac{2\log\frac{32mdH}{\delta}}{KH}} + \frac{2\sqrt{C_1 d}\log\frac{32mdH}{\delta}}{K\sqrt{H}} + \frac{2C_1 d\left(\log\frac{32mdH}{\delta}\right)^{\frac{3}{2}}}{3K^{\frac{3}{2}}\sqrt{H}}\right), \\
\mathcal{E}_\varepsilon &:= \bigcap_{j=0}^m \mathcal{E}_{\varepsilon,j}, \\
\mathcal{E} &:= \mathcal{E}_\Sigma \cap \mathcal{E}_Y \cap \mathcal{E}_\varepsilon.
\end{aligned}$$

We have the following guarantees on the above high probability events:

Lemma D.11. $\mathbb{P}\left(\mathcal{E}_\Sigma\right) \geq 1 - \frac{\delta}{4}$.

Proof. Define

$$X^{(k)} = \frac{1}{H}\sum_{h=1}^H \Sigma^{-\frac{1}{2}}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top \Sigma^{-\frac{1}{2}} \in \mathbb{R}^{d\times d}.$$

It's easy to see that $X^{(1)}, X^{(2)}, \ldots, X^{(K)}$ are independent and $\mathbb{E}\left[X^{(k)}\right] = I_d$. In the remaining part of the proof, we will apply the matrix Bernstein's inequality to analyze the concentration of $\frac{1}{K}\sum_{k=1}^{K} X^{(k)}$. We first consider the matrix-valued variance $\mathrm{Var}\left(X^{(k)}\right) = \mathbb{E}\left[\left(X^{(k)} - I_d\right)^2\right] = \mathbb{E}\left[\left(X^{(k)}\right)^2\right] - I_d$. Let

$$\Phi^{(k)} := \left[\phi\left(s_1^{(k)}, a_1^{(k)}\right), \phi\left(s_2^{(k)}, a_2^{(k)}\right), \ldots, \phi\left(s_H^{(k)}, a_H^{(k)}\right)\right] \in \mathbb{R}^{d\times H},$$

Then $X^{(k)} = \frac{1}{H}\Sigma^{-\frac{1}{2}}\Phi^{(k)}\left(\Phi^{(k)}\right)^\top \Sigma^{-\frac{1}{2}}$. For any vector $\mu \in \mathbb{R}^d$,

$$\begin{aligned}
\mu^\top \mathbb{E}\left[\left(X^{(k)}\right)^2\right]\mu =& \mathbb{E}\left[\left\|X^{(k)}\mu\right\|^2\right] = \frac{1}{H^2}\mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}\Phi^{(k)}\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2\right] \\
\leq& \frac{1}{H^2}\mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}\Phi^{(k)}\right\|^2\left\|\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2\right] \leq \frac{C_1 d}{H}\mathbb{E}\left[\left\|\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2\right] \\
=& C_1 d \mu^\top \mathbb{E}\left[X^{(k)}\right]\mu = C_1 d\|\mu\|^2,
\end{aligned}$$

where we used the identity $\left\|\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2 = \mu^\top X^{(k)}\mu$ and $\mathbb{E}\left[X^{(k)}\right] = I_d$. We have

$$\mathrm{Var}(X^{(k)}) \preceq \mathbb{E}\left[\left(X^{(k)}\right)^2\right] \preceq C_1 d I_d.$$

Additionally,

$$-I_d \preceq X^{(k)} - I_d = \frac{1}{H}\sum_{h=1}^{H}\Sigma^{-\frac{1}{2}}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\phi\left(s_h^{(k)}, a_h^{(k)}\right)^\top\Sigma^{-\frac{1}{2}} - I_d \preceq C_1 d I_d - I_d.$$

Therefore, $\|X^{(k)} - I_d\| \leq C_1 d$. Since $X^{(1)}, X^{(2)}, \ldots, X^{(K)}$ are *i.i.d.*, by the matrix-form Bernstein inequality, we have

$$\mathbb{P}\left(\left\|\sum_{k=1}^{K} X^{(k)} - I_d\right\| \geq \varepsilon\right) \leq 2d\cdot\exp\left(-\frac{\varepsilon^2/2}{C_1 dK + C_1 d\varepsilon/3}\right), \quad \forall\varepsilon > 0,$$

i.e., with probability at least $1 - \frac{\delta}{4}$,

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\left(X^{(k)} - I_d\right)\right\| \leq \sqrt{\frac{2C_1 d\log\frac{8d}{\delta}}{K}} + \frac{2C_1 d\log\frac{8d}{\delta}}{3K},$$

which has finished the proof. □

Lemma D.12. $\mathbb{P}\left(\mathcal{E}_Y\right) \geq 1 - \frac{\delta}{4}$.

Proof. Take

$$Y_\theta^{(k)} := \frac{1}{H}\sum_{h=1}^{H}\Sigma^{-\frac{1}{2}}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\phi_\theta\left(s_{h+1}^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}, \quad \forall k \in [K].$$

Then, $\Sigma^{-\frac{1}{2}}\left(\Delta Y_\theta\right)\Sigma^{-\frac{1}{2}} = \frac{1}{K}\sum_{k=1}^{K}\left(Y_\theta^{(k)} - \Sigma^{\frac{1}{2}}M_\theta\Sigma^{-\frac{1}{2}}\right)$. Note that

$$\begin{aligned}
\mathbb{E}\left[Y_\theta^{(k)}\right] =& \frac{1}{H}\sum_{h=1}^{H}\mathbb{E}\left[\Sigma^{-\frac{1}{2}}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\phi_\theta\left(s_{h+1}^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\right] \\
=& \frac{1}{H}\sum_{h=1}^{H}\mathbb{E}\left[\Sigma^{-\frac{1}{2}}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\mathbb{E}^{\pi_\theta}\left[\phi\left(s', a'\right)^\top | s_h^{(k)}, a_h^{(k)}\right]\Sigma^{-\frac{1}{2}}\right] \qquad (24) \\
=& \frac{1}{H}\sum_{h=1}^{H}\mathbb{E}\left[\Sigma^{-\frac{1}{2}}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\phi\left(s_h^{(k)}, a_h^{(k)}\right)^\top M_\theta\Sigma^{-\frac{1}{2}}\right] = \Sigma^{\frac{1}{2}}M_\theta\Sigma^{-\frac{1}{2}},
\end{aligned}$$

To this end, $\Sigma^{-\frac{1}{2}}\left(\Delta Y_\theta\right)\Sigma^{-\frac{1}{2}} = \frac{1}{K}\sum_{k=1}^K\left(Y_\theta^{(k)} - \mathbb{E}\left[Y_\theta^{(k)}\right]\right)$. Since the trajectories are i.i.d., we use the matrix-form Bernstein inequality to estimate $\left\|\Sigma^{-\frac{1}{2}}(\Delta Y_\theta)\Sigma^{-\frac{1}{2}}\right\|$. Let

$$
\begin{aligned}
\Phi^{(k)} &:= \left[\phi\left(s_1^{(k)}, a_1^{(k)}\right), \phi\left(s_2^{(k)}, a_2^{(k)}\right), \ldots, \phi\left(s_H^{(k)}, a_H^{(k)}\right)\right] \in \mathbb{R}^{d\times H}, \\
\Phi_\theta^{(k)} &:= \left[\phi_\theta\left(s_2^{(k)}\right), \phi_\theta\left(s_3^{(k)}\right), \ldots, \phi_\theta\left(s_{H+1}^{(k)}\right)\right] \in \mathbb{R}^{d\times H},
\end{aligned}
$$

We have $Y_\theta^{(k)} = \frac{1}{H}\Sigma^{-\frac{1}{2}}\Phi^{(k)}\left(\Phi_\theta^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}$. For any $\mu \in \mathbb{R}^d$, we have

$$
\begin{aligned}
\mu^\top\mathbb{E}\left[Y_\theta^{(k)}\left(Y_\theta^{(k)}\right)^\top\right]\mu =& \mathbb{E}\left[\left\|\left(Y_\theta^{(k)}\right)^\top\mu\right\|^2\right] = \frac{1}{H^2}\mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}\Phi_\theta^{(k)}\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2\right] \\
\leq& \frac{1}{H^2}\mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}\Phi_\theta^{(k)}\right\|^2\left\|\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2\right] \\
\leq& \frac{C_1 d}{H}\mathbb{E}\left[\left\|\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2\right] \\
=& \frac{C_1 d}{H}\mu^\top\Sigma^{-\frac{1}{2}}\mathbb{E}\left[\Phi^{(k)}\left(\Phi^{(k)}\right)^\top\right]\Sigma^{-\frac{1}{2}}\mu \\
=& C_1 d\|\mu\|^2,
\end{aligned}
$$

where we have used the fact $\Sigma = \frac{1}{H}\mathbb{E}\left[\Phi^{(k)}\left(\Phi^{(k)}\right)^\top\right]$. It follows that

$$
\mathrm{Var}_1\left(Y_\theta^{(k)}\right) := \mathbb{E}\left[\left(Y_\theta^{(k)} - \mathbb{E}\left[Y_\theta^{(k)}\right]\right)\left(Y_\theta^{(k)} - \mathbb{E}\left[Y_\theta^{(k)}\right]\right)^\top\right] \preceq \mathbb{E}\left[Y_\theta^{(k)}\left(Y_\theta^{(k)}\right)^\top\right] \preceq C_1 d I_d.
$$

Analogously,

$$
\begin{aligned}
\mathrm{Var}_2\left(Y_\theta^{(k)}\right) :=& \mathbb{E}\left[\left(Y_\theta^{(k)} - \mathbb{E}\left[Y_\theta^{(k)}\right]\right)^\top\left(Y_\theta^{(k)} - \mathbb{E}\left[Y_\theta^{(k)}\right]\right)\right] \preceq \mathbb{E}\left[\left(Y_\theta^{(k)}\right)^\top Y_\theta^{(k)}\right] \\
\preceq& \frac{C_1 d}{H}\Sigma^{-\frac{1}{2}}\mathbb{E}\left[\Phi_\theta^{(k)}\left(\Phi_\theta^{(k)}\right)^\top\right]\Sigma^{-\frac{1}{2}}.
\end{aligned}
$$

Therefore, $\max\left\{\left\|\mathrm{Var}_1\left(Y_\theta^{(k)}\right)\right\|, \left\|\mathrm{Var}_2\left(Y_\theta^{(k)}\right)\right\|\right\} \leq C_1 d\left(\kappa_2^2 \vee 1\right)$. It also holds that $\|Y_\theta^{(k)}\| \leq C_1 d$. Hence,

$$
\left\|Y_\theta^{(k)} - \Sigma^{\frac{1}{2}}M_\theta\Sigma^{-\frac{1}{2}}\right\| \leq 2C_1 d.
$$

Applying Matrix Bernstein's inequality, we derive for any $\varepsilon > 0$,

$$
\mathbb{P}\left(\left\|\sum_{k=1}^K\left(Y_\theta^{(k)} - \Sigma^{\frac{1}{2}}M_\theta\Sigma^{-\frac{1}{2}}\right)\right\| > \varepsilon\right) \leq 2d\exp\left(-\frac{\varepsilon^2/2}{C_1 dK\left(\kappa_2^2 \vee 1\right) + 2C_1 d\varepsilon/3}\right),
$$

which implies $\mathcal{E}_{Y,0}$ holds with probability $1 - \frac{\delta}{8}$. For $\mathcal{E}_{Y,j}, j \in [m]$, notice that for any $j \in [m]$, we have $\Sigma^{-\frac{1}{2}}(\nabla_\theta^j(\Delta Y_\theta))\Sigma^{-\frac{1}{2}} = \frac{1}{K}\sum_{k=1}^K\left(\nabla_\theta^j Y_\theta^{(k)} - \mathbb{E}\left[\nabla_\theta^j Y_\theta^{(k)}\right]\right)$, and $\nabla_\theta^j Y_\theta^{(k)} = \frac{1}{H}\Sigma^{-\frac{1}{2}}\Phi^{(k)}\left(\nabla_\theta^j\Phi_\theta^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}$. For any $\mu \in \mathbb{R}^d$, we have

$$
\begin{aligned}
\mu^\top\mathbb{E}\left[\left(\nabla_\theta^j Y_\theta^{(k)}\right)\left(\nabla_\theta^j Y_\theta^{(k)}\right)^\top\right]\mu =& \frac{1}{H^2}\mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}\left(\nabla_\theta^j\Phi_\theta^{(k)}\right)\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2\right] \\
\leq& \frac{1}{H^2}\mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}\nabla_\theta^j\Phi_\theta^{(k)}\right\|^2\left\|\left(\Phi^{(k)}\right)^\top\Sigma^{-\frac{1}{2}}\mu\right\|^2\right].
\end{aligned}
$$

Since we have

$$\begin{aligned}
&\left(\nabla_\theta^j \phi_\theta\left(s_{h+1}^{(k)}\right)\right)^\top \Sigma^{-1} \nabla_\theta^j \phi_\theta\left(s_{h+1}^{(k)}\right) \\
=&\int_{\mathcal{A}\times\mathcal{A}} \pi_\theta\left(a \left| s_{h+1}^{(k)}\right.\right) \pi_\theta\left(a' \left| s_{h+1}^{(k)}\right.\right)\left(\nabla_\theta^j \log \pi_\theta\left(a \left| s_{h+1}^{(k)}\right.\right)\right)\left(\nabla_\theta^j \log \pi_\theta\left(a' \left| s_{h+1}^{(k)}\right.\right)\right) \\
&\cdot \phi\left(s_{h+1}^{(k)}, a\right)^\top \Sigma^{-1} \phi\left(s_{h+1}^{(k)}, a'\right) \mathrm{d}a\mathrm{d}a' \\
\leq& G^2 \int_{\mathcal{A}\times\mathcal{A}} \pi_\theta\left(a | s_{h+1}^{(k)}\right) \pi_\theta\left(a' | s_{h+1}^{(k)}\right)\left\|\Sigma^{-\frac{1}{2}} \phi\left(s_{h+1}^{(k)}, a\right)\right\|\left\|\Sigma^{-\frac{1}{2}} \phi\left(s_{h+1}^{(k)}, a'\right)\right\| \mathrm{d}a\mathrm{d}a' \leq G^2 C_1 d,
\end{aligned}$$

which implies

$$\mu^\top \mathbb{E}\left[\left(\nabla_\theta^j Y_\theta^{(k)}\right)\left(\nabla_\theta^j Y_\theta^{(k)}\right)^\top\right] \mu \leq \frac{G^2 C_1 d}{H} \mathbb{E}\left[\left\|\left(\Phi^{(k)}\right)^\top \Sigma^{-\frac{1}{2}} \mu\right\|^2\right] = G^2 C_1 d \|\mu\|^2.$$

Therefore,

$$\mathrm{Var}_1\left(\nabla_\theta^j Y_\theta^{(k)}\right) := \mathbb{E}\left[\left(\nabla_\theta^j Y_\theta^{(k)} - \mathbb{E}\left[\nabla_\theta^j Y_\theta^{(k)}\right]\right)\left(\nabla_\theta^j Y_\theta^{(k)} - \mathbb{E}\left[\nabla_\theta^j Y_\theta^{(k)}\right]\right)^\top\right] \preceq \mathbb{E}\left[\left(\nabla_\theta^j Y_\theta^{(k)}\right)\left(\nabla_\theta^j Y_\theta^{(k)}\right)^\top\right] \preceq G^2 C_1 d I_d.$$

Meanwhile, we have

$$\mathrm{Var}_2\left(\nabla_\theta^j Y_\theta^{(k)}\right) \preceq \mathbb{E}\left[\left(\nabla_\theta^j Y_\theta^{(k)}\right)^\top \nabla_\theta^j Y_\theta^{(k)}\right] \preceq \frac{C_1 d}{H} \Sigma^{-\frac{1}{2}} \mathbb{E}\left[\left(\nabla_\theta^j \Phi_\theta^{(k)}\right)\left(\nabla_\theta^j \Phi_\theta^{(k)}\right)^\top\right] \Sigma^{-\frac{1}{2}}.$$

In conclusion, we get

$$\max\left\{\left\|\mathrm{Var}_1\left(\nabla_\theta^j Y_\theta^{(k)}\right)\right\|, \left\|\mathrm{Var}_2\left(\nabla_\theta^j Y_\theta^{(k)}\right)\right\|\right\} \leq G^2 C_1 d\left(\kappa_3^2 \vee 1\right),$$

Note that $\left\|\nabla_\theta^j Y_\theta^{(k)}\right\| \leq C_1 d G$, we know $\left\|\nabla_\theta^j Y_\theta^{(k)} - \mathbb{E}\left[\nabla_\theta^j Y_\theta^{(k)}\right]\right\| \leq 2 C_1 d G$. By Matrix Bernstein's inequality, we get for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^K \left(\nabla_\theta^j Y_k^\theta - \Sigma^{\frac{1}{2}}\left(\nabla_\theta^j M_\theta\right)\Sigma^{-\frac{1}{2}}\right)\right\| \geq \varepsilon\right) \leq 2d \exp\left(-\frac{\varepsilon^2/2}{G^2 C_1 d K\left(\kappa_3^2 \vee 1\right) + 2 C_1 d G \varepsilon/3}\right),$$

taking a union bound over all $j \in [m]$ proves that $\bigcap_{j=1}^m \mathcal{E}_{Y,j}$ holds with probability $1 - \frac{\delta}{8}$. Using a union bound argument again, we know with probability $1 - \frac{\delta}{4}$, $\mathcal{E}_Y$ holds, which has finished the proof. □

Lemma D.13. $\mathbb{P}(\mathcal{E}_\varepsilon) \geq 1 - \frac{\delta}{4}$.

Proof. Let $X_{\theta,h}^{(k)} := \Sigma^{-\frac{1}{2}} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^\theta \in \mathbb{R}^d$ and let $\mathcal{F}_{h,k}$ be σ-algebra generated by the history up to step h at episode k, we have $\mathbb{E}\left[X_{\theta,h}^{(k)} \middle| \mathcal{F}_{h,k}\right] = 0$. We apply matrix-form Freedman's inequality to analyze the concentration property. Consider conditional variances $\mathrm{Var}_1\left[X_{\theta,h}^{(k)} \middle| \mathcal{F}_{h,k}\right] := \mathbb{E}\left[X_{\theta,h}^{(k)}\left(X_{\theta,h}^{(k)}\right)^\top \middle| \mathcal{F}_{h,k}\right] \in \mathbb{R}^{d\times d}$ and $\mathrm{Var}_2\left[X_{\theta,h}^{(k)} \middle| \mathcal{F}_{h,k}\right] := \mathbb{E}\left[\left(X_{\theta,h}^{(k)}\right)^\top X_{\theta,h}^{(k)} | \mathcal{F}_{h,k}\right] \in \mathbb{R}$. It holds that

$$\begin{aligned}
\left\|\mathrm{Var}_1\left[X_{\theta,h}^{(k)} \middle| \mathcal{F}_{h,k}\right]\right\| =& \left\|\mathbb{E}\left[X_{\theta,h}^{(k)}\left(X_{\theta,h}^{(k)}\right)^\top \middle| \mathcal{F}_{h,k}\right]\right\| \leq \mathbb{E}\left[\left\|X_{\theta,h}^{(k)}\left(X_{\theta,h}^{(k)}\right)^\top\right\| \middle| \mathcal{F}_{h,k}\right] \\
=& \mathbb{E}\left[\left\|X_{\theta,h}^{(k)}\right\|^2 \middle| \mathcal{F}_{h,k}\right] = \mathrm{Var}_2\left[X_{\theta,h}^{(k)} \middle| \mathcal{F}_{h,k}\right]
\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]=&\mathbb{E}\left[\left\|X_{\theta,h}^{(k)}\right\|^2\middle|\mathcal{F}_{h,k}\right]=\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right)\mathrm{Var}\left[\varepsilon_{h,k}^{\theta}\middle|s_h^{(k)},a_h^{(k)}\right]\\\le&\frac{1}{(1-\gamma)^2}\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right),\end{aligned}$$

where we have used $\mathrm{Var}\left[\varepsilon_{h,k}^{\theta}\middle|s_h^{(k)},a_h^{(k)}\right]\le\frac{1}{(1-\gamma)^2}$. Note that

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right)\\=&KHd+KH\mathrm{Tr}\left(\Sigma^{-\frac12}\left(\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\right)\Sigma^{-\frac12}-I_d\right)\\\le&KHd+KHd\left\|\Sigma^{-\frac12}\left(\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)},a_h^{(k)}\right)\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\right)\Sigma^{-\frac12}-I_d\right\|.\end{aligned}$$

We take

$$\sigma^2:=\frac{KHd}{(1-\gamma)^2}\left(1+\sqrt{\frac{2C_1d\log\frac{32dH}{\delta}}{K}}+\frac{2C_1d\log\frac{32dH}{\delta}}{3K}\right).\tag{25}$$

The result of Lemma D.11 implies that

$$\mathbb{P}\left(\left\|\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|\le\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\le\sigma^2\right)\ge1-\frac{\delta}{16}.\tag{26}$$

Additionally, we have $\left\|X_{\theta,h}^{(k)}\right\|\le\frac{\sqrt{C_1d}}{1-\gamma}$. The Freedman's inequality therefore implies that for any $\varepsilon>0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{k=1}^{K}\sum_{h=1}^{H}X_{\theta,h}^{(k)}\right|\ge\varepsilon,\ \max\left\{\left\|\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|,\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\}\le\sigma^2\right)\\\le&2d\exp\left(-\frac{\varepsilon^2/2}{\sigma^2+\sqrt{C_1d}\varepsilon/(3(1-\gamma))}\right),\end{aligned}\tag{27}$$

where σ^2 is defined in (25). We take

$$\varepsilon:=\sigma\sqrt{2\log\frac{32d}{\delta}}+\frac{2\sqrt{C_1d}}{3(1-\gamma)}\log\frac{32d}{\delta}.$$

Then we get

$$\mathbb{P}\left(\left|\sum_{k=1}^{K}\sum_{h=1}^{H}X_{\theta,h}^{(k)}\right|\ge\varepsilon,\ \max\left\{\left\|\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|,\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\}\le\sigma^2\right)\le\frac{\delta}{16},$$

which implies

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{k=1}^{K}\sum_{h=1}^{H}X_{\theta,h}^{(k)}\right|\ge\varepsilon\right)\\\le&\mathbb{P}\left(\left|\sum_{k=1}^{K}\sum_{h=1}^{H}X_{\theta,h}^{(k)}\right|\ge\varepsilon,\ \max\left\{\left\|\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|,\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\}\le\sigma^2\right)\\&+\mathbb{P}\left(\max\left\{\left\|\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_1\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\|,\sum_{k=1}^{K}\sum_{h=1}^{H}\mathrm{Var}_2\left[X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\}>\sigma^2\right)\le\frac{\delta}{8}.\end{aligned}$$

which has proved $\mathbb{P}\left(\mathcal{E}_{\varepsilon,0}\right) \geq 1-\frac{\delta}{8}$. For any fixed $j \in [m]$, we use Freedman's inequality again to prove $\mathbb{P}\left(\mathcal{E}_{\varepsilon,j}\right) \geq 1-\frac{\delta}{8m}$. We have

$$\left\|\operatorname{Var}_1\left[\nabla_\theta^j X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right]\right\| = \left\|\mathbb{E}\left[\left(\nabla_\theta^j X_{\theta,h}^{(k)}\right)\left(\nabla_\theta^j X_{\theta,h}^{(k)}\right)^\top\middle|\mathcal{F}_{h,k}\right]\right\| \leq \mathbb{E}\left[\left\|\left(\nabla_\theta^j X_{\theta,h}^{(k)}\right)\left(\nabla_\theta^j X_{\theta,h}^{(k)}\right)^\top\right\|\middle|\mathcal{F}_{h,k}\right]$$
$$= \mathbb{E}\left[\left\|\nabla_\theta^j X_{\theta,h}^{(k)}\right\|^2\middle|\mathcal{F}_{h,k}\right] = \operatorname{Var}_2\left[\nabla_\theta^j X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right],$$

and

$$\operatorname{Var}_2\left[\nabla_\theta^j X_{\theta,h}^{(k)}\middle|\mathcal{F}_{h,k}\right] = \mathbb{E}\left[\left\|\nabla_\theta^j X_{\theta,h}^{(k)}\right\|^2\middle|\mathcal{F}_{h,k}\right] = \phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right)\operatorname{Var}\left[\nabla_\theta^j\varepsilon_{h,k}^\theta\middle|s_h^{(k)},a_h^{(k)}\right]$$
$$\leq \frac{4G^2}{(1-\gamma)^2}\phi\left(s_h^{(k)},a_h^{(k)}\right)^\top\Sigma^{-1}\phi\left(s_h^{(k)},a_h^{(k)}\right),$$

where we have used $\operatorname{Var}\left[\nabla_\theta^j\varepsilon_{h,k}^\theta\middle|s_h^{(k)},a_h^{(k)}\right] \leq \frac{4G^2}{(1-\gamma)^2}$. Furthermore, notice that $\left\|\nabla_\theta^j X_{\theta,h}^{(k)}\right\| \leq \frac{2G\sqrt{C_1 d}}{1-\gamma}$, the remaining steps will be exactly the same as those in the proof of the case $\mathcal{E}_{\varepsilon,0}$. Taking a union bound over $j \in [m]$ and $\mathcal{E}_{\varepsilon,0}$, we have proved $\mathbb{P}\left(\bigcap_{j=0}^m \mathcal{E}_{\varepsilon,j}\right) \geq 1-\frac{\delta}{4}$, which has finished the proof. □

Combining the results of Lemma D.11, Lemma D.12, Lemma D.13 and take a union bound, we conclude

$$\mathbb{P}\left(\mathcal{E}\right) \geq 1-\frac{3}{4}\delta.$$

Next, we prove some immediate results when the event $\mathcal{E}$ holds.

Lemma D.14. *When $\mathcal{E}_\Sigma$ holds and*

$$K \geq C_1 d\log\frac{8dmH}{\delta}, \quad \lambda \leq C_1 d\sigma_{min}(\Sigma)\cdot\log\frac{8dmH}{\delta},$$

we have

$$\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\| \leq 4\sqrt{\frac{C_1 d\log\frac{8dmH}{\delta}}{K}}.$$

Proof. Note that

$$\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\| = \left\|\Sigma^{\frac{1}{2}}\left(\widehat{\Sigma}^{-1}-\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\| \leq \left\|\Sigma^{\frac{1}{2}}\widehat{\Sigma}^{-1}\Sigma^{\frac{1}{2}}\right\|\left\|\Sigma^{-\frac{1}{2}}\widehat{\Sigma}\Sigma^{-\frac{1}{2}}-I_d\right\|. \tag{28}$$

When $\mathcal{E}_\Sigma$ holds, with the condition

$$K \geq C_1 d\log\frac{8dmH}{\delta}, \quad \lambda \leq C_1 d\sigma_{\min}(\Sigma)\cdot\log\frac{8dmH}{\delta},$$

we have

$$\left\|\Sigma^{-\frac{1}{2}}\widehat{\Sigma}\Sigma^{-\frac{1}{2}}-I_d\right\| \leq \sqrt{\frac{2C_1 d\log\frac{8dH}{\delta}}{K}}+\frac{2C_1 d\log\frac{8dH}{\delta}}{3K}+\frac{\lambda\|\Sigma^{-1}\|}{K} \leq 2\sqrt{\frac{C_1 d\log\frac{8dmH}{\delta}}{K}} \leq \frac{1}{2},$$

which further implies $\sigma_{\min}\left(\Sigma^{-\frac{1}{2}}\widehat{\Sigma}\Sigma^{-\frac{1}{2}}\right) \geq \frac{1}{2}$, and $\left\|\Sigma^{\frac{1}{2}}\widehat{\Sigma}^{-1}\Sigma^{\frac{1}{2}}\right\| \leq 2$. Combining this result with (28), we get

$$\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\| \leq 4\sqrt{\frac{C_1 d\log\frac{8dmH}{\delta}}{K}}.$$

which has finished the proof. □

Lemma D.15. *When $\mathcal{E}_\Sigma$ and $\mathcal{E}_Y$ hold, and*

$$K \geq 36\kappa_1(4+\kappa_2+\kappa_3)^2 \frac{C_1 d}{(1-\gamma)^2}\log\frac{16dmH}{\delta}, \quad \lambda \leq C_1 d\sigma_{min}(\Sigma)\cdot\log\frac{8dmH}{\delta},$$

we have

$$\left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq 6\sqrt{\kappa_1}(4+\kappa_2+\kappa_3)\sqrt{\frac{C_1 d\log\frac{16dmH}{\delta}}{K}},$$

and

$$\left\|\Sigma_\theta^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_\theta\right)}{G}\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq 6\sqrt{\kappa_1}(4+\kappa_2+\kappa_3)\sqrt{\frac{C_1 d\log\frac{16dmH}{\delta}}{K}}, \quad \forall j\in[m].$$

Proof. We have

$$\begin{aligned}\left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\| &\leq \left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\left\|\Sigma^{\frac{1}{2}}\Sigma_\theta^{-\frac{1}{2}}\right\|\left\|\Sigma^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma^{-\frac{1}{2}}\right\| = \sqrt{\kappa_1}\left\|\Sigma^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma^{-\frac{1}{2}}\right\| \\ &= \sqrt{\kappa_1}\left\|\Sigma^{\frac{1}{2}}\left(\widehat{\Sigma}^{-1}\frac{1}{HK}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)},a_h^{(k)}\right)\int_{\mathcal{A}}\phi\left(s_{h+1}^{(k)},a'\right)\pi_\theta\left(a'\middle|s_{h+1}^{(k)}\right)\mathrm{d}a' - M_\theta\right)\Sigma^{-\frac{1}{2}}\right\| \\ &\leq \sqrt{\kappa_1}\left(\left(1+\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\|\right)\left(1+\left\|\Sigma^{-\frac{1}{2}}\left(\Delta Y_\theta\right)\Sigma^{-\frac{1}{2}}\right\|\right)-1\right),\end{aligned}$$

where $\Delta Y_\theta = \frac{1}{HK}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)},a_h^{(k)}\right)\phi_\theta\left(s_{h+1}^{(k)}\right)^\top - \Sigma M_\theta$ and the last inequality uses Lemma D.9. Similarly, we have

$$\left\|\Sigma_\theta^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_\theta\right)}{G}\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq \sqrt{\kappa_1}\left(\left(1+\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma\right)\Sigma^{\frac{1}{2}}\right\|\right)\left(1+\left\|\Sigma^{-\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta Y_\theta\right)}{G}\right)\Sigma^{-\frac{1}{2}}\right\|\right)-1\right).$$

Using the result of Lemma D.14 the event $\mathcal{E}_Y$, we get

$$\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\| \leq 4\sqrt{\frac{C_1 d\log\frac{8dmH}{\delta}}{K}} \leq 1,$$

and $\forall j\in[m]$,

$$\left\|\Sigma^{\frac{1}{2}}\left(\Delta Y_\theta\right)\Sigma^{-\frac{1}{2}}\right\| \leq 2(\kappa_2+1)\sqrt{\frac{C_1 d\log\frac{16dmH}{\delta}}{K}} \leq 1, \quad \left\|\Sigma^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta Y_\theta\right)}{G}\right)\Sigma^{-\frac{1}{2}}\right\| \leq 2(\kappa_3+1)\sqrt{\frac{C_1 d\log\frac{16dmH}{\delta}}{K}} \leq 1,$$

which implies

$$\left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq 2\sqrt{\kappa_1}\left(\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\|+\left\|\Sigma^{\frac{1}{2}}\left(\Delta Y_\theta\right)\Sigma^{-\frac{1}{2}}\right\|\right) \leq 6\sqrt{\kappa_1}(4+\kappa_2+\kappa_3)\sqrt{\frac{C_1 d\log\frac{16dmH}{\delta}}{K}},$$

where we use the fact $(1+x_1)(1+x_2)-1 \leq 2(x_1+x_2)$ whenever $x_1,x_2\in[0,1]$. Similarly, we get

$$\left\|\Sigma_\theta^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_\theta\right)}{G}\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq 6\sqrt{\kappa_1}(4+\kappa_2+\kappa_3)\sqrt{\frac{C_1 d\log\frac{16dmH}{\delta}}{K}}, \quad \forall j\in[m],$$

which has finished the proof. ∎

Lemma D.16. *When $\mathcal{E}_\Sigma$ and $\mathcal{E}_Y$ hold, and*

$$K \geq 36\kappa_1(4+\kappa_2+\kappa_3)^2 \frac{C_1 d}{(1-\gamma)^2}\log\frac{16dmH}{\delta}, \quad \lambda \leq C_1 d\sigma_{min}(\Sigma)\cdot\log\frac{8dmH}{\delta},$$

we have

$$\widehat{Q}^\theta = \sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\widehat{r}, \quad \widehat{\nabla_\theta Q^\theta} = \sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\tilde{U}^\theta.$$

Proof. Firstly, note that given the conditions, the result of Lemma D.15 implies

$$\left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq 6\sqrt{\kappa_1}(4+\kappa_2+\kappa_3)\sqrt{\frac{C_1 d\log\frac{16dmH}{\delta}}{K}} \leq 1-\gamma.$$

Therefore,

$$\|\gamma\Sigma_\theta^{\frac{1}{2}}\widehat{M}_\theta\Sigma_\theta^{-\frac{1}{2}}\| \leq \gamma\left(\left\|\Sigma_\theta^{\frac{1}{2}}\widehat{M}_\theta\Sigma_\theta^{-\frac{1}{2}}\right\| + \left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta\widehat{M}_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\|\right) \leq \gamma(2-\gamma) < 1.$$

where we use the result of Lemma D.10 to get $\left\|\Sigma_\theta^{\frac{1}{2}}\widehat{M}_\theta\Sigma_\theta^{-\frac{1}{2}}\right\| \leq 1$. Therefore, we have

$$\left(I_d - \gamma\widehat{M}_\theta\right)^{-1} = \Sigma_\theta^{-\frac{1}{2}}\left(I_d - \gamma\Sigma_\theta^{\frac{1}{2}}\widehat{M}_\theta\Sigma_\theta^{-\frac{1}{2}}\right)^{-1}\Sigma_\theta^{\frac{1}{2}} = \Sigma_\theta^{-\frac{1}{2}}\sum_{h=1}^{\infty}\gamma^{h-1}\left(\Sigma_\theta^{\frac{1}{2}}\widehat{M}_\theta\Sigma_\theta^{-\frac{1}{2}}\right)^{h-1}\Sigma_\theta^{\frac{1}{2}} = \sum_{h=1}^{\infty}\gamma^{h-1}\widehat{M}_\theta^{h-1}.$$

Based on this result, we prove the main result by definition:

$$\widehat{Q}^\theta(\cdot,\cdot) = \phi(\cdot,\cdot)^\top\left(I_d - \gamma\widehat{M}_\theta\right)^{-1}\widehat{w}_r = \phi(\cdot,\cdot)^\top\sum_{h=1}^{\infty}\gamma^{h-1}\widehat{M}_\theta^{h-1}\widehat{w}_r = \left(\sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\widehat{r}\right)(\cdot,\cdot),$$

$$\widehat{\nabla_\theta Q^\theta}(\cdot,\cdot) = \phi(\cdot,\cdot)^\top\left(I_d - \gamma\widehat{M}_\theta\right)^{-1}\widehat{\nabla_\theta^j M_\theta}\widehat{w}^\theta = \phi(\cdot,\cdot)^\top\sum_{h=1}^{\infty}\gamma^{h-1}\widehat{M}_\theta^{h-1}\widehat{\nabla_\theta^j M_\theta}\widehat{w}^\theta = \left(\sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\tilde{U}^\theta\right)(\cdot,\cdot),$$

which has finished the proof. ☐

Now we consider the decomposition of $Q^\theta - \widehat{Q}^\theta$:

Lemma D.17. *Under the same condition of Lemma D.16, we have*

$$Q^\theta - \widehat{Q}^\theta = \sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\left(Q^\theta - \widehat{r} - \gamma\widehat{\mathcal{P}}_\theta Q^\theta\right).$$

Proof. Simply note that

$$\begin{aligned}
Q^\theta - \widehat{Q}^\theta &= \sum_{h=1}^{\infty}\gamma^{h-1}\left(\mathcal{P}_\theta\right)^{h-1}r - \sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\widehat{r} \\
&= \sum_{h=1}^{\infty}\gamma^{h-1}\left(\left(\mathcal{P}_\theta\right)^{h-1} - \left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\right)r + \sum_{h'=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}(r-\widehat{r}) \\
&= \sum_{h=1}^{\infty}\gamma^{h-1}\sum_{h'=1}^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h'-1}\left(\mathcal{P}_\theta - \widehat{\mathcal{P}}_\theta\right)\left(\mathcal{P}_\theta\right)^{h-h'-1}r + \sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}(r-\widehat{r}) \\
&= \sum_{h'=1}^{\infty}\left(\widehat{\mathcal{P}}_\theta\right)^{h'-1}\left(\mathcal{P}_\theta - \widehat{\mathcal{P}}_\theta\right)\sum_{h=h'+1}^{\infty}\gamma^{h-1}\left(\mathcal{P}_\theta\right)^{h-h'-1}r + \sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}(r-\widehat{r}) \\
&= \sum_{h=1}^{\infty}\gamma^{h}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\left(\mathcal{P}_\theta - \widehat{\mathcal{P}}_\theta\right)Q^\theta + \sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}(r-\widehat{r}) \\
&= \sum_{h=1}^{\infty}\gamma^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h-1}\left(Q^\theta - \widehat{r} - \gamma\widehat{\mathcal{P}}_\theta Q^\theta\right),
\end{aligned}$$

which is the desired result. □

D.6. Proofs of Main Theorems

Define $\widehat{\nu}_h^{\theta} := \left(\widehat{M}_{\theta}^{\top}\right)^{h-1} \nu_1^{\theta}$ and $\widehat{\nu}^{\theta} = \sum_{h=1}^{\infty} \gamma^{h-1} \widehat{\nu}_h^{\theta}$. We may prove the following decomposition of $\nabla_{\theta} v_{\theta} - \widehat{\nabla_{\theta} v_{\theta}}$:

Lemma D.18. *Given the same condition of Lemma D.16, we have* $\nabla_{\theta} v_{\theta} - \widehat{\nabla_{\theta} v_{\theta}} = E_1 + E_2 + E_3$*, where*

$$
\begin{aligned}
E_1 =& \nabla_{\theta}\left[\left(\nu^{\theta}\right)^{\top} \Sigma^{-1} \frac{1}{KH} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^{\theta}\right] \\
E_2 =& \nabla_{\theta}\left[\left(\left(\widehat{\nu}^{\theta}\right)^{\top} \widehat{\Sigma}^{-1} - \left(\nu^{\theta}\right)^{\top} \Sigma^{-1}\right) \frac{1}{KH} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^{\theta}\right] \\
E_3 =& \frac{\lambda}{KH} \sum_{h=1}^{T} \nabla_{\theta}\left[\left(\widehat{\nu}^{\theta}\right)^{\top} \widehat{\Sigma}^{-1} w^{\theta}\right].
\end{aligned}
$$

The proof of Lemma D.18 is deferred to appendix D.7. Based on this observation, here we show the proofs of our main theorems.

D.6.1. Proof of Theorem D.3

Proof. We use Lemma D.18 to decompose $\langle \nabla_{\theta} v_{\theta} - \widehat{\nabla_{\theta} v_{\theta}}, t\rangle = \langle E_1, t\rangle + \langle E_2, t\rangle + \langle E_3, t\rangle$. To bound each term individually, we introduce the following lemmas, whose proofs are deferred to appendix D.7.

Lemma D.19. *For any* $t \in \mathbb{R}^m$*, with probability* $1 - \frac{\delta}{4}$*, we have*

$$
|\langle E_1, t\rangle| \leq \sqrt{\frac{2t^{\top} \Lambda_{\theta} t \log(8/\delta)}{HK}} + \frac{2\log(8/\delta)\sqrt{C_1 md}\|t\| B}{3HK}.
$$

where $B = \frac{1}{1-\gamma} \max_{j\in[m]} \sqrt{\left(\nabla_{\theta}^{j} \nu^{\theta}\right)^{\top} \Sigma^{-1} \nabla_{\theta}^{j} \nu^{\theta}} + \frac{2G}{(1-\gamma)^2} \sqrt{\left(\nu^{\theta}\right)^{\top} \Sigma^{-1} \nu^{\theta}}$.

Lemma D.20. *Let* E_2^j *be the* j*th entry of* E_2*, suppose* $\mathcal{E}$ *holds and* $K \geq 36\kappa_1(4+\kappa_2+\kappa_3)^2 C_1 d(1-\gamma)^{-2} \log\frac{16dmH}{\delta}$ *and* $\lambda \leq C_1 d\sigma_{\min}(\Sigma) \log\frac{8dmH}{\delta}$*, then we have*

$$
|E_2^j| \leq \frac{240\sqrt{\kappa_1}(2+\kappa_2+\kappa_3)\sqrt{C_1}d}{(1-\gamma)^3} \left(\left\|\Sigma_{\theta}^{-\frac{1}{2}} \nabla_{\theta}^{j} \nu_1^{\theta}\right\| + \frac{G}{1-\gamma}\left\|\Sigma_{\theta}^{-\frac{1}{2}} \nu_1^{\theta}\right\|\right) \left\|\Sigma_{\theta}^{\frac{1}{2}} \Sigma^{-\frac{1}{2}}\right\| \frac{\log\frac{32dmH}{\delta}}{KH}, \quad \forall j \in [m].
$$

Lemma D.21. *Let* E_3^j *be the* j*th entry of* E_3*, suppose* $\mathcal{E}$ *holds and* $K \geq 36\kappa_1(4+\kappa_2+\kappa_3)^2 C_1 d(1-\gamma)^{-2} \log\frac{16dmH}{\delta}$ *and* $\lambda \leq C_1 d\sigma_{\min}(\Sigma) \log\frac{8dmH}{\delta}$*, we have*

$$
|E_3^j| \leq \frac{6C_1 d}{(1-\gamma)^2} \left\|\Sigma_{\theta}^{\frac{1}{2}} \Sigma^{-\frac{1}{2}}\right\| \left(\left\|\Sigma_{\theta}^{-\frac{1}{2}} \nabla_{\theta}^{j} \nu_1^{\theta}\right\| + \frac{G}{1-\gamma}\left\|\Sigma_{\theta}^{-\frac{1}{2}} \nu_1^{\theta}\right\|\right) \frac{\log\frac{8dmH}{\delta}}{KH}, \quad \forall j \in [m].
$$

Let $B_1^j = \frac{1}{1-\gamma}\sqrt{\left(\nabla_{\theta}^{j} \nu^{\theta}\right)^{\top} \Sigma^{-1} \nabla_{\theta}^{j} \nu^{\theta}}$, $B_2 = \frac{G}{(1-\gamma)^2}\sqrt{\left(\nu^{\theta}\right)^{\top} \Sigma^{-1} \nu^{\theta}}$, then we have the relation $B = \max_{j\in[m]} B_1^j + 2B_2$.

For any $j \in [m]$, note that

$$\begin{aligned}
B_1^j =& \frac{1}{1-\gamma} \left\| \Sigma^{-\frac{1}{2}} \nabla_\theta^j \nu^\theta \right\| \leq \sum_{h=1}^{\infty} \frac{\gamma^{h-1}}{1-\gamma} \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left\| \Sigma_\theta^{-\frac{1}{2}} \nabla_\theta^j \left(\left(M_\theta^{h-1} \right)^\top \nu_1^\theta \right) \right\| \\
=& \sum_{h=1}^{\infty} \frac{\gamma^{h-1}}{1-\gamma} \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \\
& \cdot \left(\left\| \Sigma_\theta^{\frac{1}{2}} M_\theta \Sigma_\theta^{-\frac{1}{2}} \right\|^{h-1} \left\| \Sigma_\theta^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta \right\| + (h-1) \left\| \Sigma_\theta^{\frac{1}{2}} M_\theta \Sigma_\theta^{-\frac{1}{2}} \right\|^{h-2} \left\| \Sigma_\theta^{\frac{1}{2}} \left(\nabla_\theta^j M_\theta \right) \Sigma_\theta^{-\frac{1}{2}} \right\| \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \right) \\
\leq& \frac{1}{(1-\gamma)^2} \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left(\left\| \Sigma_\theta^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta \right\| + \frac{G}{1-\gamma} \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \right),
\end{aligned}$$

where we use the result of Lemma D.10. Similarly,

$$\begin{aligned}
B_2 \leq& \sum_{h=1}^{\infty} \frac{G\gamma^{h-1}}{(1-\gamma)^2} \left\| \Sigma^{-\frac{1}{2}} \nu_h^\theta \right\| \leq \sum_{h=1}^{\infty} \frac{G\gamma^{h-1}}{(1-\gamma)^2} \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left\| \Sigma_\theta^{\frac{1}{2}} M_\theta \Sigma_\theta^{-\frac{1}{2}} \right\|^{h-1} \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \\
\leq& \sum_{h=1}^{\infty} \frac{G\gamma^{h-1}}{(1-\gamma)^2} \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \leq \frac{G}{(1-\gamma)^3} \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\|.
\end{aligned}$$

We conclude that when $K \geq 36 C_1 d(1-\gamma)^{-2} \kappa_1 (4+\kappa_2+\kappa_3)^2 \log \frac{16dmH}{\delta}$, we have

$$\begin{aligned}
& \left(\max_{j\in[m]} B_1^j + 2B_2 \right) \frac{2\log\frac{2}{\delta} \sqrt{C_1 md} \|t\|}{KH} \\
\leq& \frac{1}{(1-\gamma)^2} \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left(\max_{j\in[m]} \left\| \Sigma_\theta^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta \right\| + 2HG \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \right) \frac{2\log\frac{2}{\delta} \sqrt{C_1 md} \|t\|}{KH},
\end{aligned}$$

Now, take a union bound on $\mathcal{E}$ and the event in Lemma D.19, we get with probability $1-\delta$, we have

$$\begin{aligned}
& |\langle E_1, t\rangle| + |\langle E_2, t\rangle| + |\langle E_3, t\rangle| \leq \sqrt{\frac{2t^\top \Lambda_\theta t \log(8/\delta)}{HK}} \\
& + \sqrt{\kappa_1}(5+\kappa_2+\kappa_3) \left(\max_{j\in[m]} \left\| \Sigma_\theta^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta \right\| + \frac{G}{1-\gamma} \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \right) \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \frac{240 C_1 d \log \frac{32dmH}{\delta}}{(1-\gamma)^3 KH} \\
\leq& \sqrt{\frac{2t^\top \Lambda_\theta t \log(8/\delta)}{KH}} + \kappa_1(5+\kappa_2+\kappa_3) \left(\max_{j\in[m]} \left\| \Sigma_\theta^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta \right\| + \frac{G}{1-\gamma} \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \right) \frac{240 C_1 d\sqrt{m} \|t\| \log \frac{32dmH}{\delta}}{(1-\gamma)^3 HK}.
\end{aligned}$$

we have finished the proof. ∎

D.6.2. Proof of Theorem D.4

Proof. According to the result of Theorem D.3, we know

$$|\langle t, \widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta \rangle| \leq \sqrt{\frac{2t^\top \Lambda_\theta t}{HK} \cdot \log\frac{8}{\delta}} + \frac{C_\theta \|t\| \log\frac{32mdH}{\delta}}{HK},$$

Pick $t = e_j, j \in [m]$, we have

$$\begin{aligned}
t^\top \Lambda_\theta t =& \mathbb{E}\left[\frac{1}{H} \sum_{h=1}^{H} \left(\nabla_\theta^j \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)} \right)^\top \Sigma^{-1} \nu^\theta \right) \right)^2 \right] \\
\leq& 2\mathbb{E}\left[\frac{1}{H} \sum_{h=1}^{H} \left(\nabla_\theta^j \varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)} \right)^\top \Sigma^{-1} \nu^\theta \right)^2 \right] + 2\mathbb{E}\left[\frac{1}{H} \sum_{h=1}^{H} \left(\varepsilon_{h,1}^\theta \phi\left(s_h^{(1)}, a_h^{(1)} \right)^\top \Sigma^{-1} \nabla_\theta^j \nu^\theta \right)^2 \right] \\
\leq& 2\mathbb{E}\left[\sum_{h=1}^{H} \frac{G^2}{H(1-\gamma)^4} \left(\phi\left(s_h^{(1)}, a_h^{(1)} \right)^\top \Sigma^{-1} \nu^\theta \right)^2 \right] + 2\mathbb{E}\left[\sum_{h=1}^{H} \frac{1}{H(1-\gamma)^2} \left(\phi\left(s_h^{(1)}, a_h^{(1)} \right)^\top \Sigma^{-1} \nabla_\theta^j \nu^\theta \right)^2 \right] \\
\leq& 2\left(\frac{G^2}{(1-\gamma)^4} \left\| \Sigma^{-\frac{1}{2}} \nu^\theta \right\|^2 + \frac{1}{(1-\gamma)^2} \left\| \Sigma^{-\frac{1}{2}} \nabla_\theta^j \nu^\theta \right\|^2 \right).
\end{aligned}$$

Therefore, take a union bound over $j \in [m]$, we get

$$\left|\widehat{\nabla_\theta^j v_\theta} - \nabla_\theta^j v_\theta\right| \leq 4b_\theta\sqrt{\frac{\log\frac{8m}{\delta}}{HK}} + \frac{2C_\theta\sqrt{m}\log\frac{32mdH}{\delta}}{HK}, \quad \forall j \in [m],$$

where

$$b_\theta = \frac{G}{(1-\gamma)^2}\left\|\Sigma^{-\frac{1}{2}}\nu^\theta\right\| + \frac{1}{1-\gamma}\left\|\Sigma^{-\frac{1}{2}}\nabla_\theta^j\nu^\theta\right\|.$$

When $\phi(s',a')^\top\Sigma^{-1}\phi(s,a) \geq 0,\ \forall (s,a),(s',a') \in \mathcal{S}\times\mathcal{A}$, for any $h \in \mathbb{N}_+$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}
\left|\left(\nabla_\theta^j \nu_h^\theta\right)^\top \Sigma^{-1}\phi(s,a)\right| =& \left|\mathbb{E}^{\pi_\theta}\left[\phi(s_h,a_h)\Sigma^{-1}\phi(s,a)\sum_{h'=1}^{h}\nabla_\theta^j \log\pi_{\theta,h'}(a_{h'}|s_{h'})\right]\right| \\
\leq& \mathbb{E}^{\pi_\theta}\left[\phi(s_h,a_h)\Sigma^{-1}\phi(s,a)\sum_{h'=1}^{h}\left|\nabla_\theta^j \log\pi_{\theta,h'}(a_{h'}|s_{h'})\right|\right] \\
\leq& Gh\mathbb{E}^{\pi_\theta}\left[\phi(s_h,a_h)\Sigma^{-1}\phi(s,a)\right] \\
=& Gh\left(\nu_h^\theta\right)^\top\Sigma^{-1}\phi(s,a).
\end{aligned}$$

Meanwhile, for any positive integer $\tilde{H}$, we have

$$\begin{aligned}
\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nabla_\theta^j\nu^\theta &= \sum_{h=1}^{\infty}\gamma^{h-1}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nabla_\theta^j\nu_h^\theta \\
&= \sum_{h=1}^{\tilde{H}}\gamma^{h-1}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nabla_\theta^j\nu_h^\theta + \sum_{h=\tilde{H}+1}^{\infty}\gamma^{h-1}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nabla_\theta^j\nu_h^\theta.
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
\sum_{h=\tilde{H}+1}^{\infty}\gamma^{h-1}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nabla_\theta^j\nu_h^\theta &\leq \sum_{h=\tilde{H}+1}^{\infty}Gh\gamma^{h-1}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nu_h^\theta \\
&\leq GC_1d\sum_{h=\tilde{H}+1}^{\infty}h\gamma^{h-1} \\
&= GC_1d\left(\tilde{H}+\frac{1}{1-\gamma}\right)\gamma^{\tilde{H}}.
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\sum_{h=1}^{\tilde{H}}\gamma^{h-1}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nabla_\theta^j\nu_h^\theta &\leq G\tilde{H}\sum_{h=1}^{\tilde{H}}\gamma^{h-1}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nu_h^\theta \leq G\tilde{H}\sum_{h=1}^{\infty}\gamma^{h-1}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nu_h^\theta \\
&= G\tilde{H}\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nu^\theta.
\end{aligned}$$

Therefore,

$$\begin{aligned}
t^\top\Lambda_\theta t =& 2\mathbb{E}\left[\sum_{h=1}^{H}\frac{G^2}{H(1-\gamma)^4}\left(\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nu^\theta\right)^2\right] + 2\mathbb{E}\left[\sum_{h=1}^{H}\frac{1}{H(1-\gamma)^2}\left(\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nabla_\theta^j\nu^\theta\right)^2\right] \\
\leq& 4\mathbb{E}\left[\sum_{h=1}^{H}\frac{G^2}{H(1-\gamma)^2}\left(\tilde{H}^2+\frac{1}{(1-\gamma)^2}\right)\left(\phi\left(s_h^{(1)},a_h^{(1)}\right)^\top\Sigma^{-1}\nu^\theta\right)^2\right] + \frac{4G^2C_1^2d^2}{(1-\gamma)^2}\left(\tilde{H}+\frac{1}{1-\gamma}\right)^2\gamma^{2\tilde{H}} \\
\leq& \frac{4G^2}{(1-\gamma)^2}\left(\tilde{H}+\frac{1}{1-\gamma}\right)^2\left(\left\|\Sigma^{-\frac{1}{2}}\nu^\theta\right\|^2 + C_1^2d^2\gamma^{2\tilde{H}}\right).
\end{aligned}$$

In particular, pick

$$\tilde{H} = \frac{2\log(C_1 d)}{1-\gamma} \geq \frac{\log \frac{C_1 d}{\left\|\Sigma^{-\frac{1}{2}}\nu^\theta\right\|}}{\log \frac{1}{\gamma}},$$

where we use the fact $\log \frac{1}{\gamma} \geq \frac{1-\gamma}{2\gamma}$ whenever $\gamma \in (\frac{1}{2}, 1)$, and

$$\left(\nu^\theta\right)^\top \Sigma^{-1}\nu^\theta = \sup_{w\in\mathbb{R}^d} \frac{\left(w^\top \nu^\theta\right)^2}{w^\top \Sigma w} \geq \frac{1}{(1-\gamma)^2} \geq 1,$$

where the inequality is due to the fact that $\mathcal{F}$ includes the constant functions. We have

$$t^\top \Lambda_\theta t \leq \frac{8G^2}{(1-\gamma)^2}\left(\tilde{H} + \frac{1}{(1-\gamma)}\right)^2 \left\|\Sigma^{-\frac{1}{2}}\nu^\theta\right\|^2 \leq \frac{128G^2\left(\log(C_1 d)\right)^2}{(1-\gamma)^4}\left\|\Sigma^{-\frac{1}{2}}\nu^\theta\right\|^2.$$

Taking a union bound w.r.t. m, we get

$$\left|\widehat{\nabla^j_\theta v_\theta} - \nabla^j_\theta v_\theta\right| \leq \frac{16G\log(C_1 d)}{(1-\gamma)^2}\left\|\Sigma^{-\frac{1}{2}}\nu^\theta\right\|\sqrt{\frac{\log\frac{8m}{\delta}}{HK}} + \frac{2C_\theta\sqrt{m}\log\frac{32mdH}{\delta}}{HK}.$$

□

D.6.3. PROOF OF THEOREM D.5

Proof. We use the same decomposition as in Theorem D.3. Define a martingale difference sequence $\{e_k^\theta\}_{k=1}^K$ by

$$\begin{aligned}
e^\theta_{h,k} &= \frac{1}{\sqrt{HK}}\nabla_\theta\left(\left(\nu^\theta\right)^\top\Sigma^{-1}\phi(s_h^{(k)}, a_h^{(k)})\varepsilon^\theta_{h,k}\right) \\
&= \frac{1}{\sqrt{HK}}\left(\nabla_\theta\nu^\theta\right)^\top\Sigma^{-1}\phi(s_h^{(k)}, a_h^{(k)})\varepsilon^\theta_{h,k} + \frac{1}{\sqrt{HK}}\sum_{h=1}^H\left(\nu^\theta\right)^\top\Sigma^{-1}\phi(s_h^{(k)}, a_h^{(k)})\nabla_\theta\varepsilon^\theta_{h,k},
\end{aligned}$$

we have

$$\|e^\theta_{h,k}\|_\infty \leq \frac{1}{\sqrt{HK}(1-\gamma)}\max_{j\in[m]}\|\Sigma^{-\frac{1}{2}}\nabla^j_\theta\nu^\theta\|\sqrt{C_1 d} + \frac{2}{\sqrt{HK}(1-\gamma)^2}\|\Sigma^{-\frac{1}{2}}\nu^\theta\|\sqrt{C_1 d}G \to 0,$$

where we use the result of Lemma D.8. Furthermore,

$$\begin{aligned}
\sum_{h=1}^H \mathbb{E}\left[e^\theta_{h,k}\left(e^\theta_{h,k}\right)^\top\right]_{ij} =& \frac{1}{HK}\mathbb{E}\left[\sum_{h=1}^H\left[\nabla^i_{\theta_1}\left(\left(\nu^{\theta_1}\right)^\top\Sigma^{-1}\phi(s_h^{(k)}, a_h^{(k)})\varepsilon^{\theta_1}_{h,k}\right)\right]\left[\nabla^j_{\theta_2}\left(\left(\nu^{\theta_2}\right)^\top\Sigma^{-1}\phi(s_h^{(k)}, a_h^{(k)})\varepsilon^{\theta_2}_{h,k}\right)\right]^\top\right]\Bigg|_{\theta_1=\theta_2=\theta} \\
=& \frac{[\Lambda_\theta]_{ij}}{K}.
\end{aligned}$$

Therefore,by WLLN, we have

$$\sum_{k=1}^K\sum_{h=1}^H\left[e^\theta_{h,k}\left(e^\theta_{h,k}\right)^\top\right]_{ij} \to_p \sum_{k=1}^K\sum_{h=1}^H\mathbb{E}\left[e^\theta_{h,k}\left(e^\theta_{h,k}\right)^\top\right]_{ij} = [\Lambda_\theta]_{ij},$$

To finish the rest of the proof, we introduce the following lemmas,

Lemma D.22 (Martingale CLT, Corollary 2.8 in (McLeish et al., 1974))**.** *Let $\{X_{mn}, n = 1, \ldots, k_m\}$ be a martingale difference array (row-wise) on the probability triple $(\Omega, \mathcal{F}, P)$.Suppose X_{mn} satisfy the following two conditions:*

$$\max_{1\leq n\leq k_m}|X_{mn}| \xrightarrow{p} 0, \text{ and } \sum_{n=1}^{k_m} X_{mn}^2 \xrightarrow{p} \sigma^2$$

for $k_m \to \infty$. Then $\sum_{n=1}^{k_m} X_{mn} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$.

Lemma D.23 (Cramér–Wold Theorem)**.** *Let $X_n = (X_n^1, X_n^2, \ldots, X_n^k)^\top$ be a k-dimensional random vector series and $X = (X^1, X^2, \ldots, X^k)^\top$ be a random vector of same dimension. Then X_n converges in distribution to X if and only if for any constant vector $t = (t_1, t_2, \ldots, t_k)^\top$, $t^\top X_n$ converges to $t^\top X$ in distribution.*

Lemma D.22 implies $\sum_{k=1}^K \sum_{h=1}^H t^\top e_{h,k} \to_d \mathcal{N}(0, t^\top \Lambda_\theta t)$ for any t, and Lemma D.23 implies

$$\sqrt{HK} E_1 = \sum_{k=1}^K \sum_{h=1}^H e_{h,k} \to_p \mathcal{N}(0, \Lambda_\theta).$$

Furthermore, notice that the results of Lemma D.20 and Lemma D.21 imply $\sqrt{HK} E_2 \to_p 0, \sqrt{HK} E_3 \to_p 0$. Combining the above results, we have finished the proof. □

D.6.4. Proof of Theorem D.6

Proof. We first derive the influence function of policy gradient estimator for sake of completeness. We denote each of the K sampled trajectories as

$$\boldsymbol{\tau} := (s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_H, a_H, r_H, s_{H+1})$$

We denote $\bar{\pi}(a \mid s)$ as the behavior policy. The distribution of trajectory is then given by

$$\mathcal{P}(d\boldsymbol{\tau}) = \bar{\xi}\left(ds_1, da_1\right) p\left(ds_2 \mid s_1, a_1\right) \bar{\pi}\left(da_2 \mid s_2\right) \ldots \bar{\pi}\left(da_H \mid s_H\right) p\left(ds_{H+1} \mid s_H, a_H\right)$$

Define $p_\eta = p + \eta \Delta p$ as a new transition probability function and $\mathcal{P}_\eta := \mathcal{P} + \eta \Delta \mathcal{P}$ where $\Delta \mathcal{P}$ satisfies

$$(\Delta \mathcal{P}) \mathcal{F} \subseteq \mathcal{F}.$$

Define $g_\eta\left(s' \mid s, a\right) := \frac{\partial}{\partial \eta} \log p_\eta\left(s' \mid s, a\right)$ and the score function as

$$g_\eta(\boldsymbol{\tau}) := \frac{\partial}{\partial \eta} \log \mathcal{P}_\eta(d\boldsymbol{\tau}) = \sum_{h=1}^H g_\eta\left(s_{h+1} \mid s_h, a_h\right).$$

Without loss of generality, we assume p_η is continuously derivative with respect to η. This guarantees that we can change the order of taking derivatives with respect to η and θ. When the subscript η vanishes, it means $\eta = 0$ and the underlying transition probability is $p(s'|s,a)$, i.e. $p_0(s'|s,a) = p(s'|s,a)$. Then we denote $g(s'|s,a) := \frac{\partial}{\partial \eta} \log p_\eta(s'|s,a)\Big|_{\eta=0}$, and $g(\boldsymbol{\tau}) = \sum_{h=1}^H g(s_{h+1}|s_h, a_h)$. We define the policy value under new transition kernel is

$$v_{\theta,\eta} := \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^\infty \gamma^{h-1} r(s_h, a_h) \middle| s_1 \sim \xi, \mathcal{P}_\eta\right]$$

Then, our objective function is

$$\psi_\eta := \nabla_\theta v_{\theta,\eta} = \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^\infty \nabla_\theta \log \pi_\theta\left(a_h \mid s_h\right) \cdot \left(\sum_{h'=h}^\infty \gamma^{h'-1} r\left(s_{h'}, a_{h'}\right)\right) \middle| s_1 \sim \xi, \mathcal{P}_\eta\right].$$

We are going to compute the influence function with respect to the above objective function. We denote this influence function as $\mathcal{I}(\boldsymbol{\tau})$. By definition, it satisfies that

$$\frac{\partial}{\partial \eta} \psi_\eta \bigg|_{\eta=0} = \mathbb{E}\left[g(\boldsymbol{\tau}) \mathcal{I}(\boldsymbol{\tau})\right].$$

By exchanging the order of derivatives, we find that

$$\frac{\partial}{\partial \eta} \psi_\eta \bigg|_{\eta=0} = \nabla_\theta \left[\frac{\partial}{\partial \eta} v_{\theta,\eta} \bigg|_{\eta=0}\right].$$

Therefore, we calculate the derivatives.

$$\begin{aligned}
\frac{\partial}{\partial \eta} v_{\theta,\eta} &= \frac{\partial}{\partial \eta}\left[\sum_{h=1}^{\infty} \gamma^{h-1} \int_{(\mathcal{S}\times\mathcal{A})^h} r\left(s_h, a_h\right) \xi(s_1) \prod_{j=1}^{h-1} p_\eta\left(s_{j+1} \mid s_j, a_j\right) \prod_{j=1}^{h} \pi_\theta\left(a_j \mid s_j\right) d\boldsymbol{\tau}_h\right] \\
&= \sum_{h=1}^{\infty} \gamma^{h-1} \int_{(\mathcal{S}\times\mathcal{A})^h} r\left(s_h, a_h\right)\left(\sum_{j=1}^{h-1} g_\eta\left(s_{j+1} \mid s_j, a_j\right)\right) \xi(s_1) \prod_{j=1}^{h-1} p_\eta\left(s_{j+1} \mid s_j, a_j\right) \prod_{j=1}^{h} \pi_\theta\left(a_j \mid s_j\right) d\boldsymbol{\tau}_h \\
&= \int \sum_{h=1}^{\infty} \gamma^{h-1} r\left(s_h, a_h\right)\left(\sum_{j=1}^{h-1} g_\eta\left(s_{j+1} \mid s_j, a_j\right)\right)\left[\xi(s_1) \prod_{j=1}^{h} p_\eta\left(s_{j+1} \mid s_j, a_j\right) \prod_{j=1}^{h} \pi_\theta\left(a_j \mid s_j\right)\right] d\boldsymbol{\tau}.
\end{aligned}$$

We denote Q_η^θ and $\nabla_\theta Q_\eta^\theta$ as the state-action function and its gradient with underlying transition probability being p_η. For sake of simplicity, we define the state value function as

$$V^\theta(s) := \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{\infty} \gamma^{h-1} r(s_h, a_h) \middle| s_1 = s, \mathcal{P}\right].$$

We denote $V_\eta^\theta(s)$ as the same function except for transition probability substituted by p_η. Therefore,

$$\begin{aligned}
\frac{\partial}{\partial \eta} v_{\theta,\eta} &= \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{\infty} \gamma^{h-1} r\left(s_h, a_h\right)\left(\sum_{j=1}^{h-1} g_\eta\left(s_{j+1} \mid s_j, a_j\right)\right) \middle| s_1 \sim \xi, \mathcal{P}_\eta\right] \\
&= \mathbb{E}^{\pi_\theta}\left[\sum_{j=1}^{\infty} g_\eta\left(s_{j+1} \mid s_j, a_j\right) \sum_{h=j+1}^{\infty} \gamma^{h-1} r\left(s_h, a_h\right) \middle| s_1 \sim \xi, \mathcal{P}_\eta\right] \\
&= \mathbb{E}^{\pi_\theta}\left[\sum_{j=1}^{\infty} g_\eta\left(s_{j+1} \mid s_j, a_j\right) \cdot \mathbb{E}^{\pi_\theta}\left[\sum_{h=j+1}^{\infty} \gamma^{h-1} r\left(s_h, a_h\right) \middle| s_{j+1}\right] \middle| s_1 \sim \xi, \mathcal{P}_\eta\right] \\
&= \mathbb{E}^{\pi_\theta}\left[\sum_{j=1}^{\infty} \mathbb{E}\left[\gamma^j g_\eta\left(s_{j+1} \mid s_j, a_j\right) V_\eta^\theta(s_{j+1}) \middle| s_j, a_j\right] \middle| s_1 \sim \xi, \mathcal{P}_\eta\right].
\end{aligned}$$

Therefore,

$$\left.\frac{\partial}{\partial \eta} v_{\theta,\eta}\right|_{\eta=0} = \mathbb{E}^{\pi_\theta}\left[\sum_{h=1}^{\infty} \mathbb{E}\left[\gamma^h g\left(s_{h+1} \mid s_h, a_h\right) V^\theta(s_{h+1}) \middle| s_h, a_h\right] \middle| s_1 \sim \xi, \mathcal{P}_\eta\right]. \tag{29}$$

We notice that $\Sigma = \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H} \phi\left(s_h^{(1)}, a_h^{(1)}\right)\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top\right]$. We denote $w_h(s,a) := \phi^\top(s,a)\Sigma^{-1}\nu_h^\theta = \phi^\top(s,a)\Sigma^{-1}\mathbb{E}^{\pi_\theta}\left[\phi(s_h, a_h) \mid s_1 \sim \xi\right]$. We leverage the following fact to rewrite (29): for any $f(s,a) = w_f^\top \phi(s,a) \in \mathcal{F}$ where $w_f \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbb{E}^{\pi_\theta}\left[f(s_h, a_h)\right] &= \mathbb{E}^{\pi_\theta}\left[w_f^\top \phi(s_h, a_h)\right] \\
&= \mathbb{E}^{\pi_\theta}\left[w_f^\top \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^{H} \phi\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\phi^\top\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\right]\Sigma^{-1}\phi(s_h, a_h)\right] \\
&= \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^{H} w_f^\top \phi\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\phi^\top\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\Sigma^{-1}\mathbb{E}^{\pi_\theta}\left[\phi(s_h, a_h)\right]\right] \\
&= \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^{H} f\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right) w_h\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\right]
\end{aligned}$$

Since

$$\mathbb{E}\left[g\left(s' \mid s, a\right) V^{\theta}(s')|s,a\right] = \frac{\partial}{\partial \eta}\left(Q_{\eta}^{\theta}(s,a) - r_{\eta}(s,a)\right)\Big|_{\eta=0} \in \mathcal{F},$$

we have

$$\begin{aligned}\frac{\partial}{\partial \eta} v_{\theta,\eta}\Big|_{\eta=0} &= \mathbb{E}\left[\sum_{h=1}^{\infty}\gamma^{h}\frac{1}{H}\sum_{h'=1}^{H} w_h\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\mathbb{E}\left[g\left(s' \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right)\cdot V^{\theta}\left(s'\right) \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right]\right] \\ &= \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^{H}\mathbb{E}_{s'\sim p(\cdot|s_{h'}^{(1)},a_{h'}^{(1)})}\left[\sum_{h=1}^{\infty}\gamma^{h} w_h\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right) g\left(s' \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right)\cdot V^{\theta}\left(s'\right)\right]\right] \\ &= \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^{H}\mathbb{E}_{s'\sim p(\cdot|s_{h'}^{(1)},a_{h'}^{(1)})}\left[\sum_{h=1}^{\infty}\gamma^{h} w_h\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right) g\left(s' \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right)\left(V^{\theta}\left(s'\right) - \mathbb{E}\left[V^{\theta}\left(s'\right) \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right]\right)\right]\right] \\ &= \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^{H}\sum_{h=1}^{\infty}\gamma^{h} w_h\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right) g\left(s_{h'+1}^{(1)} \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right)\left(V^{\theta}\left(s_{h'+1}^{(1)}\right) - \mathbb{E}\left[V^{\theta}\left(s_{h'+1}^{(1)}\right) \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right]\right)\right] \\ &= \mathbb{E}\left[g\left(\boldsymbol{\tau}\right)\frac{1}{H}\sum_{h'=1}^{H}\sum_{h=1}^{\infty}\gamma^{h} w_h\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\left(V^{\theta}\left(s_{h'+1}^{(1)}\right) - \mathbb{E}\left[V^{\theta}\left(s_{h'+1}^{(1)}\right) \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right]\right)\right].\end{aligned}$$

Taking gradient in both sides and we have

$$\nabla_{\theta}\left(\frac{\partial}{\partial \eta} v_{\theta,\eta}\Big|_{\eta=0}\right) = \mathbb{E}\left\{g\left(\boldsymbol{\tau}\right)\cdot\nabla_{\theta}\left[\frac{1}{H}\sum_{h'=1}^{H}\sum_{h=1}^{H}\gamma^{h} w_h\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\left(V^{\theta}\left(s_{h'+1}^{(1)}\right) - \mathbb{E}\left[V^{\theta}\left(s_{h'+1}^{(1)}\right) \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right]\right)\right]\right\}.$$

The implies that the influence function we want is

$$\mathcal{I}(\boldsymbol{\tau}) = \nabla_{\theta}\left[\frac{1}{H}\sum_{h'=1}^{H}\sum_{h=1}^{\infty}\gamma^{h} w_h\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)\left(V^{\theta}\left(s_{h'+1}^{(1)}\right) - \mathbb{E}\left[V^{\theta}\left(s_{h'+1}^{(1)}\right) \mid s_{h'}^{(1)}, a_{h'}^{(1)}\right]\right)\right].$$

Insert the expression of $w_h(s,a)$ and exploit $\varepsilon_{h,k}^{\theta} = Q^{\theta}(s_h^{(k)}, a_h^{(k)}) - r_h^{(k)} - \gamma\int_{\mathcal{A}}\pi_{\theta}\left(a' \mid s_{h+1}^{(k)}\right) Q^{\theta}\left(s_{h+1}^{(k)}, a'\right)\mathrm{d}a'$, we can rewrite the influence function as

$$\mathcal{I}(\boldsymbol{\tau}) = -\nabla_{\theta}\left[\frac{1}{H}\sum_{h'=1}^{H}\sum_{h=1}^{\infty}\gamma^{h-1}\phi\left(s_{h'}^{(1)}, a_{h'}^{(1)}\right)^{\top}\Sigma^{-1}\varepsilon_{h',1}^{\theta}\nu_h^{\theta}\right] = -\nabla_{\theta}\left[\frac{1}{H}\sum_{h=1}^{H}\phi\left(s_{h}^{(1)}, a_{h}^{(1)}\right)^{\top}\Sigma^{-1}\varepsilon_{h,1}^{\theta}\nu^{\theta}\right]$$

Therefore, since the cross terms vanish by taking conditional expectation, we have

$$\mathbb{E}\left[\mathcal{I}(\boldsymbol{\tau})^{\top}\mathcal{I}(\boldsymbol{\tau})\right] = \mathbb{E}\left[\frac{1}{H^2}\sum_{h=1}^{H}\left(\nabla_{\theta}\left(\varepsilon_{h,1}^{\theta}\phi\left(s_h^{(1)}, a_h^{(1)}\right)^{\top}\Sigma^{-1}\nu^{\theta}\right)\right)^{\top}\nabla_{\theta}\left(\varepsilon_{h,1}^{\theta}\phi\left(s_h^{(1)}, a_h^{(1)}\right)^{\top}\Sigma^{-1}\nu^{\theta}\right)\right] = \frac{1}{H}\Lambda_{\theta}.$$

For any vector $t \in \mathbb{R}^m$, when it comes to $\langle t, \psi_{\eta}\rangle$, by linearity we have

$$\frac{\partial}{\partial \eta}\langle t, \psi_{\eta}\rangle\Big|_{\eta=0} = \mathbb{E}[g(\boldsymbol{\tau})\langle t, \mathcal{I}(\boldsymbol{\tau})\rangle].$$

Then the influence function of $\langle t, \nabla_{\theta} v_{\theta}\rangle$ is $\langle t, \mathcal{I}(\boldsymbol{\tau})\rangle$. The Cramer-Rao lower bound for $\langle t, \nabla_{\theta} v_{\theta}\rangle$ is

$$\mathbb{E}\left[\langle t, \mathcal{I}(\boldsymbol{\tau})\rangle^2\right] = t^{\top}\mathbb{E}\left[\mathcal{I}(\boldsymbol{\tau})^{\top}\mathcal{I}(\boldsymbol{\tau})\right] t = \frac{1}{H} t^{\top}\Lambda_{\theta} t.$$

By continuous mapping theorem, a trivial corollary of Theorem D.6 is that for any $t \in \mathbb{R}^m$,

$$\sqrt{HK}\left(\left\langle t, \widehat{\nabla_{\theta} v_{\theta}} - \nabla_{\theta} v_{\theta}\right\rangle\right) \xrightarrow{d} \mathcal{N}\left(0, t^{\top}\Lambda_{\theta} t\right).$$

This implies that the variance of any unbiased estimator for $\langle t, \nabla_{\theta} v_{\theta}\rangle \in \mathbb{R}$ is lower bounded by $\frac{1}{\sqrt{HK}} t^{\top}\Lambda_{\theta} t$. □

D.7. Missing Proofs

D.7.1. Proof of Lemma D.18

Proof. Note that

$$\begin{aligned}
\nabla_\theta Q^\theta - \widehat{\nabla_\theta Q^\theta} &= \sum_{h=1}^{\infty} \gamma^{h-1} \left(\mathcal{P}_\theta\right)^{h-1} U^\theta - \sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \tilde{U}^\theta \\
&= \sum_{h=1}^{\infty} \gamma^{h-1} \left(\mathcal{P}_\theta\right)^{h-1} U^\theta - \sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \widehat{U}^\theta + \sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \left(\widehat{U}^\theta - \tilde{U}^\theta\right) \\
&= \sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \left(\nabla_\theta Q^\theta - \widehat{U}^\theta - \widehat{\mathcal{P}}_\theta \nabla_\theta Q^\theta\right) + \sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \left(\widehat{U}^\theta - \tilde{U}^\theta\right).
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \left(\nabla_\theta Q^\theta - \widehat{U}^\theta - \gamma \widehat{\mathcal{P}}_\theta \nabla_\theta Q^\theta\right) \\
&= \sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \phi^\top \widehat{\Sigma}^{-1} \frac{1}{KH} \sum_{k=1}^{K} \sum_{h'=1}^{H} \phi\left(s_{h'}^{(k)}, a_{h'}^{(k)}\right) \\
&\cdot \left(\nabla_\theta Q^\theta\left(s_{h'}^{(k)}, a_{h'}^{(k)}\right) - \gamma \int_{\mathcal{A}} \left(\left(\nabla_\theta \pi_\theta\left(a' \middle| s_{h'+1}^{(k)}\right)\right) Q^\theta\left(s_{h'+1}^{(k)}, a'\right) + \pi_\theta\left(a' \middle| s_{h'+1}^{(k)}\right) \nabla_\theta Q^\theta\left(s_{h'+1}^{(k)}, a'\right)\right) \mathrm{d}a'\right) \\
&+ \frac{\lambda}{KH} \sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \phi^\top \widehat{\Sigma}^{-1} \nabla_\theta w^\theta \\
&= \sum_{h=1}^{\infty} \gamma^{h-1} \phi^\top \left(\widehat{M}_\theta\right)^{h-1} \widehat{\Sigma}^{-1} \frac{1}{KH} \sum_{k=1}^{K} \sum_{h'=1}^{H} \phi\left(s_{h'}^{(k)}, a_{h'}^{(k)}\right) \nabla_\theta \varepsilon_{h',k}^\theta + \frac{\lambda}{KH} \sum_{h=1}^{\infty} \gamma^{h-1} \phi^\top \left(\widehat{M}_\theta\right)^{h-1} \widehat{\Sigma}^{-1} \nabla_\theta w^\theta.
\end{aligned}$$

Using the definition of $\widehat{\nu}^\theta$, we get

$$\begin{aligned}
&\int_{\mathcal{S}\times\mathcal{A}} \xi(s) \pi_\theta(a|s) \left(\sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \left(\nabla_\theta Q^\theta - \widehat{U}^\theta - \gamma \widehat{\mathcal{P}}_\theta \nabla_\theta Q^\theta\right)\right)(s,a) \mathrm{d}s \mathrm{d}a \\
&= \left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} \frac{1}{KH} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \nabla_\theta \varepsilon_{h,k}^\theta + \frac{\lambda}{KH} \left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} \nabla_\theta w^\theta.
\end{aligned} \tag{30}$$

For the second term, by Lemma D.17, we have

$$\begin{aligned}
\sum_{h=1}^{\infty} \gamma^{h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h-1} \left(\widehat{U}^\theta - \tilde{U}^\theta\right) &= \sum_{h=1}^{\infty} \gamma^{h} \left(\widehat{\mathcal{P}}_\theta\right)^{h} \left(\nabla_\theta \log \Pi_\theta\right) \left(Q^\theta - \widehat{Q}^\theta\right) \\
&= \sum_{h=1}^{\infty} \gamma^{h} \left(\widehat{\mathcal{P}}_\theta\right)^{h} \left(\nabla_\theta \log \Pi_\theta\right) \sum_{h'=h}^{\infty} \gamma^{h'-h-1} \left(\widehat{\mathcal{P}}_\theta\right)^{h'-h-1} \left(Q^\theta - \widehat{r} - \gamma \widehat{\mathcal{P}}_\theta Q^\theta\right) \\
&= \sum_{h=1}^{\infty} \sum_{h'=1}^{h-1} \left(\gamma \widehat{\mathcal{P}}_\theta\right)^{h'} \left(\nabla_\theta \log \Pi_\theta\right) \left(\gamma \widehat{\mathcal{P}}_\theta\right)^{h-h'-1} \left(Q^\theta - \widehat{r} - \gamma \widehat{\mathcal{P}}_\theta Q^\theta\right).
\end{aligned}$$

Meanwhile, again by Lemma D.17, we have

$$\left(\nabla_\theta \log \Pi_\theta\right) \left(Q^\theta - \widehat{Q}^\theta\right) = \left(\nabla_\theta \log \Pi_\theta\right) \sum_{h=1}^{\infty} \left(\gamma \widehat{\mathcal{P}}_\theta\right)^{h-1} \left(Q^\theta - \widehat{r} - \gamma \widehat{\mathcal{P}}_\theta Q^\theta\right),$$

which implies

$$
\begin{aligned}
&\sum_{h=1}^{\infty}\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h-1}\left(\widehat{U}^\theta-\tilde{U}^\theta\right)+\left(\nabla_\theta\log\Pi_\theta\right)\left(Q^\theta-\widehat{Q}^\theta\right)\\
=&\sum_{h=1}^{\infty}\left(\left(\nabla_\theta\log\Pi_\theta\right)\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h-1}+\sum_{h'=1}^{h-1}\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h'}\left(\nabla_\theta\log\Pi_\theta\right)\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h-h'-1}\right)\left(Q^\theta-\widehat{r}-\gamma\widehat{\mathcal{P}}_\theta Q^\theta\right)\\
=&\sum_{h=1}^{\infty}\sum_{h'=0}^{h-1}\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h'}\left(\nabla_\theta\log\Pi_\theta\right)\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h-h'-1}\left(Q^\theta-\widehat{r}-\gamma\widehat{\mathcal{P}}_\theta Q^\theta\right)\\
=&\sum_{h=1}^{\infty}\sum_{h'=0}^{h-1}\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h'}\left(\nabla_\theta\log\Pi_\theta\right)\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h-h'-1}\phi^\top\widehat{\Sigma}^{-1}\\
&\cdot\frac{1}{KH}\sum_{k=1}^{K}\sum_{h'=1}^{H}\phi\left(s_{h'}^{(k)},a_{h'}^{(k)}\right)\left(Q^\theta\left(s_{h'}^{(k)},a_{h'}^{(k)}\right)-r_{h'}^{(k)}-\gamma\int_{\mathcal{A}}\pi_\theta\left(a'\,\middle|s_{h'+1}^{(k)}\right)Q^\theta\left(s_{h'+1}^{(k)},a'\right)\mathrm{d}a'\right)\\
&+\frac{\lambda}{KH}\sum_{h=1}^{\infty}\sum_{h'=0}^{h-1}\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h'}\left(\nabla_\theta\log\Pi_\theta\right)\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h-h'-1}\phi^\top\widehat{\Sigma}^{-1}w^\theta.
\end{aligned}
$$

For each $j\in[m]$, notice the relation

$$
\begin{aligned}
\left(\nabla_\theta^j\widehat{\nu}_h^\theta\right)^\top&=\left(\nabla_\theta^j\nu_1^\theta\right)^\top\left(\widehat{M}_\theta\right)^{h-1}+\sum_{h'=1}^{h-1}\left(\nu_1^\theta\right)^\top\left(\widehat{M}_\theta\right)^{h'-1}\left(\widehat{\nabla_\theta^j M_\theta}\right)\left(\widehat{M}_\theta\right)^{h-h'-1}\\
&=\int_{\mathcal{S}\times\mathcal{A}}\xi(s)\pi_\theta(a|s)\left(\sum_{h'=0}^{h-1}\left(\widehat{\mathcal{P}}_\theta\right)^{h'}\left(\nabla_\theta\log\Pi_\theta\right)\left(\widehat{\mathcal{P}}_\theta\right)^{h-h'-1}\phi^\top\right)(s,a)\mathrm{d}s\mathrm{d}a.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\left[\int\xi(s)\pi_\theta(a|s)\left(\sum_{h=1}^{\infty}\left(\gamma\widehat{\mathcal{P}}_\theta\right)^{h-1}\left(\widehat{U}^\theta-\tilde{U}^\theta\right)+\left(\nabla_\theta\log\Pi_\theta\right)\left(Q^\theta-\widehat{Q}^\theta\right)\right)(s,a)\mathrm{d}s\mathrm{d}a\right]_j\\
=&\sum_{h=1}^{\infty}\gamma^{h-1}\left(\nabla_\theta^j\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^{K}\sum_{h'=1}^{H}\phi\left(s_{h'}^{(k)},a_{h'}^{(k)}\right)\left(Q^\theta\left(s_{h'}^{(k)},a_{h'}^{(k)}\right)-r_{h'}^{(k)}-\int_{\mathcal{A}}\pi_\theta\left(a'\,\middle|s_{h'+1}^{(k)}\right)Q^\theta\left(s_{h'+1}^{(k)},a'\right)\mathrm{d}a'\right)\\
&+\frac{\lambda}{KH}\sum_{h=1}^{\infty}\gamma^{h-1}\left(\nabla_\theta^j\widehat{\nu}_h^\theta\right)^\top\widehat{\Sigma}^{-1}w^\theta\\
=&\left(\nabla_\theta^j\widehat{\nu}^\theta\right)^\top\widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)},a_h^{(k)}\right)\varepsilon_{h,k}^\theta+\frac{\lambda}{KH}\left(\nabla_\theta^j\widehat{\nu}^\theta\right)^\top\widehat{\Sigma}^{-1}w^\theta.
\end{aligned}
\tag{31}
$$

Combing the results of (30) and (31), we get for each $j \in [m]$,

$$\nabla_\theta^j v_\theta - \widehat{\nabla_\theta^j v_\theta} = \int_{\mathcal{S}\times\mathcal{A}} \xi(s)\pi_\theta(a|s)\left(\nabla_\theta^j Q^\theta - \widehat{\nabla_\theta^j Q^\theta} + \left(\nabla_\theta^j \log \Pi_\theta\right)\left(Q^\theta - \widehat{Q}^\theta\right)\right)(s,a)\mathrm{d}s\mathrm{d}a$$

$$= \left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\nabla_\theta^j \varepsilon_{h,k}^\theta + \frac{\lambda}{KH}\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1}\nabla_\theta^j w^\theta + \left(\nabla_\theta^j \nu^\theta\right)^\top \widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta$$
$$+ \frac{\lambda}{KH}\left(\nabla_\theta^j \widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} w^\theta$$
$$= \nabla_\theta^j\left(\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta + \frac{\lambda}{KH}\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} w^\theta\right)$$
$$= \nabla_\theta^j\Bigg(\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta + \frac{\lambda}{KH}\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} w^\theta$$
$$+ \left(\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} - \left(\nu^\theta\right)^\top \Sigma^{-1}\right)\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\Bigg).$$

Rewriting the above decomposition in a vector form, we get

$$\nabla_\theta v_\theta - \widehat{\nabla_\theta v_\theta} = \nabla_\theta\Bigg(\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta$$
$$+ \frac{\lambda}{KH}\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} w^\theta + \left(\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} - \left(\nu^\theta\right)^\top \Sigma^{-1}\right)\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\Bigg),$$

which is the desired result. □

D.7.2. PROOF OF LEMMA D.19

Proof. Note that,

$$\langle E_1, t\rangle = \left\langle \nabla_\theta\left(\left(\nu^\theta\right)^\top \Sigma^{-1}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right), t\right\rangle$$
$$= \left\langle \left(\nabla_\theta \nu^\theta\right)^\top \Sigma^{-1}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta, t\right\rangle + \left\langle \left(\nu^\theta\right)^\top \Sigma^{-1}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\nabla_\theta\varepsilon_{h,k}^\theta, t\right\rangle.$$

Let $e_{h,k} = \left\langle \nabla_\theta\left(\left(\nu^\theta\right)^\top \Sigma^{-1}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right), t\right\rangle$, we have

$$|e_{h,k}| \le \sqrt{C_1 dm}\|t\|\frac{1}{1-\gamma}\max_{j\in[m]}\sqrt{\left(\nabla_\theta^j \nu^\theta\right)^\top \Sigma^{-1}\nabla_\theta^j \nu^\theta} + 2G\sqrt{C_1 dm}\|t\|\frac{1}{(1-\gamma)^2}\sqrt{\left(\nu^\theta\right)^\top \Sigma^{-1}\nu^\theta} = B\sqrt{C_1 dm}\|t\|.$$

We have

$$\sum_{k=1}^K\sum_{h=1}^H \mathrm{Var}[e_{h,k}|\mathcal{F}_{h,k}] = \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}\left[\left\langle \nabla_\theta\left(\left(\nu^\theta\right)^\top \Sigma^{-1}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right), t\right\rangle^2 \middle| \mathcal{F}_{h,k}\right] = HKt^\top \Lambda_\theta t.$$

We pick $\sigma^2 = HKt^\top\Lambda_\theta t$, the Bernstein's inequality implies that for any $\varepsilon \in \mathbb{R}$,

$$\mathbb{P}\left(\left|\sum_{k=1}^K\sum_{h=1}^H e_{h,k}\right| \ge \varepsilon\right) \le 2\exp\left(-\frac{\varepsilon^2/2}{\sigma^2 + \sqrt{C_1 dm}\|t\|B\varepsilon/3}\right).$$

Therefore, if we pick $\varepsilon = \sigma\sqrt{2\log(2/\delta)} + 2\log(2/\delta)\sqrt{C_1 dm}\|t\|B/3$, we get

$$\mathbb{P}\left(\left|\sum_{k=1}^K\sum_{h=1}^H e_{h,k}\right| \ge \varepsilon\right) \le \delta,$$

i.e., we have with probability $1-\frac{\delta}{4}$,

$$\left|\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}e_{h,k}\right| \le \sqrt{\frac{2t^\top \Lambda_\theta t\log(8/\delta)}{HK}} + \frac{2\log(8/\delta)\sqrt{C_1 dm}\|t\|B}{3HK}$$

□

D.7.3. Proof of Lemma D.20

Proof. For an arbitrarily given θ_0, let $\Sigma_{\theta_0} = \mathbb{E}^{\pi_{\theta_0}}[\phi(s,a)\phi(s,a)^\top | s \sim \xi_{\theta_0}, a \sim \pi_{\theta_0}(\cdot|s)]$, we have

$$\begin{aligned}
&\left(\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} - \left(\nu^\theta\right)^\top \Sigma^{-1}\right)\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \\
&= \sum_{h'=1}^{\infty}\gamma^{h'-1}\left(\nu_1^\theta\right)^\top\left(\left(\widehat{M}_\theta\right)^{h'-1}\widehat{\Sigma}^{-1} - \left(M_\theta\right)^{h'-1}\Sigma^{-1}\right)\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \\
&= \sum_{h'=1}^{\infty}\gamma^{h'-1}\left(\Sigma_{\theta_0}^{-\frac{1}{2}}\nu_1^\theta\right)^\top\left(\left(\Sigma_{\theta_0}^{\frac{1}{2}}\widehat{M}_\theta\Sigma_{\theta_0}^{-\frac{1}{2}}\right)^{h'-1}\Sigma_{\theta_0}^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\widehat{\Sigma}^{-1}\Sigma^{\frac{1}{2}} - \left(\Sigma_{\theta_0}^{\frac{1}{2}}M_\theta\Sigma_{\theta_0}^{-\frac{1}{2}}\right)^{h'-1}\Sigma_{\theta_0}^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right) \\
&\quad\cdot\Sigma^{-\frac{1}{2}}\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta.
\end{aligned}$$

Taking derivatives on both sides, and let $\theta_0=\theta$, we get

$$\nabla_\theta^j E_2 = \nabla_\theta^j\left(\left(\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1} - \left(\nu^\theta\right)^\top \Sigma^{-1}\right)\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right) = E_{21}^j + E_{22}^j + E_{23}^j,$$

where

$$\begin{aligned}
E_{21}^j &= \sum_{h'=1}^{\infty}\gamma^{h'-1}\left(\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right)^\top\left(\left(\Sigma_\theta^{\frac{1}{2}}\widehat{M}_\theta\Sigma_\theta^{-\frac{1}{2}}\right)^{h'-1}\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\widehat{\Sigma}^{-1}\Sigma^{\frac{1}{2}} - \left(\Sigma_\theta^{\frac{1}{2}}M_\theta\Sigma_\theta^{-\frac{1}{2}}\right)^{h'-1}\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right) \\
&\quad\cdot\Sigma^{-\frac{1}{2}}\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\nabla_\theta^j\varepsilon_{h,k}^\theta \\
E_{22}^j &= \sum_{h'=1}^{\infty}\gamma^{h'-1}\left(\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right)^\top\left(\left(\Sigma_\theta^{\frac{1}{2}}\widehat{M}_\theta\Sigma_\theta^{-\frac{1}{2}}\right)^{h'-1}\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\widehat{\Sigma}^{-1}\Sigma^{\frac{1}{2}} - \left(\Sigma_\theta^{\frac{1}{2}}M_\theta\Sigma_\theta^{-\frac{1}{2}}\right)^{h'-1}\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right) \\
&\quad\cdot\Sigma^{-\frac{1}{2}}\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta \\
E_{23}^j &= \sum_{h'=1}^{\infty}\gamma^{h'-1}\left(\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right)^\top \\
&\quad\cdot\left(\nabla_\theta^j\left(\Sigma_{\theta_0}^{\frac{1}{2}}\widehat{M}_\theta\Sigma_{\theta_0}^{-\frac{1}{2}}\right)^{h'-1}\Bigg|_{\theta_0=\theta}\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\widehat{\Sigma}^{-1}\Sigma^{\frac{1}{2}} - \nabla_\theta^j\left(\Sigma_{\theta_0}^{\frac{1}{2}}M_\theta\Sigma_{\theta_0}^{-\frac{1}{2}}\right)^{h'-1}\Bigg|_{\theta_0=\theta}\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right) \\
&\quad\cdot\Sigma^{-\frac{1}{2}}\frac{1}{HK}\sum_{k=1}^{K}\sum_{h=1}^{H}\phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta.
\end{aligned}$$

Therefore, using the result of Lemma D.9, we get

$$
\begin{aligned}
|E_{21}^j| \leq & \sum_{h'=1}^{\infty} \gamma^{h'-1} \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left(\left(1 + \left\| \Sigma_\theta^{\frac{1}{2}} \left(\Delta M_\theta \right) \Sigma_\theta^{-\frac{1}{2}} \right\| \right)^{h'-1} \left(1 + \left\| \Sigma^{\frac{1}{2}} \left(\Delta \Sigma^{-1} \right) \Sigma^{\frac{1}{2}} \right\| \right) - 1 \right) \\
& \cdot \left\| \Sigma^{-\frac{1}{2}} \frac{1}{HK} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \nabla_\theta^j \varepsilon_{h,k}^\theta \right\| \\
= & \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left(\left(1 - \gamma - \gamma \left\| \Sigma_\theta^{\frac{1}{2}} \left(\Delta M_\theta \right) \Sigma_\theta^{-\frac{1}{2}} \right\| \right)^{-1} \left(1 + \left\| \Sigma^{\frac{1}{2}} \left(\Delta \Sigma^{-1} \right) \Sigma^{\frac{1}{2}} \right\| \right) - (1-\gamma)^{-1} \right) \\
& \cdot \left\| \Sigma^{-\frac{1}{2}} \frac{1}{HK} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \nabla_\theta^j \varepsilon_{h,k}^\theta \right\| \\
\leq & \frac{1}{1-\gamma} \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left(\left(1 - \frac{\left\| \Sigma_\theta^{\frac{1}{2}} \left(\Delta M_\theta \right) \Sigma_\theta^{-\frac{1}{2}} \right\|}{1-\gamma} \right)^{-1} \left(1 + \left\| \Sigma^{\frac{1}{2}} \left(\Delta \Sigma^{-1} \right) \Sigma^{\frac{1}{2}} \right\| \right) - 1 \right) \\
& \cdot \left\| \Sigma^{-\frac{1}{2}} \frac{1}{HK} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \nabla_\theta^j \varepsilon_{h,k}^\theta \right\| \\
|E_{22}^j| \leq & \sum_{h'=1}^{\infty} \gamma^{h'-1} \left\| \Sigma_\theta^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta \right\| \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left(\left(1 + \left\| \Sigma_\theta^{\frac{1}{2}} \left(\Delta M_\theta \right) \Sigma_\theta^{-\frac{1}{2}} \right\| \right)^{h'-1} \left(1 + \left\| \Sigma^{\frac{1}{2}} \left(\Delta \Sigma^{-1} \right) \Sigma^{\frac{1}{2}} \right\| \right) - 1 \right) \\
& \cdot \left\| \Sigma^{-\frac{1}{2}} \frac{1}{HK} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^\theta \right\| \\
\leq & \frac{1}{1-\gamma} \left\| \Sigma_\theta^{-\frac{1}{2}} \nabla_\theta^j \nu_1^\theta \right\| \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \left(\left(1 - \frac{\left\| \Sigma_\theta^{\frac{1}{2}} \left(\Delta M_\theta \right) \Sigma_\theta^{-\frac{1}{2}} \right\|}{1-\gamma} \right)^{-1} \left(1 + \left\| \Sigma^{\frac{1}{2}} \left(\Delta \Sigma^{-1} \right) \Sigma^{\frac{1}{2}} \right\| \right) - 1 \right) \\
& \cdot \left\| \Sigma^{-\frac{1}{2}} \frac{1}{HK} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^\theta \right\| \\
|E_{23}^j| \leq & \sum_{h'=2}^{\infty} (h'-1) \gamma^{h'-1} G \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \\
& \cdot \left(\left(1 + \left\| \Sigma_\theta^{\frac{1}{2}} \left(\Delta M_\theta \right) \Sigma_\theta^{-\frac{1}{2}} \right\| \right)^{h'-2} \left(1 + \left\| \Sigma_\theta^{\frac{1}{2}} \left(\frac{\nabla_\theta^j \left(\Delta M_\theta \right)}{G} \right) \Sigma_\theta^{-\frac{1}{2}} \right\| \right) \left(1 + \left\| \Sigma^{\frac{1}{2}} \left(\Delta \Sigma^{-1} \right) \Sigma^{\frac{1}{2}} \right\| \right) - 1 \right) \\
& \cdot \left\| \Sigma^{-\frac{1}{2}} \frac{1}{HK} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^\theta \right\| \\
\leq & \frac{G}{(1-\gamma)^2} \left\| \Sigma_\theta^{-\frac{1}{2}} \nu_1^\theta \right\| \left\| \Sigma_\theta^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \right\| \\
& \cdot \left(\left(1 - \frac{\left\| \Sigma_\theta^{\frac{1}{2}} \left(\Delta M_\theta \right) \Sigma_\theta^{-\frac{1}{2}} \right\|}{1-\gamma} \right)^{-2} \left(1 + \left\| \Sigma_\theta^{\frac{1}{2}} \left(\frac{\nabla_\theta^j \left(\Delta M_\theta \right)}{G} \right) \Sigma_\theta^{-\frac{1}{2}} \right\| \right) \left(1 + \left\| \Sigma^{\frac{1}{2}} \left(\Delta \Sigma^{-1} \right) \Sigma^{\frac{1}{2}} \right\| \right) - 1 \right) \\
& \cdot \left\| \Sigma^{-\frac{1}{2}} \frac{1}{HK} \sum_{k=1}^{K} \sum_{h=1}^{H} \phi\left(s_h^{(k)}, a_h^{(k)}\right) \varepsilon_{h,k}^\theta \right\|,
\end{aligned}
$$

where $\Delta\Sigma^{-1} = \widehat{\Sigma}^{-1} - \Sigma$ and we use the fact $\left\|\Sigma_\theta^{\frac{1}{2}} M_\theta \Sigma_\theta^{-\frac{1}{2}}\right\| \leq 1$ and $\left\|\Sigma_\theta^{\frac{1}{2}}\left(\nabla_\theta^j M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq G$ from Lemma D.10. Now, define $\alpha = 6\sqrt{\kappa_1}(4+\kappa_2+\kappa_3)\sqrt{\frac{C_1 d \log\frac{16dmH}{\delta}}{K}}$ and pick

$$K \geq 36\kappa_1(4+\kappa_2+\kappa_3)^2 \frac{C_1 d}{(1-\gamma)^2}\log\frac{16dmH}{\delta}, \quad \lambda \leq C_1 d\sigma_{\min}(\Sigma)\cdot\log\frac{8dmH}{\delta},$$

we get $\alpha \leq \frac{1-\gamma}{2}$. Using the results of Lemma D.14 and Lemma D.15, we get

$$\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\| \leq 4\sqrt{\frac{C_1 d\log\frac{8dmH}{\delta}}{K}} \leq \alpha \leq 1, \tag{32}$$

and

$$\left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq \alpha, \tag{33}$$

$$\left\|\Sigma_\theta^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_\theta\right)}{G}\right)\Sigma_\theta^{-\frac{1}{2}}\right\| \leq \alpha, \quad \forall j \in [m]. \tag{34}$$

Meanwhile, the event $\mathcal{E}_\varepsilon$ implies

$$\left\|\Sigma^{-\frac{1}{2}}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\varepsilon_{h,k}^\theta\right\| \leq \frac{4\sqrt{d}}{1-\gamma}\sqrt{\frac{\log\frac{32dmH}{\delta}}{KH}} \tag{35}$$

$$\left\|\Sigma^{-\frac{1}{2}}\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H \phi\left(s_h^{(k)}, a_h^{(k)}\right)\nabla_\theta^j\varepsilon_{h,k}^\theta\right\| \leq \frac{8\sqrt{d}G}{(1-\gamma)^2}\sqrt{\frac{\log\frac{32dmH}{\delta}}{KH}}. \tag{36}$$

Combining the results of (32), (33), (34), (35), (36) and use a union bound, we have with probability $1-3\delta$,

$$|E_{21}^j| \leq \frac{16\alpha\sqrt{d}G}{(1-\gamma)^4}\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{32dmH}{\delta}}{KH}}$$

$$|E_{22}^j| \leq \frac{8\alpha\sqrt{d}}{(1-\gamma)^3}\left\|\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{32dmH}{\delta}}{KH}}$$

$$|E_{23}^j| \leq \frac{16G\alpha\sqrt{d}}{(1-\gamma)^4}\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\sqrt{\frac{\log\frac{32dmH}{\delta}}{KH}}$$

where we use the fact $(1-\frac{\alpha}{1-\gamma})^{-1}(1+\alpha) \leq 1 + \frac{3\alpha}{1-\gamma}$ whenever $\alpha \leq \frac{1}{2(1-\gamma)}$. Summing up the above terms and using the definition of α, we get

$$|E_2^j| \leq \frac{240\sqrt{\kappa_1}(2+\kappa_2+\kappa_3)\sqrt{C_1}d}{(1-\gamma)^3}\left(\left\|\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\| + \frac{G}{1-\gamma}\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\right)\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\frac{\log\frac{32dmH}{\delta}}{KH}, \quad \forall j \in [m],$$

which finished the proof. □

D.7.4. Proof of Lemma D.21

Proof. Similar to the decomposition in the proof of Lemma D.20, we have

$$
\begin{aligned}
&\left|E_3^j\right| \\
=&\frac{\lambda}{KH}\left|\nabla_\theta^j\left(\left(\widehat{\nu}^\theta\right)^\top \widehat{\Sigma}^{-1}w^\theta\right)\right| \\
\leq&\frac{\lambda}{KH}\sum_{h=1}^{\infty}\gamma^{h-1}\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\left(1+\left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\|\right)^{h-1}\left(1+\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\|\right)\left\|\Sigma^{-1}\right\|\left\|\Sigma^{\frac{1}{2}}\nabla_\theta^j w^\theta\right\| \\
+&\frac{\lambda}{KH}\sum_{h=1}^{\infty}\gamma^{h-1}\left\|\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\left(1+\left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\|\right)^{h-1}\left(1+\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\|\right)\left\|\Sigma^{-1}\right\|\left\|\Sigma^{\frac{1}{2}}w^\theta\right\| \\
+&\frac{\lambda}{KH}\sum_{h=1}^{\infty}\gamma^{h-1}(h-1)G\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\| \\
&\cdot\left(1+\left\|\Sigma_\theta^{\frac{1}{2}}\left(\Delta M_\theta\right)\Sigma_\theta^{-\frac{1}{2}}\right\|\right)^{h-2}\left(1+\left\|\Sigma_\theta^{\frac{1}{2}}\left(\frac{\nabla_\theta^j\left(\Delta M_\theta\right)}{G}\right)\Sigma_\theta^{-\frac{1}{2}}\right\|\right)\left(1+\left\|\Sigma^{\frac{1}{2}}\left(\Delta\Sigma^{-1}\right)\Sigma^{\frac{1}{2}}\right\|\right)\left\|\Sigma^{-1}\right\|\left\|\Sigma^{\frac{1}{2}}w^\theta\right\| \\
\leq&\frac{\lambda}{KH}\sum_{h=1}^{\infty}\gamma^{h-1}\left(1+\alpha\right)^h\left\|\Sigma^{-1}\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma^{\frac{1}{2}}\nabla_\theta^j w_h^\theta\right\|+\left\|\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\left\|\Sigma^{\frac{1}{2}}w^\theta\right\|+G(h-1)\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma^{\frac{1}{2}}w^\theta\right\|\right) \\
\leq&\frac{2\lambda}{KH(1-\gamma)}\left\|\Sigma^{-1}\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma^{\frac{1}{2}}\nabla_\theta^j w^\theta\right\|+\left\|\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\left\|\Sigma^{\frac{1}{2}}w^\theta\right\|+\frac{2G}{1-\gamma}\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\left\|\Sigma^{\frac{1}{2}}w^\theta\right\|\right),
\end{aligned}
$$

where α is defined in the same way as that in the proof of Lemma D.20. Similarly, we have $\alpha \leq \frac{1-\gamma}{2}$ and we have

$$
\begin{aligned}
\left\|\Sigma^{\frac{1}{2}}\nabla_\theta^j w^\theta\right\|^2 &= \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H}\left(\nabla_\theta^j Q^\theta\left(s_h^{(1)},a_h^{(1)}\right)\right)^2\right] \leq \frac{G^2}{(1-\gamma)^4} \\
\left\|\Sigma^{\frac{1}{2}}w^\theta\right\|^2 &= \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H}\left(Q^\theta\left(s_h^{(1)},a_h^{(1)}\right)\right)^2\right] \leq \frac{1}{(1-\gamma)^2}.
\end{aligned}
$$

We conclude

$$
\begin{aligned}
|E_3^j| \leq&\frac{2\lambda}{KH(1-\gamma)^2}\left\|\Sigma^{-1}\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\frac{3G}{1-\gamma}+\left\|\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\right) \\
\leq&\frac{6C_1 d\sigma_{\min}(\Sigma)\cdot\log\frac{8dmH}{\delta}}{KH(1-\gamma)^2}\left\|\Sigma^{-1}\right\|\left\|\Sigma_\theta^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}\right\|\left(\left\|\Sigma_\theta^{-\frac{1}{2}}\nu_1^\theta\right\|\frac{G}{1-\gamma}+\left\|\Sigma_\theta^{-\frac{1}{2}}\nabla_\theta^j\nu_1^\theta\right\|\right).
\end{aligned}
$$

which has finished the proof. ∎