NORTHWESTERN UNIVERSITY

On the Optimality and Complexity of Reinforcement Learning

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Industrial Engineering and Management Sciences

By

Zuyue Fu

EVANSTON, ILLINOIS

August 2022

ABSTRACT

On the Optimality and Complexity of Reinforcement Learning

Zuyue Fu

In this dissertation, we aim to develop algorithms that achieve optimality with provable complexity guarantees under various settings in reinforcement learning (RL). Specifically, in Markov decision processes (MDPs), we study single-agent and multi-agent online RL, respectively, and offline RL under the presence of unobserved confounders.

- Single-agent online RL. We design a single-timescale actor-critic method to solve single-agent RL, where the actor and critic are updated simultaneously. Specifically, in each iteration, the critic update is obtained by applying the Bellman evaluation operator only once while the actor is updated in the policy gradient direction computed using the critic. We prove that the actor sequence converges to a globally optimal policy at a sublinear $O(K^{-1/2})$ rate, where K is the number of iterations.
- Multi-agent online RL. We study discrete-time mean-field Markov games with infinite numbers of agents, where each agent aims to minimize its ergodic cost. Specifically, we consider a linear-quadratic case, where the agents have identical

linear state transitions and quadratic cost functions. For such a game, we provide sufficient conditions for the existence and uniqueness of its Nash equilibrium, and also propose a model-free mean-field actor-critic algorithm. In particular, we prove that our algorithm converges to the Nash equilibrium at a linear rate.

- Offline RL with unobserved confounders. We study offline RL in the face of unobserved confounders. Offline RL is typically facing the following two significant challenges: (i) the agent may be confounded by the unobserved state variables; (ii) the offline data collected a prior does not provide sufficient coverage for the environment. To tackle the above challenges, we study the policy learning in the confounded MDPs with the aid of instrumental variables (IVs). Specifically, we propose value- and ratio-based identification results for the identification of the expected total reward. Then by leveraging pessimism and our identification results, we propose various policy learning methods with the finite-sample suboptimality guarantee of finding the optimal in-class policy under minimal data coverage and modeling assumptions.

Acknowledgements

It is only with the unwavering support of many people that I can achieve this milestone in my life, and I would like to express my sincerest gratitude to them.

First and foremost, I would like to thank my advisor Professor Zhaoran Wang. It is a great honor for me to be advised by and work alongside one of the great minds in the theory of machine learning. Thank you for being a great advisor who has patiently guided me, once an outsider to the field of machine learning, into such a beautiful world. Thank you for being a great mentor and offering me so much advice, guidance, and opportunities for my academic and career development. Thank you for being a great friend, who is so caring, understanding, supporting, encouraging, and always willing to help. Your knowledge, sharpness, dedication, and passion, inspire me forward in my academic journey. You are great role models to whom I shall always look up.

I would like to thank many faculties I have had throughout my Ph.D. studies, all of whom I have learned a great deal from. Thank you, Professor Ermin Wei and Professor Zhengling Qi, for your dedicated guidance and serving on my dissertation committee. Thank you, Professor Zhuoran Yang, for teaching me so much and for your help with many challenging problems I have encountered. Thank you, Professor Yongxin Chen and Professor Yang Liu, for the valuable opportunities to collaborate with you and leading me to various fields in machine learning. I have also had the privilege of collaborating with many other exceptional researchers: Lingxiao, Yufeng, Hongyi, Zhihan, Zhihong, and Luofeng. Thank you all for your efforts in our research projects.

I want to thank the IEMS faculty for offering me admission to the Ph.D. program several years ago, teaching me so much about industrial engineering, and being role models as exceptional researchers. I also want to thank the IEMS staff members for always being so warm and helpful.

The past five years have been incredibly happy, for which I have my dear friends to thank. Lingxiao and Boyi, when I began my Ph.D. studies five years ago, I could never have imagined that I could make such dear friends like you, with whom I would share so many memories (food and car, primarily) and from whom I would learn so much. Bing, thanks for being my roommate and helping me so much in life during my earliest year in the country. We will for sure watch more soccer games in the future. Alex, you are so interesting and helpful. I really enjoy the days that we do homework together. Shukai, thanks for tirelessly organizing various activities during the past three years. Oliver, you are always so dedicated and passionate about everything. It is always a pleasure talking with you. Michelle and Xinyi, thanks for all the help you have given me (including the fancy food). Xiaochun, thank you for being in my life. It is the greatest joy to always have you by my side. There are so many friends not only in the IEMS family but also throughout the university that I have drank and ate and played with. I could not list all the names here, but the memories with these friends will always be with me.

Last but not least, I want to thank all my family members. Thank you for bringing me to the wonderful world. Thank you for your unconditional love, trust and support.

Without your extraordinary efforts on my education since my childhood, this dissertation would be impossible to accomplish. I would like to dedicate this work to you.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

In reinforcement learning (RL, Sutton and Barto (2018)), the agent aims to make sequential decisions that maximize the expected total reward through interacting with the environment and learning from the experiences, where the environment is modeled as a Markov decision process (MDP, Puterman (2014)). RL with deep neural networks achieves tremendous successes in practice, e.g., video games (Silver et al., 2016; OpenAI, 2018), robotics (Kalashnikov et al., 2018), solving social dilemmas (de Cote et al., 2006; Leibo et al., 2017; Hughes et al., 2018), etc. However, there still lack a theoretical understanding on the optimality, i.e., how good is a learned policy compared with the optimal policy, and complexity, i.e, how large a dataset is required to learn a good policy, of various types of RL methods. In this dissertation, we aim to develop and study algorithms that provably achieve optimality and complexity guarantees under various settings in RL.

In single-agent online RL, where the agent can actively interact with the environment to collect new data, to achieve the highest possible total reward in expectation, the actor-critic method (Konda and Tsitsiklis, 2000) is the among the most commonly used algorithms. Specifically, to establish convergence guarantees for actor-critic, most existing works either focus on the bi-level setting or the two-timescale setting, which are seldom adopted in practice. In Chapter 2, we propose a single-timescale actor-critic method, and investigate its convergence and global optimality under linear and deep neural network function approximation. In particular, we focus on the family of energy-based policies and aim to find the optimal policy within this class. In our actor-critic algorithm, the actor update follows proximal policy optimization (PPO, Schulman et al. (2017)) and the critic update is obtained by applying the Bellman evaluation operator only once to the current critic iterate, which is more closed to the practical algorithms.

Multi-agent RL extends single-agent RL to sequential decision-making problems involving multiple agents. Mean-field game (MFG, Huang et al. (2003, 2006); Lasry and Lions (2006a,b, 2007)) is a specific form of multi-agent RL, which utilizes mean-field approximation to model the strategic interactions within a large population. In Chapter 3, we develop an efficient model-free RL approach to solve MFGs, which provably attains the Nash equilibrium. In particular, we focus on discrete-time MFGs with linear state transitions and quadratic cost functions, where the aggregated effect of the population is quantified by the mean-field state. In detail, we propose a mean-field actor-critic algorithm, which alternatingly updates the policy and mean-field state. In theory, we prove that the sequence of policies and its corresponding sequence of mean-field states converge to the unique Nash equilibrium at a linear rate.

Since actively interacting with the environment in an MDP is usually either expansive or unethical (e.g., in healthcare (Raghu et al., 2017; Komorowski et al., 2018; Gottesman et al., 2019), autonomous driving (Shalev-Shwartz et al., 2016)), a growing body of literature focus on designing RL methods in the offline setting, where the agent aims to learn an optimal policy π^* in the infinite-horizon Markov decision process (MDP, Puterman (2014)) only through observational data. In Chapter 4, we aim to solve the following two challenges in offline RL: (i) The agent may be confounded by unobserved variables (confounders) in the observational data; (ii) Due to insufficient data coverage, typical methods (Precup, 2000; Antos et al., 2008b; Chen and Jiang, 2019) may fail to converge without imposing strong assumptions on the data generation process. To tackle such challenges, by leveraging instrumental variables (IVs) and the principle of pessimism, we propose value-based and ratio-based estimators of the optimal policy, which are shown enjoy provable optimality and complexity guarantees.

1.1. Single-Timescale Actor-Critic Provably Finds Globally Optimal Policy

In reinforcement learning (RL, Sutton et al. (1998)), the agent aims to make sequential decisions that maximize the expected total reward through interacting with the environment and learning from the experiences, where the environment is modeled as a Markov Decision Process (MDP, (Puterman, 2014)). To learn a policy that achieves the highest possible total reward in expectation, the actor-critic method (Konda and Tsitsiklis, 2000) is among the most commonly used algorithms. In actor-critic, the actor refers to the policy and the critic corresponds to the value function that characterizes the performance of the actor. This method directly optimizes the expected total return over the policy class by iteratively improving the actor, where the update direction is determined by the critic. In particular, recently, actor-critic combined with deep neural networks (LeCun et al., 2015) achieves tremendous empirical successes in solving large-scale RL tasks, such as the game of Go (Silver et al., 2017), StarCraft (Vinyals et al., 2019), Dota (OpenAI, 2018), Rubik's cube (Agostinelli et al., 2019; Akkaya et al., 2019), and autonomous driving (Sallab et al., 2017). See Li (2017) for a detailed survey of the recent developments of deep reinforcement learning.

Despite these great empirical successes of actor-critic, there is still an evident chasm between theory and practice. Specifically, to establish convergence guarantees for actor-critic, most existing works either focus on the bi-level setting or the two-timescale setting, which are seldom adopted in practice. In particular, under the bi-level setting (Yang et al., 2019a; Wang et al., 2019; Agarwal et al., 2019; Fu et al., 2019; Liu et al., 2019; Abbasi-Yadkori et al., 2019a,b; Cai et al., 2019; Hao et al., 2020; Mei et al., 2020; Bhandari and Russo, 2020), the actor is updated only after the critic solves the policy evaluation sub-problem completely, which is equivalent to applying the Bellman evaluation operator to the previous critic for infinite times. Consequently, actor-critic under the bi-level setting is a double-loop iterative algorithm where the inner loop is allocated for solving the policy evaluation sub-problem of the critic. In terms of theoretical analysis, such a double-loop structure decouples the analysis for the actor and critic. For the actor, the problem is essentially reduced to analyzing the convergence of a variant of the policy gradient method (Sutton et al., 2000; Kakade, 2002) where the error of the gradient estimate depends on the policy evaluation error of the critic. Besides, under the two-timescale setting (Borkar and Konda, 1997; Konda and Tsitsiklis, 2000; Xu et al., 2020; Wu et al., 2020; Hong et al., 2020), the actor and the critic are updated simultaneously, but with disparate stepsizes. More concretely, the stepsize of the actor is set to be much smaller than that of the critic, with the ratio between these stepsizes converging to zero. In an asymptotic sense, such a separation between stepsizes ensures that the critic completely solves its policy evaluation sub-problem asymptotically. In other words, such a two-timescale scheme results in a separation between actor and critic in an asymptotic sense, which leads to asymptotically unbiased policy gradient estimates. In sum, in terms of convergence analysis, the existing theory of actor-critic hinges on decoupling the analysis for critic and actor, which is ensured via focusing on the bi-level or two-timescale settings.

However, most practical implementations of actor-critic are under the single-timescale setting (Peters and Schaal, 2008a; Schulman et al., 2015; Mnih et al., 2016; Schulman et al., 2017; Haarnoja et al., 2018), where the actor and critic are simultaneously updated, and particularly, the actor is updated without the critic reaching an approximate solution to the policy evaluation sub-problem. Meanwhile, in comparison with the two-timescale setting, the actor is equipped with a much larger stepsize in the the single-timescale setting such that the asymptotic separation between the analysis of actor and critic is no longer valid.

Furthermore, when it comes to function approximation, most existing works only analyze the convergence of actor-critic with either linear function approximation (Xu et al., 2020; Wu et al., 2020; Hong et al., 2020), or shallow-neural-network parameterization (Wang et al., 2019; Liu et al., 2019). In contrast, practically used actor-critic methods such as asynchronous advantage actor-critic (Mnih et al., 2016) and soft actor-critic (Haarnoja et al., 2018) oftentimes represent both the actor and critic using deep neural networks.

Thus, the following question is left open:

Does single-timescale actor-critic provably find a globally optimal policy under the function approximation setting, especially when deep neural networks are employed?

To answer such a question, we make the first attempt to investigate the convergence and global optimality of single-timescale actor-critic with linear and neural network function approximation. In particular, we focus on the family of energy-based policies and aim to find the optimal policy within this class. Here we represent both the energy function and the critic as linear or deep neural network functions. In our actor-critic algorithm, the actor update follows proximal policy optimization (PPO) (Schulman et al., 2017) and the critic update is obtained by applying the Bellman evaluation operator only once to the current critic iterate. As a result, the actor is updated before the critic solves the policy evaluation sub-problem. Such a coupled updating structure persists even when the number of iterations goes to infinity, which implies that the update direction of the actor is always biased compared with the policy gradient direction. This brings an additional challenge that is absent in the bi-level and the two-timescale settings, where the actor and critic are decoupled asymptotically.

To tackle such a challenge, our analysis captures the joint effect of actor and critic updates on the objective function, dubbed as the "double contraction" phenomenon, which plays a pivotal role for the success of single-timescale actor-critic. Specifically, thanks to the discount factor of the MDP, the Bellman evaluation operator is contractive, which implies that, after each update, the critic makes noticeable progress by moving towards the value function associated with the current actor. As a result, although we use a biased estimate of the policy gradient, thanks to the contraction brought by the discount factor, the accumulative effect of the biases is controlled. Such a phenomenon enables us to characterize the progress of each iteration of joint actor and critic update, and thus yields the convergence to the globally optimal policy. In particular, for both the linear and neural settings, we prove that, single-timescale actor-critic finds a $O(K^{-1/2})$-globally optimal policy after K iterations. To the best of our knowledge, we seem to establish the first theoretical guarantee of global convergence and global optimality for actor-critic with function approximation in the single-timescale setting. Moreover, under the broader scope of policy optimization with nonlinear function approximation, our work seems to prove convergence and optimality guarantees for actor-critic with deep neural network for the first time.

Contribution. Our contribution is two-fold. First, in the single-timescale setting with linear function approximation, we prove that, after K iterations of actor and critic updates, actor-critic returns a policy that is at most $O(K^{-1/2})$ inferior to the globally optimal policy. Second, when both the actor and critic are represented by deep neural networks, we prove a similar $O(K^{-1/2})$ rate of convergence to the globally optimal policy when the architecture of the neural networks are properly chosen.

Related Work. Our work extends the line of works on the convergence of actor-critic under the function approximation setting. In particular, actor-critic is first introduced in Sutton et al. (2000); Konda and Tsitsiklis (2000). Later, Kakade (2002); Peters and Schaal (2008b) propose the natural actor-critic method which updates the policy via the natural gradient (Amari, 1998) direction. The convergence of (natural) actor-critic with linear function approximation are studied in Bhatnagar et al. (2008, 2009); Bhatnagar (2010); Castro and Meir (2010); Maei (2018). However, these works only characterize the asymptotic convergence of actor-critic and their proofs all resort to tools from stochastic approximation via ordinary differential equations (Borkar, 2008). As a result, these works only show that actor-critic with linear function approximation converges to the set of stable equilibria of a set of ordinary differential equations. Recently, Zhang et al. (2019a) propose a variant of actor-critic where Monte-Carlo sampling is used to ensure the critic and the policy gradient estimates are unbiased. Although they incorporate nonlinear function approximation in the actor, they only establish finite-time convergence result to a stationary point of the expected total reward. Moreover, due to having an inner loop for solving the policy evaluation sub-problem, they focus on the bi-level setting. Moreover, under the two-timescale setting, Wu et al. (2020); Xu et al. (2020) show that actor-critic with linear function approximation finds an ε-stationary point with $\widetilde{O}(\varepsilon^{-5/2})$ samples, where ε measures the squared norm of the policy gradient. All of these results establish the convergence of actor-critic, without characterizing the optimality of the policy obtained by actor-critic.

In terms of the global optimality of actor-critic, Fazel et al. (2018); Malik et al. (2018); Tu and Recht (2018); Yang et al. (2019a); Bu et al. (2019); Fu et al. (2019) show that policy gradient and bi-level actor-critic methods converge to the globally optimal policies under the linear-quadratic setting, where the state transitions follow a linear dynamical system and the reward function is quadratic. For general MDPs, Bhandari and Russo (2019) recently prove the global optimality of vanilla policy gradient under the assumption that the families of policies and value functions are both convex. In addition, our work is also related to Liu et al. (2019) and Wang et al. (2019), where they establish the global optimality of proximal policy optimization and (natural) actor-critic, respectively, where both the actor and critic are parameterized by two-layer neural networks. Our work is also related to Agarwal et al. (2019); Abbasi-Yadkori et al. (2019a,b); Cai et al. (2019); Hao et al. (2020); Mei et al. (2020); Bhandari and Russo (2020), which focus on characterizing the optimality of natural policy gradient in tabular and/or linear settings. However, these aforementioned works all focus on bi-level actor-critic, where the actor is updated only after the critic solves the policy evaluation sub-problem to an approximate optimum.

Besides, these works consider linear or two-layer neural network function approximations whereas we focus on the setting with deep neural networks. Furthermore, under the two-timescale setting, Xu et al. (2020); Hong et al. (2020) prove that linear actor-critic requires a sample complexity of $\widetilde{O}(\varepsilon^{-4})$ for obtaining an ε-globally optimal policy. In comparison, our $O(K^{-1/2})$ convergence for single-timescale actor-critic can be translated into a similar $\widetilde{O}(\varepsilon^{-4})$ sample complexity directly. Moreover, when reusing the data, our result leads to an improved $\widetilde{O}(\varepsilon^{-2})$ sample complexity. In addition, our work is also related to Geist et al. (2019), which proposes a variant of policy iteration algorithm with Bregman divergence regularization. Without considering an explicit form of function approximation, their algorithm is shown to converge to the globally optimal policy at a similar $O(K^{-1/2})$ rate, where K is the number of policy updates. In contrast, our method is single-timescale actor-critic with linear or deep neural network function approximation, which enjoys both global convergence and global optimality. Meanwhile, our proof is based on a finite-sample analysis, which involves dealing with the algorithmic errors that track the performance of actor and critic updates as well as the statistical error due to having finite data.

Our work is also related to the literature on deep neural networks. Previous works (Daniely, 2017; Jacot et al., 2018; Wu et al., 2018; Allen-Zhu et al., 2018a,b; Du et al., 2018; Zou et al., 2018; Chizat and Bach, 2018; Jacot et al., 2018; Li and Liang, 2018; Cao and Gu, 2019a,b; Arora et al., 2019; Lee et al., 2019; Gao et al., 2019) analyze the computational and statistical rates of supervised learning methods with overparameterized neural networks. In contrast, our work employs overparameterized deep neural networks in actor-critic for solving RL tasks, which is significantly more challenging than supervised learning due to the interplay between the actor and the critic.

1.2. Actor-Critic Provably Finds Nash Equilibria of Linear-Quadratic Mean-Field Games

In reinforcement learning (RL) (Sutton and Barto, 2018), an agent learns to make decisions that minimize its expected total cost through sequential interactions with the environment. Multi-agent reinforcement learning (MARL) (Shoham et al., 2003, 2007; Busoniu et al., 2008) aims to extend RL to sequential decision-making problems involving multiple agents. In a non-cooperative game, we are interested in the Nash equilibrium (Nash, 1951), which is a joint policy of all the agents such that each agent cannot decrease its expected total cost by unilaterally deviating from its Nash policy. The Nash equilibrium plays a critical role in understanding the social dynamics of self-interested agents (Ash, 2000; Axtell, 2002) and constructing the optimal policy of a particular agent via fictitious self-play (Bowling and Veloso, 2000; Ganzfried and Sandholm, 2009). With the recent development in deep learning (LeCun et al., 2015), MARL with function approximation achieves tremendous empirical successes in applications, including Go (Silver et al., 2016, 2017), Poker (Heinrich and Silver, 2016; Moravčík et al., 2017), Star Craft (Vinyals et al., 2019), Dota (OpenAI, 2018), autonomous driving (Shalev-Shwartz et al., 2016), multi-robotic systems (Yang and Gu, 2004), and solving social dilemmas (de Cote et al., 2006; Leibo et al., 2017; Hughes et al., 2018). However, since the capacity of the joint state and action spaces grows exponentially in the number of agents, such MARL approaches become computationally intractable when the number of agents is large, which is common in real-world applications (Sandholm, 2010; Calderone, 2017; Wang et al., 2017a).

Mean-field game is proposed by Huang et al. (2003, 2006); Lasry and Lions (2006a,b, 2007) with the idea of utilizing mean-field approximation to model the strategic interactions within a large population. In a mean-field game, each agent has the same cost function and state transition, which depend on the other agents only through their aggregated effect. As a result, the optimal policy of each agent depends solely on its own state and the aggregated effect of the population, and such an optimal policy is symmetric across all the agents. Moreover, if the aggregated effect of the population corresponds to the Nash equilibrium, then the optimal policy of each agent jointly constitutes a Nash equilibrium. Although such a Nash equilibrium corresponds to an infinite number of agents, it well approximates the Nash equilibrium for a sufficiently large number of agents (Bensoussan et al., 2016). Also, as the aggregated effect of the population abstracts away the strategic interactions between individual agents, it circumvents the computational intractability of the MARL approaches that do not exploit symmetry.

However, most existing work on mean-field games focuses on characterizing the existence and uniqueness of the Nash equilibrium rather than designing provably efficient algorithms. In particular, most existing work considers the continuous-time setting, which requires solving a pair of Hamilton-Jacobi-Bellman (HJB) and Fokker-Planck (FP) equations, whereas the discrete-time setting is more common in practice, e.g., in the aforementioned applications. Moreover, most existing approaches, including the ones based on solving the HJB and FP equations, require knowing the model of dynamics (Bardi and Priuli, 2014), or having the access to a simulator, which generates the next state given any state-action pair and aggregated effect of the population (Guo et al., 2019), which is often unavailable in practice.

To address these challenges, we develop an efficient model-free RL approach to mean-field game, which provably attains the Nash equilibrium. In particular, we focus on discrete-time mean-field games with linear state transitions and quadratic cost functions, where the aggregated effect of the population is quantified by the mean-field state. Such games capture the fundamental difficulties of general mean-field games and well approximates a variety of real-world systems such as power grids (Minciardi and Sacile, 2011), swarm robots (Fang, 2014; Araki et al., 2017; Doerr et al., 2018), and financial systems (Zhou and Li, 2000; Huang and Li, 2018). In detail, based on the Nash certainty equivalence (NCE) principle (Huang et al., 2006, 2007), we propose a mean-field actor-critic algorithm which, at each iteration, given the mean-field state μ, approximately attains the optimal policy π_μ^* of each agent, and then updates the mean-field state μ assuming that all the agents follow π_μ^*. We parametrize the actor and critic by linear and quadratic functions, respectively, and prove that such a parameterization encompasses the optimal policy of each agent. Specifically, we update the actor parameter using policy gradient (Sutton et al., 2000) and natural policy gradient (Kakade, 2002; Peters and Schaal, 2008a; Bhatnagar et al., 2009) and update the critic parameter using primal-dual gradient temporal difference (Sutton et al., 2009a,b). In particular, we prove that given the mean-field state μ, the sequence of policies generated by the actor converges linearly to the optimal policy π_μ^*. Moreover, when alternatingly update the policy and mean-field state, we prove that the sequence of policies and its corresponding sequence of mean-field states converge to the unique Nash equilibrium at a linear rate. Our approach can be interpreted from both “passive” and “active” perspectives: (i) Assuming that each self-interested agent employs the single-agent actor-critic algorithm, the policy of each agent converges to the unique Nash policy, which characterizes the social dynamics of a large population of model-free RL agents. (ii) For a particular agent, our approach serves as a fictitious self-play method for it to find its Nash policy, assuming the other agents give their best responses. To the best of our knowledge, our work establishes the first efficient model-free RL approach with function approximation that provably attains the Nash equilibrium of a discrete-time mean-field game. As a byproduct, we also show that the sequence of policies generated by the single-agent actor-critic algorithm converges at a linear rate to the optimal policy of a linear-quadratic regulator (LQR) problem in the presence of drift, which may be of independent interest.

Related Work. Mean-field game is first introduced in Huang et al. (2003, 2006); Lasry and Lions (2006a,b, 2007). In the last decade, there is growing interest in understanding continuous-time mean-field games. See, e.g., Guéant et al. (2011); Bensoussan et al. (2013); Gomes et al. (2014); Carmona and Delarue (2013, 2018) and the references therein. Due to their simple structures, continuous-time linear-quadratic mean-field games are extensively studied under various model assumptions. See Li and Zhang (2008); Bardi (2011); Wang and Zhang (2012); Bardi and Priuli (2014); Huang et al. (2016a,b); Bensoussan et al. (2016, 2017); Caines and Kizilkale (2017); Huang and Huang (2017); Moon and Başar (2018); Huang and Zhou (2019) for examples of this line of work. Meanwhile, the literature on discrete-time linear-quadratic mean-field games remains relatively scarce. Most of this line of work focuses on characterizing the existence of a Nash equilibrium and the behavior of such a Nash equilibrium when the number of agents goes to infinity

(Gomes et al., 2010; Tembine and Huang, 2011; Moon and Başar, 2014; Biswas, 2015; Saldi et al., 2018a,b, 2019). See also Yang et al. (2018a), which applies maximum entropy inverse RL (Ziebart et al., 2008) to infer the cost function and social dynamics of discrete-time mean-field games with finite state and action spaces. Our work is most related to Guo et al. (2019), where they propose a mean-field Q-learning algorithm (Watkins and Dayan, 1992) for discrete-time mean-field games with finite state and action spaces. Such an algorithm requires the access to a simulator, which, given any state-action pair and mean-field state, outputs the next state. In contrast, both our state and action spaces are infinite, and we do not require such a simulator but only observations of trajectories under given mean-field state. Correspondingly, we study the mean-field actor-critic algorithm with linear function approximation, whereas their algorithm is tailored to the tabular setting. Also, our work is closely related to Mguni et al. (2018), which focuses on a more restrictive setting where the state transition does not involve the mean-field state. In such a setting, mean-field games are potential games, which is, however, not true in more general settings (Li et al., 2017; Briani and Cardaliaguet, 2018). In comparison, we allow the state transition to depend on the mean-field state. Meanwhile, they propose a fictitious self-play method based on the single-agent actor-critic algorithm and establishes its asymptotic convergence. However, their proof of convergence relies on the assumption that the single-agent actor-critic algorithm converges to the optimal policy, which is unverified therein. In addition, our work is related to Jayakumar and Aditya (2019), where the proposed algorithm is only shown to converge asymptotically to a stationary point of the mean-field game.

Our work also extends the line of work on finding the Nash equilibria of Markov games using MARL. Due to the computational intractability introduced by the large number of agents, such a line of work focuses on finite-agent Markov games (Littman, 1994, 2001; Hu and Wellman, 1998; Bowling, 2001; Lagoudakis and Parr, 2002; Hu and Wellman, 2003; Conitzer and Sandholm, 2007; Perolat et al., 2015; Pérolat et al., 2016b,a, 2018; Wei et al., 2017; Zhang et al., 2018; Zou et al., 2019; Casgrain et al., 2019). See also Shoham et al. (2003, 2007); Busoniu et al. (2008); Li (2018) for detailed surveys. Our work is related to Yang et al. (2018b), where they combine the mean-field approximation of actions (rather than states) and Nash Q-learning (Hu and Wellman, 2003) to study general-sum Markov games with a large number of agents. However, the Nash Q-learning algorithm is only applicable to finite state and action spaces, and its convergence is established under rather strong assumptions. Also, when the number of agents goes to infinity, their approach yields a variant of tabular Q-learning, which is different from our mean-field actor-critic algorithm.

For policy optimization, based on the policy gradient theorem, Sutton et al. (2000); Konda and Tsitsiklis (2000) propose the actor-critic algorithm, which is later generalized to the natural actor-critic algorithm (Peters and Schaal, 2008a; Bhatnagar et al., 2009). Most existing results on the convergence of actor-critic algorithms are based on stochastic approximation using ordinary differential equations (Bhatnagar et al., 2009; Castro and Meir, 2010; Konda and Tsitsiklis, 2000; Maei, 2018), which are asymptotic in nature. For policy evaluation, the convergence of primal-dual gradient temporal difference is studied in Liu et al. (2015); Du et al. (2017); Wang et al. (2017b); Yu (2017); Wai et al. (2018). However, this line of work assumes that the feature mapping is bounded, which is not the case in our setting. Thus, the existing convergence results are not applicable to analyzing the critic update in our setting. To handle the unbounded feature mapping, we utilize a truncation argument, which requires more delicate analysis.

Finally, our work extends the line of work that studies model-free RL for LQR. For example, Bradtke (1993); Bradtke et al. (1994) show that policy iteration converges to the optimal policy, Tu and Recht (2017); Dean et al. (2017) study the sample complexity of least-squares temporal-difference for policy evaluation. More recently, Fazel et al. (2018); Malik et al. (2018); Tu and Recht (2018) show that the policy gradient algorithm converges at a linear rate to the optimal policy. See as also Hardt et al. (2016); Dean et al. (2018) for more in this line of work. Our work is also closely related to Yang et al. (2019b), where they show that the sequence of policies generated by the natural actor-critic algorithm enjoys a linear rate of convergence to the optimal policy. Compared with this work, when fixing the mean-field state, we use the actor-critic algorithm to study LQR in the presence of drift, which introduces significant difficulties in the analysis. As we show in §3.2, the drift causes the optimal policy to have an additional intercept, which makes the state- and action-value functions more complicated.

1.3. Offline Reinforcement Learning with Instrumental Variables in Confounded Markov Decision Processes

Reinforcement learning (RL, Sutton and Barto (2018)) with deep neural networks gains tremendous successes in practice, e.g., games (Silver et al., 2016; OpenAI, 2018), robotics (Kalashnikov et al., 2018), etc. Most RL methods heavily rely on an efficient data generator, e.g., game engines (Bellemare et al., 2013) and physics simulators (Todorov et al., 2012), which serve as an environment to be interacted with the agent. Since interacting with such an environment is usually either expansive or unethical (e.g., in healthcare (Raghu et al., 2017; Komorowski et al., 2018; Gottesman et al., 2019), autonomous driving (Shalev-Shwartz et al., 2016)), RL methods requiring actively collecting data in an online fashion is impractical under this scenario. Thus, a growing body of literature focus on designing RL methods in the offline setting, where the agent aims to learn an optimal policy π^* in the infinite-horizon Markov decision process (MDP, Puterman (2014)) only through observational data, which consists of N trajectories generated by a behavior policy b with a finite horizon T.

However, applying RL methods in the offline setting still possesses the following challenges: (i) The agent may be confounded by unobserved variables (confounders) in the observational data. We refer to the MDP with such confounders as confounded MDP. Such confounders usually comes from private data or heuristic information not recorded (Brookhart et al., 2010). In the confounded MDP, the causal effects of actions on the transitions and rewards are not identifiable from the observational data, leaving most offline RL methods assuming unconfoundedness not applicable in our setting. (ii) To learn an optimal policy from observational data, many prior methods (Precup, 2000; Antos et al., 2008b; Chen and Jiang, 2019) require a data coverage assumption for any policy π, i.e., the density ratio between the state-action visitation measure induced by π and that induced by the behavior policy b is uniformly upper bounded for any π. However, such a data coverage assumption is hard to satisfy in practice, especially when the state or action spaces are large. Further, many existing methods developed under this assumption are not stable or even do not converge when the assumption is violated (Wang et al., 2020, 2021c).

To tackle the above challenge (i), we study the confounded MDP via instrumental variables (IVs, Pearl (2009)). Informally, IVs are variables that affect the transitions and rewards only through actions. With the aid of IVs, we introduce two types of identification results: value function (VF)-based and marginalized importance sampling (MIS)-based. Specifically, with only finite-horizon data, VF-based identification result helps identify the state-value function in the infinite-horizon confounded MDP and establish a new Bellman equation by leveraging IVs, which relies on the memoryless assumption on the unmeasured confounders. The memoryless assumption rules out the existence of unmeasured confounders that can affect future rewards and dynamics. On the other hand, we establish the MIS-based identification result for estimating the expected total reward. Interestingly, our MIS-based identification result does not require the memoryless assumption and allow the existence of unmeasured confounders that affects future rewards and dynamics. Therefore our identification result via MIS can be applied in a more general confounded MDP.

In the meanwhile, to tackle the above challenge (ii), we employ pessimism to achieve policy learning. Specifically, when using VF-based identification, we first formulate a minimax estimator of the state-value function via the newly established estimating equation. Then, we construct a confidence set of such a minimax estimator, so that the true state-value function lies within the confidence set with a high probability. Finally, we search for the best policy that maximizes the most conservative expected total reward associated with the estimated state-value function within the confidence set. As a theoretical contribution, under data coverage assumption only for an optimal policy π^* and realizability assumption for all policies, we show that the suboptimality of the learned best policy is upper bounded by $O(\log(NT)(NT)^{-1/2})$. It is worth noting that our theoretical analysis does not assume that the observational data are generated from stationary distribution or even independent, which is widely imposed in related literature (Farahmand et al., 2016; Nachum et al., 2019; Tang et al., 2019; Xie et al., 2021; Kallus and Uehara, 2022). Without imposing such a restrictive assumption, inspired by Wang et al. (2021a), our convergence analysis relies on novel concentration inequalities for geometrically ergodic sequences, which significantly increases the applicability of our results in practice. In the meanwhile, pessimism with MIS-based identification achieves a similar result by imposing realizability assumption only for an optimal policy π^* and data coverage assumption for all policies. Further, by combining VF- and MIS-based identification results, we propose a doubly robust (DR) estimator of the optimal policy π^*. Theoretically, for such a DR estimator, we show a similar suboptimality at a rate of $O(\log(NT)(NT)^{-1/2})$, but only requiring that either the assumptions imposed in VF-based method or those imposed in MIS-based one hold. See Table 1.1 for an overview of our theoretical results.

Contribution. As a summary, our contribution is three-fold. First, by leveraging IVs, we provide VF- and MIS-based identification results. Second, by employing pessimism, we construct estimators of the optimal policy π^* via VF- and MIS-based identification. Further, by combining VF- and MIS-based identification, we propose a DR-based algorithm for estimating π^*. Third, under mild conditions on data coverage and realizability, we show that the suboptimalities of the proposed estimators are upper bounded by

Methods	Data Coverage	Realizability	Identifiability
VF-based	$\Vert w^{\pi^*}\Vert_\infty \leq C_*$	$w^{\pi^*} \in \mathcal{W}$, $V^\pi \in \mathcal{V}\ \forall \pi \in \Pi$	$J(\pi^*)$ is identifiable
MIS-based	$\Vert w^{\pi}\Vert_\infty \leq C_*\ \forall \pi \in \Pi$	$V^{\pi^*} \in \mathcal{V}$, $w^\pi \in \mathcal{W}\ \forall \pi \in \Pi$	$J(\pi)$ is identifiable $\forall \pi \in \Pi$
DR-based	$\Vert w^{\pi^*}\Vert_\infty \leq C_*$, $w^{\pi^*} \in \mathcal{W}$, $V^\pi \in \mathcal{V}\ \forall \pi \in \Pi$, and $J(\pi^*)$ is identifiable; ***OR*** $\Vert w^{\pi}\Vert_\infty \leq C_*\ \forall \pi \in \Pi$, $V^{\pi^*} \in \mathcal{V}$, $w^\pi \in \mathcal{W}\ \forall \pi \in \Pi$, and $J(\pi)$ is identifiable $\forall \pi \in \Pi$		

Table 1.1. Assumptions required by our VF-, MIS-, and DR-based methods, where w^π is the density ratio between visitation measures induced by the policy π and the behavior policy b (see (4.2.1) for a detailed definition), and V^π is the state-value function of the policy π. Here $\mathcal{V}$ and Π are function classes, and C_* is a positive absolute constant.

$O(\log(NT)(NT)^{-1/2})$, without requiring that the observational data is generated from stationary distribution or even independent.

Related Work. Our work is related to the line of works that study RL under the presence of unobserved confounders. Zhang and Bareinboim (2019) propose an online RL method to solve dynamic treatment regimes in a finite-horizon setting with the presence of confounded observational data. Their method relies on sensitivity analysis, which constructs a set of possible models based on the confounded observational data to obtain partial identification. Also, to incorporate the observational data into the finite-horizon RL, Wang et al. (2021b) propose deconfounded optimistic value iteration, which is an online algorithm with a provable regret guarantee. To ensure the identifiability through the observational data, they impose the backdoor criterion (Pearl, 2009; Peters et al., 2017) when confounders are partially observed, and also the frontdoor criterion when confounders are unobserved. Our work is also closely related to Liao et al. (2021a), where they propose an IV-aided value iteration algorithm to study confounded MDPs in the offline setting. It is worth mentioning that they only consider the finite-horizon MDP, where the transition dynamics is a linear function of some known feature map. In the meanwhile, to ensure identifiability, they assume that the unobserved confounders are Gaussian noise, which does not affect the immediate reward and only affect the transition dynamics in an additive manner. In contrast, we consider infinite-horizon confounded MDP without such specific assumptions on the structure of the model, which brings significant technical challenges. With unobserved confounders, Kallus and Zhou (2020) study off-policy evaluation in the infinite-horizon setting based on sensitivity analysis, which imposes additional assumptions on how strong the unobserved confounding can possibly be. Bennett et al. (2021) also study off-policy evaluation in the infinite-horizon MDP via a conditional moment restriction. In the meanwhile, to ensure identifiability, Namkoong et al. (2020) consider the case where the unmeasured confounders affect only one of the decisions made.

Our work is also related to the line of research on policy evaluation and policy learning in the offline setting assuming unconfoundedness. In terms of off-policy evaluation, most works either employs a VF-based method (Ernst et al., 2005; Shi et al., 2020; Liao et al., 2021b; Uehara et al., 2021), or an MIS-based method (Liu et al., 2018; Nachum et al., 2019; Zhang et al., 2020; Wang et al., 2021a; Uehara et al., 2021). Our work is also related to research that propose DR estimators in off-policy evaluation. See Jiang and Li (2016); Thomas and Brunskill (2016); Tang et al. (2019); Kallus and Uehara (2020); Uehara et al. (2020); Kallus and Uehara (2022) for this line of research. As for policy learning in the offline setting, Munos and Szepesvári (2008)and Antos et al. (2008b) prove that fitted value and policy iterations converge to an optimal policy under the data coverage assumption and realizability assumption for all policies, respectively. By employing pessimism, Xie et al. (2021) guarantee a near-optimal policy under the realizability and the completeness assumptions for all policies, while Jiang and Huang (2020) provide a similar guarantee under the data coverage assumption for the optimal policy and the realizability assumption for all policies. More recently, Zhan et al. (2022) claims that they can learn a near-optimal policy under the data coverage and realizability assumptions for the optimal policy. Their method is built upon a regularized version of the LP formulation of MDPs and thus working on a class of regularized policies. However, due to regularization, the policy learned by Zhan et al. (2022) is typically suboptimal even given infinite data. Moreover, their realizability assumption is imposed on the regularized value function, making it difficult to interpret and compare with other works. In contrast, our work still focuses on a non-regularized setting under standard realizability and data coverage assumptions, even under the presence of unobserved confounders and non-stationary dependent observational data.

CHAPTER 2

Single-Timescale Actor-Critic Provably Finds Globally Optimal Policy

2.1. Background

In this section, we introduce the background on discounted Markov decision processes (MDPs) and actor-critic methods.

Notation. We denote by $[n]$ the set $\{1, 2, \ldots, n\}$. For any measure ν and $1 \le p \le \infty$, we denote by $\|f\|_{\nu,p} = (\int_{\mathcal{X}} |f(x)|^p \mathrm{d}\nu)^{1/p}$ and $\|f\|_p = (\int_{\mathcal{X}} |f(x)|^p \mathrm{d}\mu)^{1/p}$, where μ is the Lebesgue measure.

2.1.1. Discounted MDP

A discounted MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, P, \zeta, \mathcal{R}, \gamma)$. Here $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, $P \colon \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the Markov transition kernel, $\zeta \colon \mathcal{S} \to [0, 1]$ is the initial state distribution, $\mathcal{R} \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the deterministic reward function, and $\gamma \in [0, 1)$ is the discount factor. A policy $\pi(a \,|\, s)$ measures the probability of taking the action a at the state s. We focus on a family of parameterized policies defined as follows,

$$\Pi = \{\pi_\theta(\cdot \,|\, s) \in \mathcal{P}(\mathcal{A}) \colon s \in \mathcal{S}\}, \tag{2.1.1}$$

where $\mathcal{P}(\mathcal{A})$ is the probability simplex on the action space $\mathcal{A}$ and θ is the parameter of the policy π_θ. For any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define the action-value function as follows,

$$Q^\pi(s, a) = (1 - \gamma) \cdot \mathbb{E}_\pi \Bigl[\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{R}(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a \Bigr], \tag{2.1.2}$$

where $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ and $a_{t+1} \sim \pi(\cdot \,|\, s_{t+1})$ for any $t \geq 0$. We use $\mathbb{E}_\pi[\cdot]$ to denote that the actions follow the policy π, which further affect the transition of the states. We aim to find an optimal policy π^* such that $Q^{\pi^*}(s, a) \geq Q^\pi(s, a)$ for any policy π and state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. That is to say, such an optimal policy π^* attains a higher expected total reward than any other policy π, regardless of the initial state-action pair (s, a). For notational convenience, we denote by $Q^*(s, a) = Q^{\pi^*}(s, a)$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ hereafter.

Meanwhile, we denote by $\nu_\pi(s)$ and $\rho_\pi(s, a) = \nu_\pi(s) \cdot \pi(a \,|\, s)$ the stationary state distribution and stationary state-action distribution of the policy π, respectively, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. Correspondingly, we denote by $\nu^*(s)$ and $\rho^*(s, a)$ the stationary state distribution and stationary state-action distribution of the optimal policy π^*, respectively, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. For ease of presentation, given any functions $g_1 \colon \mathcal{S} \to \mathbb{R}$ and $g_2 \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define two operators $\mathbb{P}$ and $\mathbb{P}^\pi$ as follows,

$$[\mathbb{P} g_1](s, a) = \mathbb{E}[g_1(s_1) \,|\, s_0 = s, a_0 = a], \quad [\mathbb{P}^\pi g_2](s, a) = \mathbb{E}_\pi[g_2(s_1, a_1) \,|\, s_0 = s, a_0 = a], \tag{2.1.3}$$

where $s_1 \sim P(\cdot \,|\, s_0, a_0)$ and $a_1 \sim \pi(\cdot \,|\, s_1)$. Intuitively, given the current state-action pair (s_0, a_0), the operator $\mathbb{P}$ pushes the agent to its next state s_1 following the Markov transition kernel $P(\cdot\,|\,s_0, a_0)$, while the operator $\mathbb{P}^\pi$ pushes the agent to its next state-action pair (s_1, a_1) following the Markov transition kernel $P(\cdot\,|\,s_0, a_0)$ and policy $\pi(\cdot\,|\,s_1)$. These operators also relate to the Bellman evaluation operator $\mathcal{T}^\pi$, which is defined for any function $g\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows,

$$\mathcal{T}^\pi g = (1-\gamma) \cdot \mathcal{R} + \gamma \cdot \mathbb{P}^\pi g. \tag{2.1.4}$$

The Bellman evaluation operator $\mathcal{T}^\pi$ is used to characterize the actor-critic method in the following section. By the definition in (2.1.2), it is straightforward to verify that the action-value function Q^π is the fixed point of the Bellman evaluation operator $\mathcal{T}^\pi$ defined in (2.1.4), that is, $Q^\pi = \mathcal{T}^\pi Q^\pi$ for any policy π. For notational convenience, we let $\mathbb{P}^\ell$ denote the ℓ-fold composition $\underbrace{\mathbb{P}\mathbb{P}\cdots\mathbb{P}}_{\ell}$. Such notation is also adopted for other linear operators such as $\mathbb{P}^\pi$ and $\mathcal{T}^\pi$.

2.1.2. Actor-Critic Method

To obtain an optimal policy π^*, the actor-critic method (Konda and Tsitsiklis, 2000) aims to maximize the expected total reward as a function of the policy, which is equivalent to solving the following maximization problem,

$$\max_{\pi\in\Pi} J(\pi) = \mathbb{E}_{s\sim\zeta, a\sim\pi(\cdot\,|\,s)}\bigl[Q^\pi(s,a)\bigr], \tag{2.1.5}$$

where ζ is the initial state distribution, Q^π is the action-value function defined in (2.1.2), and the family of parameterized polices Π is defined in (2.1.1). The actor-critic method solves the maximization problem in (2.1.5) via first-order optimization using an estimator of the policy gradient $\nabla_\theta J(\pi)$. Here θ is the parameter of the policy π. In detail, by the policy gradient theorem (Sutton et al., 2000), we have

$$\nabla_\theta J(\pi) = \mathbb{E}_{(s,a)\sim \varrho_\pi}\big[Q^\pi(s,a)\cdot \nabla_\theta \log \pi(a\,|\,s)\big]. \tag{2.1.6}$$

Here ϱ_π is the state-action visitation measure of the policy π, which is defined as $\varrho_\pi(s,a) = (1-\gamma)\cdot\sum_{t=0}^\infty \gamma^t \cdot \Pr[s_t = s, a_t = a]$. Based on the closed form of the policy gradient in (2.1.6), the actor-critic method consists of the following two parts: (i) the critic update, where a policy evaluation algorithm is invoked to estimate the action-value function Q^π, e.g., by applying the Bellman evaluation operator $\mathcal{T}^\pi$ to the current estimator of Q^π, and (ii) the actor update, where a policy improvement algorithm, e.g., the policy gradient method, is invoked using the updated estimator of Q^π.

In this paper, we consider the following variant of the actor-critic method,

$$\begin{aligned} &\pi_{k+1} \leftarrow \mathop{\mathrm{argmax}}_{\pi\in\Pi} \mathbb{E}_{\nu_{\pi_k}}\big[\langle Q_k(s,\cdot), \pi(\cdot\,|\,s)\rangle - \beta\cdot \mathrm{KL}\big(\pi(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s)\big)\big], \\ &Q_{k+1}(s,a) \leftarrow \mathbb{E}_{\pi_{k+1}}\big[(1-\gamma)\cdot\mathcal{R}(s_0,a_0) + \gamma\cdot Q_k(s_1,a_1)\,\big|\,s_0 = s, a_0 = a\big], \end{aligned} \tag{2.1.7}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, where $s_1\sim P(\cdot\,|\,s_0,a_0)$, $a_1\sim\pi_{k+1}(\cdot\,|\,s_1)$, and we write $\mathbb{E}_{\nu_{\pi_k}}[\cdot] = \mathbb{E}_{s\sim\nu_{\pi_k}}[\cdot]$ for notational convenience. Here Π is defined in (2.1.1) and $\mathrm{KL}(\pi(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s))$ is the Kullback-Leibler (KL) divergence between $\pi(\cdot\,|\,s)$ and $\pi_k(\cdot\,|\,s)$, which is defined for any $s\in\mathcal{S}$ as follows,

$$\mathrm{KL}\big(\pi(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s)\big) = \sum_{a\in\mathcal{A}} \log\Big(\frac{\pi(a\,|\,s)}{\pi_k(a\,|\,s)}\Big)\cdot\pi(a\,|\,s).$$

In (2.1.7), the actor update uses the proximal policy optimization (PPO) method (Schulman et al., 2017), while the critic update applies the Bellman evaluation operator $\mathcal{T}^{\pi_{k+1}}$

defined in (2.1.4) to Q_k only once, which is the current estimator of the action-value function. Furthermore, we remark that the updates in (2.1.7) provide a general framework in the following two aspects. First, the critic update can be extended to letting $Q_{k+1} \leftarrow (\mathcal{T}^{\pi_{k+1}})^\tau Q_k$ for any fixed $\tau \geq 1$, which corresponds to updating the value function via τ-step rollouts following π_{k+1}. Here we only focus on the case with $\tau = 1$ for simplicity. Our theory can be easily modified for any fixed τ. Moreover, the KL divergence used in the actor step can also be replaced by other Bregman divergences between probability distributions over $\mathcal{A}$. Second, the actor and critic updates in (2.1.7) is a general template that admits both on- and off-policy evaluation methods and various function approximators in the actor and critic. In the next section, we present an incarnation of (2.1.7) with on-policy sampling and linear and neural network function approximation.

Furthermore, for analyzing the actor-critic method, most existing works (Yang et al., 2019a; Wang et al., 2019; Agarwal et al., 2019; Fu et al., 2019; Liu et al., 2019) rely on (approximately) obtaining $Q^{\pi_{k+1}}$ at each iteration, which is equivalent to applying the Bellman evaluation operator $\mathcal{T}^{\pi_{k+1}}$ infinite times to Q_k. This is usually achieved by minimizing the mean-squared Bellman error $\|Q - \mathcal{T}^{\pi_{k+1}} Q\|^2_{\rho_{\pi_{k+1}},2}$ using stochastic semi-gradient descent, e.g., as in the temporal-difference method (Sutton, 1988), to update the critic for sufficiently many iterations. The unique global minimizer of the mean-squared Bellman error gives the action-value function $Q^{\pi_{k+1}}$, which is used in the actor update. Meanwhile, the two-timescale setting is also considered in existing works (Borkar and Konda, 1997; Konda and Tsitsiklis, 2000; Xu et al., 2019, 2020; Wu et al., 2020; Hong et al., 2020), which require the actor to be updated more slowly than the critic in an asymptotic sense. Such a requirement is usually satisfied by forcing the ratio between the stepsizes of the actor and critic updates to go to zero asymptotically.

In comparison with the setting with bi-level updates, we consider the single-timescale actor and critic updates in (2.1.7), where the critic involves only one step of update, that is, applying the Bellman evaluation operator $\mathcal{T}^{\pi}$ to Q_k only once. Meanwhile, in comparison with the two-timescale setting, where the actor and critic are updated simultaneously but with the ratio between their stepsizes asymptotically going to zero, the single-timescale setting is able to achieve a faster rate of convergence by allowing the actor to be updated with a larger stepsize, while updating the critic simultaneously. In particular, such a single-timescale setting better captures a broader range of practical algorithms (Peters and Schaal, 2008a; Schulman et al., 2015; Mnih et al., 2016; Schulman et al., 2017; Haarnoja et al., 2018), where the stepsize of the actor is not asymptotically zero. In §2.2, we discuss the implementation of the updates in (2.1.7) for different schemes of function approximation. In §2.3, we compare the rates of convergence between the two-timescale and single-timescale settings.

2.2. Algorithms

We consider two settings, where the actor and critic are parameterized using linear functions and deep neural networks, respectively. We consider the energy-based policy $\pi_\theta(a \,|\, s) \propto \exp(\tau^{-1} f_\theta(s,a))$, where the energy function $f_\theta(s,a)$ is parameterized with the parameter θ. Also, for the (estimated) action-value function, we consider the parameterization $Q_\omega(s,a)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where ω is the parameter. For such parameterizations of the actor and critic, the updates in (2.1.7) have the following forms.

Actor Update. The following proposition gives the closed form of π_{k+1} in (2.1.7).

Proposition 2.2.1. Let $\pi_{\theta_k}(a \,|\, s) \propto \exp(\tau_k^{-1} f_{\theta_k}(s,a))$ be an energy-based policy and

$$\widetilde{\pi}_{k+1} = \operatorname*{argmax}_{\pi} \mathbb{E}_{\nu_k}\bigl[\langle Q_{\omega_k}(s,\cdot), \pi(\cdot \,|\, s)\rangle - \beta \cdot \mathrm{KL}\bigl(\pi(\cdot \,|\, s) \,\|\, \pi_{\theta_k}(\cdot \,|\, s)\bigr)\bigr].$$

Then $\widetilde{\pi}_{k+1}$ has the following closed form,

$$\widetilde{\pi}_{k+1}(a \,|\, s) \propto \exp\bigl(\beta^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a)\bigr),$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $\nu_k = \nu_{\pi_{\theta_k}}$ is the stationary state distribution of π_{θ_k}.

Proof. See §A.5.1 for a detailed proof. □

Motivated by Proposition 2.2.1, to implement the actor update in (2.1.7), we update the actor parameter θ by solving the following minimization problem,

$$\theta_{k+1} \leftarrow \operatorname*{argmin}_{\theta} \mathbb{E}_{\rho_k}\Bigl[\bigl(f_\theta(s,a) - \tau_{k+1} \cdot \bigl(\beta^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a)\bigr)\bigr)^2\Bigr], \tag{2.2.1}$$

where $\rho_k = \rho_{\pi_{\theta_k}}$ is the stationary state-action distribution of π_{θ_k}.

Critic Update. To implement the critic update in (2.1.7), we update the critic parameter ω by solving the following minimization problem,

$$\omega_{k+1} \leftarrow \operatorname*{argmin}_{\omega} \mathbb{E}_{\rho_{k+1}}\Bigl[\bigl([Q_\omega - (1-\gamma) \cdot \mathcal{R} - \gamma \cdot \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}](s,a)\bigr)^2\Bigr], \tag{2.2.2}$$

where $\rho_{k+1} = \rho_{\pi_{\theta_{k+1}}}$ is the stationary state-action distribution of $\pi_{\theta_{k+1}}$ and the operator $\mathbb{P}^\pi$ is defined in (2.1.3).

2.2.1. Linear Function Approximation

In this section, we consider linear function approximation. More specifically, we parameterize the action-value function using $Q_\omega(s,a) = \omega^\top \varphi(s,a)$ and the energy function of the energy-based policy π_θ using $f_\theta(s,a) = \theta^\top \varphi(s,a)$. Here $\varphi(s,a) \in \mathbb{R}^d$ is the feature vector, where $d > 0$ is the dimension. Without loss of generality, we assume that $\|\varphi(s,a)\|_2 \le 1$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, which can be achieved by normalization.

Actor Update. The minimization problem in (2.2.1) admits the following closed-form solution,

$$\theta_{k+1} = \tau_{k+1} \cdot (\beta^{-1}\omega_k + \tau_k^{-1}\theta_k), \tag{2.2.3}$$

which corresponds to a step of the natural policy gradient method (Kakade, 2002).

Critic Update. The minimization problem in (2.2.2) admits the following closed-form solution,

(2.2.4)

$$\widetilde{\omega}_{k+1} = \bigl(\mathbb{E}_{\rho_{k+1}}[\varphi(s,a)\varphi(s,a)^\top]\bigr)^{-1} \cdot \mathbb{E}_{\rho_{k+1}}\bigl[[(1-\gamma)\cdot \mathcal{R} + \gamma \cdot \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}](s,a) \cdot \varphi(s,a)\bigr].$$

Since the closed-form solution $\widetilde{\omega}_{k+1}$ in (2.2.4) involves the expectation over the stationary state-action distribution ρ_{k+1} of $\pi_{\theta_{k+1}}$, we use data to approximate such an expectation. More specifically, we sample $\{(s_{\ell,1}, a_{\ell,1})\}_{\ell\in[N]}$ and $\{(s_{\ell,2}, a_{\ell,2}, r_{\ell,2}, s'_{\ell,2}, a'_{\ell,2})\}_{\ell\in[N]}$ such that $(s_{\ell,1}, a_{\ell,1}) \sim \rho_{k+1}$, $(s_{\ell,2}, a_{\ell,2}) \sim \rho_{k+1}$, $r_{\ell,2} = \mathcal{R}(s_{\ell,2}, a_{\ell,2})$, $s'_{\ell,2} \sim P(\cdot \,|\, s_{\ell,2}, a_{\ell,2})$, and $a'_{\ell,2} \sim \pi_{\theta_{k+1}}(\cdot \,|\, s'_{\ell,2})$, where N is the sample size. We approximate $\widetilde{\omega}_{k+1}$ using ω_{k+1}, which is defined as follows,

$$\omega_{k+1} = \Gamma_R \Big\{ \Big(\sum_{\ell=1}^{N} \varphi(s_{\ell,1}, a_{\ell,1}) \varphi(s_{\ell,1}, a_{\ell,1})^\top \Big)^{-1} \cdot \sum_{\ell=1}^{N} \big((1-\gamma) \cdot r_{\ell,2} + \gamma \cdot Q_{\omega_k}(s'_{\ell,2}, a'_{\ell,2}) \big) \cdot \varphi(s_{\ell,2}, a_{\ell,2}) \Big\}. \tag{2.2.5}$$

Here Γ_R is the projection operator, which projects the parameter onto the centered ball with radius R in $\mathbb{R}^d$. Such a projection operator stabilizes the algorithm (Konda and Tsitsiklis, 2000; Bhatnagar et al., 2009). It is worth mentioning that one may also view the update in (2.2.5) as one step of the least-squares temporal difference method (Bradtke and Barto, 1996), which can be modified for the off-policy setting (Antos et al., 2007; Yu, 2010; Liu et al., 2018; Nachum et al., 2019; Xie et al., 2019; Zhang et al., 2020; Uehara and Jiang, 2019; Nachum and Dai, 2020). Such a modification allows the data points in (2.2.5) to be reused in the subsequent iterations, which further improves the sample complexity. Specifically, let $\rho_{\text{bhv}} \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be the stationary state-action distribution induced by a behavioral policy π_{bhv}. We replace the actor and critic updates in (2.2.1) and (2.2.2) by

$$\theta_{k+1} \leftarrow \operatorname*{argmin}_{\theta} \mathbb{E}_{\rho_{\text{bhv}}} \Big[\big(f_\theta(s,a) - \tau_{k+1} \cdot \big(\beta^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a) \big) \big)^2 \Big], \tag{2.2.6}$$

$$\omega_{k+1} \leftarrow \operatorname*{argmin}_{\omega} \mathbb{E}_{\rho_{\text{bhv}}} \Big[\big([Q_\omega - (1-\gamma) \cdot \mathcal{R} - \gamma \cdot \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}](s,a) \big)^2 \Big], \tag{2.2.7}$$

respectively. With linear function approximation, the actor update in (2.2.6) is reduced to (2.2.3), while the critic update in (2.2.7) admits a closed form solution

$$\widetilde{\omega}_{k+1} = \bigl(\mathbb{E}_{\rho_{\mathrm{bhv}}}[\varphi(s,a)\varphi(s,a)^\top]\bigr)^{-1} \cdot \mathbb{E}_{\rho_{\mathrm{bhv}}}\bigl[[(1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi_{\theta_{k+1}}}Q_{\omega_k}](s,a)\cdot\varphi(s,a)\bigr],$$

which can be well approximated using state-action pairs drawn from ρ_{bhv}. See §2.3 for a detailed discussion.

Finally, by assembling the updates in (2.2.3) and (2.2.5), we present the linear actor-critic method in Algorithm 1 as follows.

Algorithm 1 Linear Actor-Critic Method

```
**Input:** Number of iterations $K$, sample size $N$, temperature parameter $\beta$.
**Initialization:** Set $\tau_0 \leftarrow \infty$, and randomly initialize the actor parameter $\theta_0$ and the critic parameter $\omega_0$.
**for** $k = 0, 1, 2, \ldots, K$ **do**
    **Actor Update:** Update $\theta_{k+1}$ via (2.2.3) with $\tau_{k+1}^{-1} = (k+1)\cdot\beta^{-1}$.
    **Critic Update:** Sample $\{(s_{\ell,1}, a_{\ell,1})\}_{\ell\in[N]}$ and $\{(s_{\ell,2}, a_{\ell,2}, r_{\ell,2}, s'_{\ell,2}, a'_{\ell,2})\}_{\ell\in[N]}$ as specified in §2.2.1. Update $\omega_{k+1}$ via (2.2.5).
**end for**
**Output:** $\{\pi_{\theta_k}\}_{k\in[K+1]}$, where $\pi_{\theta_k} \propto \exp(\tau_k^{-1} f_{\theta_k})$.
```

2.2.2. Deep Neural Network Approximation

In this section, we consider deep neural network approximation. We first formally define deep neural networks. Then we introduce the actor-critic method under such a parameterization.

A deep neural network (DNN) $u_\theta(x)$ with the input $x \in \mathbb{R}^d$, depth H, and width m is defined as

$$x^{(0)} = x, \quad x^{(h)} = \frac{1}{\sqrt{m}} \cdot \sigma(W_h^\top x^{(h-1)}), \text{ for } h \in [H], \quad u_\theta(x) = b^\top x^{(H)}. \tag{2.2.8}$$

Here $\sigma\colon \mathbb{R}^m \to \mathbb{R}^m$ is the rectified linear unit (ReLU) activation function, which is define as $\sigma(y) = (\max\{0, y_1\}, \ldots, \max\{0, y_m\})^\top$ for any $y = (y_1, \ldots, y_m)^\top \in \mathbb{R}^m$. Also, we have $b \in \{-1, 1\}^m$, $W_1 \in \mathbb{R}^{d\times m}$, and $W_h \in \mathbb{R}^{m\times m}$ for $2 \leq h \leq H$. Meanwhile, we denote the parameter of the DNN u_θ as $\theta = (\mathrm{vec}(W_1)^\top, \ldots, \mathrm{vec}(W_H)^\top)^\top \in \mathbb{R}^{m_{\text{all}}}$ with $m_{\text{all}} = md + (H-1)m^2$. We call $\{W_h\}_{h\in[H]}$ the weight matrices of θ. Without loss of generality, we normalize the input x such that $\|x\|_2 = 1$.

We initialize the DNN such that each entry of W_h follows the standard Gaussian distribution $\mathcal{N}(0, 1)$ for any $h \in [H]$, while each entry of b follows the uniform distribution $\mathrm{Unif}(\{-1, 1\})$. Without loss of generality, we fix b during training and only optimize $\{W_h\}_{h\in[H]}$. We denote the initialization of the parameter θ as $\theta_0 = (\mathrm{vec}(W_1^0)^\top, \ldots, \mathrm{vec}(W_H^0)^\top)^\top$. Meanwhile, we restrict θ within the ball $\mathcal{B}(\theta_0, R)$ during training, which is defined as follows,

$$\mathcal{B}(\theta_0, R) = \bigl\{\theta \in \mathbb{R}^{m_{\text{all}}}\colon \|W_h - W_h^0\|_{\mathrm{F}} \leq R, \text{ for } h \in [H]\bigr\}. \tag{2.2.9}$$

Here $\{W_h\}_{h\in[H]}$ and $\{W_h^0\}_{h\in[H]}$ are the weight matrices of θ and θ_0, respectively. By (2.2.9), we have $\|\theta - \theta_0\|_2 \leq R\sqrt{H}$ for any $\theta \in \mathcal{B}(\theta_0, R)$. Now, we define the family of DNNs as

$$\mathcal{U}(m, H, R) = \bigl\{u_\theta\colon \theta \in \mathcal{B}(\theta_0, R)\bigr\}, \tag{2.2.10}$$

where u_θ is a DNN with depth H and width m.

We parameterize the action-value function using $Q_\omega(s,a) \in \mathcal{U}(m_\mathrm{c}, H_\mathrm{c}, R_\mathrm{c})$ and the energy function of the energy-based policy π_θ using $f_\theta(s,a) \in \mathcal{U}(m_\mathrm{a}, H_\mathrm{a}, R_\mathrm{a})$. Here $\mathcal{U}(m_\mathrm{c}, H_\mathrm{c}, R_\mathrm{c})$ and $\mathcal{U}(m_\mathrm{a}, H_\mathrm{a}, R_\mathrm{a})$ are the families of DNNs defined in (2.2.10). Hereafter we assume that the energy function f_θ and the action-value function Q_ω share the same architecture and initialization, i.e., $m_\mathrm{a} = m_\mathrm{c}$, $H_\mathrm{a} = H_\mathrm{c}$, $R_\mathrm{a} = R_\mathrm{c}$, and $\theta_0 = \omega_0$. Such shared architecture and initialization of the DNNs ensure that the parameterizations of the policy and the action-value function are approximately compatible. See Sutton et al. (2000); Konda and Tsitsiklis (2000); Kakade (2002); Peters and Schaal (2008a); Wang et al. (2019) for a detailed discussion.

Actor Update. To solve (2.2.1), we use projected stochastic gradient descent, whose n-th iteration has the following form,

$$\begin{aligned}\theta(n+1) \leftarrow &\Gamma_{\mathcal{B}(\theta_0, R_\mathrm{a})}\big(\theta(n) \\ &\qquad - \alpha \cdot \big(f_{\theta(n)}(s,a) - \tau_{k+1} \cdot \big(\beta^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a)\big)\big) \cdot \nabla_\theta f_{\theta(n)}(s,a)\big).\end{aligned}$$

Here $\Gamma_{\mathcal{B}(\theta_0, R_\mathrm{a})}$ is the projection operator, which projects the parameter onto the ball $\mathcal{B}(\theta_0, R_\mathrm{a})$ defined in (2.2.9). The state-action pair (s,a) is sampled from the stationary state-action distribution ρ_k. We summarize the update in Algorithm 5, which is deferred to §A.1 of the appendix.

Critic Update. To solve (2.2.2), we apply projected stochastic gradient descent. More specifically, at the n-th iteration of projected stochastic gradient descent, we sample a tuple (s, a, r, s', a'), where $(s,a) \sim \rho_{k+1}$, $r = \mathcal{R}(s,a)$, $s' \sim P(\cdot \,|\, s,a)$, and $a' \sim \pi_{\theta_{k+1}}(\cdot \,|\, s')$.

We define the residual at the n-th iteration as $\delta(n) = Q_{\omega(n)}(s,a) - (1-\gamma)\cdot r - \gamma \cdot Q_{\omega_k}(s', a')$. Then the n-th iteration of projected stochastic gradient descent has the following form,

$$\omega(n+1) \leftarrow \Gamma_{\mathcal{B}(\omega_0, R_{\mathrm{c}})}\big(\omega(n) - \eta \cdot \delta(n) \cdot \nabla_\omega Q_{\omega(n)}(s,a)\big).$$

Here $\Gamma_{\mathcal{B}(\omega_0, R_{\mathrm{c}})}$ is the projection operator, which projects the parameter onto the ball $\mathcal{B}(\omega_0, R_{\mathrm{c}})$ defined in (2.2.9). We summarize the update in Algorithm 6, which is deferred to §A.1 of the appendix.

By assembling Algorithms 5 and 6, we present the deep neural actor-critic method in Algorithm 4, which is deferred to §A.1 of the appendix.

Finally, we remark that the off-policy actor and critic updates given in (2.2.6) and (2.2.7) can also incorporate deep neural network approximation with a slight modification, which enables data reuse in the algorithm.

2.3. Theoretical Results

In this section, we upper bound the regret of the linear actor-critic method. We defer the analysis of the deep neural actor-critic method to §A.2 of the appendix. Hereafter we assume that $|\mathcal{R}(s,a)| \leq \mathcal{R}_{\max}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $\mathcal{R}_{\max}$ is a positive absolute constant. First, we impose the following assumptions. Recall that ρ^* is the stationary state-action distribution of π^*, while ρ_k is the stationary state-action distribution of π_{θ_k}. Moreover, let $\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be a state-action distribution with respect to which we aim to characterize the performance of the actor-critic algorithm. Specifically, after $K+1$ actor updates, we are interest in upper bounding the following regret

$$\mathbb{E}\Big[\sum_{k=0}^{K}\big(\|Q^* - Q^{\pi_{\theta_{k+1}}}\|_{\rho,1}\big)\Big] = \mathbb{E}\Big[\sum_{k=0}^{K}\big(Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a)\big)\Big], \tag{2.3.1}$$

where the expectation is taken with respect to $\{\theta_k\}_{k\in[K+1]}$ and $(s,a) \sim \rho$. Here we allow ρ to be any fixed distribution for generality, which might be different from ρ^*.

Assumption 2.3.1 (Concentrability Coefficient)**.** The following statements hold.

(i) There exists a positive absolute constant ϕ^* such that $\phi_k^* \le \phi^*$ for any $k \ge 1$, where $\phi_k^* = \|\mathrm{d}\rho^*/\mathrm{d}\rho_k\|_{\rho_k,2}$.

(ii) We assume that for any $k \ge 1$ and a sequence of policies $\{\pi_i\}_{i\ge 1}$, the k-step future-state-action distribution $\rho\mathbb{P}^{\pi_1}\cdots\mathbb{P}^{\pi_k}$ is absolutely continuous with respect to ρ^*, where ρ is the same as the one in (2.3.1) Also, it holds for such ρ that

$$C_{\rho,\rho^*} = (1-\gamma)^2 \sum_{k=1}^{\infty} k^2\gamma^k \cdot c(k) < \infty,$$

where $c(k) = \sup_{\{\pi_i\}_{i\in[k]}} \|\mathrm{d}(\rho\mathbb{P}^{\pi_1}\cdots\mathbb{P}^{\pi_k})/\mathrm{d}\rho^*\|_{\rho^*,\infty}$.

In Assumption 2.3.1, C_{ρ,ρ^*} is known as the discounted-average concentrability coefficient of the future-state-action distributions. Similar assumptions are commonly imposed in the literature (Szepesvári and Munos, 2005; Munos and Szepesvári, 2008; Antos et al., 2008a,b; Scherrer, 2013; Scherrer et al., 2015; Farahmand et al., 2016; Yang et al., 2019c; Geist et al., 2019; Chen and Jiang, 2019).

Assumption 2.3.2 (Zero Approximation Error)**.** It holds for any $\omega, \theta \in \mathcal{B}(0,R)$ that

$$\inf_{\bar\omega\in\mathcal{B}(0,R)} \mathbb{E}_{\rho_{\pi_\theta}}\big[\big([\mathcal{T}^{\pi_\theta}Q_\omega - \bar\omega^\top\varphi](s,a)\big)^2\big] = 0,$$

where $\mathcal{T}^{\pi_\theta}$ is defined in (2.1.4).

Assumption 2.3.2 states that the Bellman evaluation operator maps a linear function to a linear function. Such an assumption only aims to simplify the presentation of our results. If the approximation error is nonzero, we only need to incorporate an additional bias term into the rate of convergence.

Assumption 2.3.3 (Well-Conditioned Feature)**.** The minimum singular value of the matrix $\mathbb{E}_{\rho_k}[\varphi(s,a)\varphi(s,a)^\top]$ is uniformly lower bounded by a positive absolute constant σ^* for any $k \geq 1$.

Assumption 2.3.3 ensures that the minimization problem in (2.2.2) admits a unique minimizer, which is used in the critic update. Similar assumptions are commonly imposed in the literature (Bhandari et al., 2018; Zou et al., 2019).

Under Assumptions 2.3.1, 2.3.2, and 2.3.3, we upper bound the regret of Algorithm 1 in the following theorem.

Theorem 2.3.4. We assume that Assumptions 2.3.1, 2.3.2, and 2.3.3 hold. Let ρ be a state-action distribution satisfying (ii) of Assumption 2.3.1. Also, for any confidence parameter $\delta \in (0,1)$ and sufficiently large number of iterations $K > 0$, let $\beta = K^{1/2}$, $N = \Omega(KC^2_{\rho,\rho^*} \cdot (\phi^*/\sigma^*)^2 \cdot \log^2(KN/\delta))$, and the sequence of policy parameters $\{\theta_k\}_{k\in[K+1]}$ be generated by Algorithm 1. It holds with probability at least $1-\delta$ that

$$\mathbb{E}_\rho\Big[\sum_{k=0}^{K}\big(Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a)\big)\Big] \leq \big(2(1-\gamma)^{-3} \cdot \log|\mathcal{A}| + O(1)\big) \cdot K^{1/2}, \tag{2.3.2}$$

where the expectation is taken with respect $(s,a) \sim \rho$.

Proof. We sketch the proof in §2.4. See §A.3.1 for a detailed proof. □

Theorem 2.3.4 establishes an $O(K^{1/2})$ regret of Algorithm 1, where K is the total number of iterations. Here $O(\cdot)$ omits terms involving $(1-\gamma)^{-1}$ and $\log|\mathcal{A}|$. To better understand Theorem 2.3.4, we consider the ideal setting, where we have access to the action-value function Q^{π} of any policy π. In such an ideal setting, the critic update is unnecessary. However, the natural policy gradient method, which only uses the actor update, achieves the same $O(K^{1/2})$ regret (Liu et al., 2019; Agarwal et al., 2019; Cai et al., 2019). In other words, in terms of the iteration complexity, Theorem 2.3.4 shows that in the single-timescale setting, using only one step of the critic update along with one step of the actor update is as efficient as the natural policy gradient method in the ideal setting.

Furthermore, by the regret bound in (2.3.2), to obtain an ε-globally optimal policy, it suffices to set $K \asymp (1-\gamma)^{-6} \cdot \varepsilon^{-2} \cdot \log^2|\mathcal{A}|$ in Algorithm 1 and output a randomized policy that is drawn from $\{\pi_{\theta_k}\}_{k=1}^{K+1}$ uniformly. Plugging such a K into $N = \Omega(KC^2_{\rho,\rho^*}(\phi^*/\sigma^*)^2 \cdot \log^2(KN/\delta)))$, we obtain that $N = \widetilde{O}(\varepsilon^{-2})$, where $\widetilde{O}(\cdot)$ omits the logarithmic terms. Thus, to achieve an ε-globally optimal policy, the total sample complexity of Algorithm 1 is $\widetilde{O}(\varepsilon^{-4})$. This matches the sample complexity results established in Xu et al. (2020); Hong et al. (2020) for two-timescale actor-critic methods. Meanwhile, notice that here the critic updates are on-policy and we draw N new data points in each critic update. As discussed in §2.2.1, under the off-policy setting, the critic updates given in (2.2.7) can be implemented using a fixed dataset sampled from ρ_{bhv}, the stationary state-action distribution induced by the behavioral policy. Under this scenario, the total number of data points used by the algorithm is equal to N. Moreover, by imposing similar assumptions on ρ_{bhv} as in (i) of Assumption 2.3.1 and Assumption 2.3.3, we can establish a similar $O(K^{1/2})$ regret as in (2.3.2) for the off-policy setting. As a result, with data reuse, to obtain an ε-globally optimal policy, the sample complexity of Algorithm 1 is essentially $\widetilde{O}(\varepsilon^{-2})$, which demonstrates the advantage of our single-timescale actor-critic method. Besides, only focusing on the convergence to an ε-stationary point, Wu et al. (2020); Xu et al. (2020) establish the sample complexity of $\widetilde{O}(\varepsilon^{-5/2})$ for two-timescale actor-critic, where ε measures the squared Euclidean norm of the policy gradient. In contrast, by adopting the natural policy gradient (Kakade, 2002) in actor updates, we achieve convergence to the globally optimal policy. To the best of our knowledge, we establish the rate of convergence and global optimality of the actor-critic method with function approximation in the single-timescale setting for the first time.

Furthermore, as we will show in Theorem A.2.5 of §A.1, when both the actor and the critic are represented using overparameterized deep neural networks, we establish a similar $O((1-\gamma)^{-3} \cdot \log|\mathcal{A}| \cdot K^{1/2})$ regret when the architecture of the actor and critic neural networks are properly chosen. To our best knowledge, this seems the first theoretical guarantee for the actor-critic method with deep neural network function approximation in terms of the rate of convergence and global optimality.

2.4. Proof Sketch of Theorem 2.3.4

In this section, we sketch the proof of Theorem 2.3.4. Recall that ρ is a state-action distribution satisfying (ii) of Assumption 2.3.1. We first upper bound $\sum_{k=0}^{K}(Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a))$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ in part 1. Then by further taking the expectation over ρ in part 2, we conclude the proof of Theorem 2.3.4. See §A.3.1 for a detailed proof.

Part 1. In the sequel, we upper bound $\sum_{k=0}^{K}(Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a))$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$. We first decompose $Q^* - Q^{\pi_{\theta_{k+1}}}$ into the following three terms,

$$\sum_{k=0}^{K}[Q^* - Q^{\pi_{\theta_{k+1}}}](s,a) = \sum_{k=0}^{K}\big[(I - \gamma\mathbb{P}^{\pi^*})^{-1}(A_{1,k} + A_{2,k} + A_{3,k})\big](s,a), \tag{2.4.1}$$

the proof of which is deferred to (A.3.1) and (A.3.2) in §A.3.1 of the appendix. Here the operator $\mathbb{P}^{\pi^*}$ is defined in (2.1.3), $(I - \gamma\mathbb{P}^{\pi^*})^{-1} = \sum_{i=0}^{\infty}(\gamma\mathbb{P}^{\pi^*})^i$, and $A_{1,k}$, $A_{2,k}$, and $A_{3,k}$ are defined as follows,

$$A_{1,k}(s,a) = [\gamma(\mathbb{P}^{\pi^*} - \mathbb{P}^{\pi_{\theta_{k+1}}})Q_{\omega_k}](s,a), \tag{2.4.2}$$

$$A_{2,k}(s,a) = \big[\gamma\mathbb{P}^{\pi^*}(Q^{\pi_{\theta_{k+1}}} - Q_{\omega_k})\big](s,a), \tag{2.4.3}$$

$$A_{3,k}(s,a) = [\mathcal{T}^{\pi_{\theta_{k+1}}}Q_{\omega_k} - Q^{\pi_{\theta_{k+1}}}](s,a). \tag{2.4.4}$$

To understand the intuition behind $A_{1,k}$, $A_{2,k}$, and $A_{3,k}$, we interpret them as follows.

Interpretation of $A_{1,k}$. As defined in (2.4.2), $A_{1,k}$ arises from the actor update and measures the convergence of the policy $\pi_{\theta_{k+1}}$ towards a globally optimal policy π^*, which implies the convergence of $\mathbb{P}^{\pi_{\theta_{k+1}}}$ towards $\mathbb{P}^{\pi^*}$.

Interpretation of $A_{3,k}$. Note that by (2.1.2) and (2.1.4), we have $Q^{\pi_{\theta_{k+1}}} = \mathcal{T}^{\pi_{\theta_{k+1}}}Q^{\pi_{\theta_{k+1}}}$ and $\mathcal{T}^{\pi_{\theta_{k+1}}}$ is a γ-contraction, which implies that applying the Bellman evaluation operator $\mathcal{T}^{\pi_{\theta_{k+1}}}$ to any Q, e.g., Q_{ω_k}, infinite times yields $Q^{\pi_{\theta_{k+1}}}$. As defined in (2.4.4), $A_{3,k}$ measures the error of tracking the action-value function $Q^{\pi_{\theta_{k+1}}}$ of $\pi_{\theta_{k+1}}$ by applying the Bellman evaluation operator $\mathcal{T}^{\pi_{\theta_{k+1}}}$ to Q_{ω_k} only once, which arises from the critic update. Also, as $A_{3,k} = \mathcal{T}^{\pi_{\theta_{k+1}}}(Q_{\omega_k} - Q^{\pi_{\theta_{k+1}}})$, $A_{3,k}$ measures the difference between $Q^{\pi_{\theta_k}}$, which is approximated by Q_{ω_k} as discussed subsequently, and $Q^{\pi_{\theta_{k+1}}}$. Such a difference can also be viewed as the difference between π_{θ_k} and $\pi_{\theta_{k+1}}$, which arises from the actor update. Therefore, the convergence of $A_{3,k}$ to zero implies the contractions of not only the critic update but also the actor update, which illustrates the "double contraction" phenomenon. We establish the convergence of $A_{3,k}$ to zero in (2.4.10) subsequently.

Interpretation of $A_{2,k}$. Assuming that $A_{3,k-1}$ converges to zero, we have $\mathcal{T}^{\pi_{\theta_k}} Q_{\omega_{k-1}} \approx Q^{\pi_{\theta_k}}$. Moreover, assuming that the number of data points N is sufficiently large and ignoring the projection in (2.2.5), we have $\mathcal{T}^{\pi_{\theta_k}} Q_{\omega_{k-1}} = Q_{\widetilde{\omega}_k} \approx Q_{\omega_k}$ as $\widetilde{\omega}_k$ defined in (2.2.4) is an estimator of ω_k. Hence, we have $Q^{\pi_{\theta_k}} \approx Q_{\omega_k}$. Such an approximation error is characterized by ϵ_k^{c} defined in (2.4.5) subsequently. Hence, $A_{2,k}$ measures the difference between π_{θ_k} and $\pi_{\theta_{k+1}}$ through the difference between $Q^{\pi_{\theta_k}} \approx Q_{\omega_k}$ and $Q^{\pi_{\theta_{k+1}}}$, which relies on the convergence of $A_{3,k-1}$ to zero.

In the sequel, we upper bound $A_{1,k}$, $A_{2,k}$, and $A_{3,k}$, respectively. To establish such upper bounds, we define the following quantities,

$$\epsilon_{k+1}^{\mathrm{c}}(s,a) = [\mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k} - Q_{\omega_{k+1}}](s,a), \tag{2.4.5}$$

$$e_{k+1}(s,a) = [Q_{\omega_k} - \mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k}](s,a), \tag{2.4.6}$$

$$\vartheta_k(s) = \mathrm{KL}\big(\pi^*(\cdot \,|\, s) \,\|\, \pi_{\theta_k}(\cdot \,|\, s)\big) - \mathrm{KL}\big(\pi^*(\cdot \,|\, s) \,\|\, \pi_{\theta_{k+1}}(\cdot \,|\, s)\big). \tag{2.4.7}$$

To understand the intuition behind $\epsilon_{k+1}^{\mathrm{c}}$, e_{k+1}, and ϑ_k, we interpret them as follows.

Interpretation of $\epsilon_{k+1}^{\mathrm{c}}$. Recall that $\widetilde{\omega}_{k+1}$ is defined in (2.2.4), which parameterizes $\mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k}$ (ignoring the projection in (2.2.5)). Here $\epsilon_{k+1}^{\mathrm{c}}$ arises from approximating $\widetilde{\omega}_{k+1}$ using ω_{k+1} as an estimator, which is constructed based on ω_k and the N data points. In

$$\begin{tikzcd}
\text{Critic Update:} & & & Q_{\omega_k} \arrow[rrr] \arrow[drr, "e_{k+1}"] & & & Q_{\omega_{k+1}} \\
& & \mathbb{T}^{\pi_{\theta_k}} Q_{\omega_{k-1}} \arrow[ur, "\varepsilon_k^{\mathrm{c}}"] \arrow[dl, "A_{3,k-1}"'] \arrow[ddr, "A_{1,k}"] & & & \mathbb{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k} \arrow[ur, "\varepsilon_{k+1}^{\mathrm{c}}"] \arrow[dl, "A_{3,k}"'] \arrow[ddr, "A_{1,k+1}"] & \\
\text{Actor Update:} & Q^{\pi_{\theta_k}} \arrow[rrr, "\vartheta_k"] & & & Q^{\pi_{\theta_{k+1}}} & & \\
& & & \mathbb{T}^{\pi^*} Q_{\omega_k} & & & \mathbb{T}^{\pi^*} Q_{\omega_{k+1}} \\
& \mathbb{T}^{\pi^*} Q^{\pi_{\theta_{k-1}}} \arrow[urr, "A_{2,k-1}"] & & & \mathbb{T}^{\pi^*} Q^{\pi_{\theta_k}} \arrow[urr, "A_{2,k}"] & &
\end{tikzcd}$$

Figure 2.1. Illustration of the relationship among $A_{1,k}$, $A_{2,k}$, $A_{3,k}$, $\epsilon_{k+1}^{\mathrm{c}}$, e_{k+1}, and ϑ_k. Here $\{\theta_k, \omega_k\}$ and $\{\theta_{k+1}, \omega_{k+1}\}$ are two consecutive iterates of actor-critic. The red arrow from Q_{ω_k} to $Q_{\omega_{k+1}}$ represents the critic update and the red arrow from $Q^{\pi_{\theta_k}}$ to $Q^{\pi_{\theta_{k+1}}}$ represents the action-value functions associated with the two policies in any actor update. Here ϑ_k given in (2.4.7) quantifies the difference between π_{θ_k} and $\pi_{\theta_{k+1}}$ in terms of their KL distances to π^*. In addition, the cyan arrows represent quantities $A_{1,k}$, $A_{2,k}$, and $A_{3,k}$ introduced in (2.4.2)–(2.4.4), which are intermediate terms used for analyzing the error $Q^* - Q^{\pi_{k+1}}$. Finally, the blue arrows represent ε_{k+1}^{c} and e_{k+1} defined in (2.4.5) and (2.4.6), respectively. Here ε_{k+1}^{c} corresponds to the statistical error due to having finite data whereas e_{k+1} essentially quantifies the difference between π_{θ_k} and $\pi_{\theta_{k+1}}$.

particular, $\epsilon_{k+1}^{\mathrm{c}}$ decreases to zero as $N \to \infty$, which is used in characterizing $A_{2,k}$ defined in (2.4.3).

Interpretation of e_{k+1}. Assuming that $A_{3,k-1}$ defined in (2.4.4) and ϵ_k^{c} defined in (2.4.5) converge to zero, which implies $\mathcal{T}^{\pi_{\theta_k}} Q_{\omega_{k-1}} \approx Q^{\pi_{\theta_k}}$ and $\mathcal{T}^{\pi_{\theta_k}} Q_{\omega_{k-1}} \approx Q_{\omega_k}$, respectively, we have $Q_{\omega_k} \approx Q^{\pi_{\theta_k}}$. Therefore, as defined in (2.4.6), $e_{k+1} = Q_{\omega_k} - \mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k} \approx Q^{\pi_{\theta_k}} -$

$\mathcal{T}^{\pi_{\theta_{k+1}}} Q^{\pi_{\theta_k}} = (\mathcal{T}^{\pi_{\theta_k}} - \mathcal{T}^{\pi_{\theta_{k+1}}}) Q^{\pi_{\theta_k}}$ measures the difference between π_{θ_k} and $\pi_{\theta_{k+1}}$, which implies the difference between $\mathcal{T}^{\pi_{\theta_k}}$ and $\mathcal{T}^{\pi_{\theta_{k+1}}}$. We remark that e_{k+1} fully characterizes $A_{3,k}$ defined in (2.4.4) as shown in (2.4.8) subsequently.

Interpretation of ϑ_k. As defined in (2.4.7), ϑ_k measures the difference between π_{θ_k} and $\pi_{\theta_{k+1}}$ in terms of their differences with π^*, which are measured by the corresponding KL-divergences. In particular, ϑ_k is used in characterizing $A_{1,k}$ and $A_{2,k}$ defined in (2.4.2) and (2.4.3), respectively.

We remark that $\epsilon^{\mathrm{c}}_{k+1}$ measures the statistical error in the critic update, while ϑ_k measures the optimization error in the actor update. As discussed above, the convergence of $A_{3,k}$ to zero implies the contraction of both the actor update and the critic update, which illustrates the "double contraction" phenomenon. Meanwhile, since e_{k+1} fully characterizes $A_{3,k}$ as shown in (2.4.8) subsequently, e_{k+1} plays a key role in the "double contraction" phenomenon. In particular, the convergence of e_{k+1} to zero is established in (2.4.9) subsequently. See Figure 2.1 for an illustration of these quantities.

With the quantities defined in (2.4.5), (2.4.6), and (2.4.7), we upper bound $A_{1,k}$, $A_{2,k}$, and $A_{3,k}$ as follows,

$$\begin{aligned}
&A_{1,k}(s,a) \le \gamma\beta \cdot [\mathbb{P}\vartheta_k](s,a),\\
&A_{2,k}(s,a) \le \big[(\gamma\mathbb{P}^{\pi^*})^{k+1}(Q^* - Q_{\omega_0})\big](s,a) + \gamma\beta \cdot \sum_{i=0}^{k-1} \big[(\gamma\mathbb{P}^{\pi^*})^{k-i}\mathbb{P}\vartheta_i\big](s,a)\\
&\qquad\qquad + \sum_{i=0}^{k-1} \big[(\gamma\mathbb{P}^{\pi^*})^{k-i}\epsilon^{\mathrm{c}}_{i+1}\big](s,a),
\end{aligned}$$

$$A_{3,k}(s,a) = \big[\gamma\mathbb{P}^{\pi_{\theta_{k+1}}}(I - \gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1} e_{k+1}\big](s,a), \tag{2.4.8}$$

the proof of which is deferred to Lemmas A.3.1, A.3.2, and A.3.3 in §A.3.1 of the appendix, respectively. Meanwhile, by recursively expanding (2.4.5) and (2.4.6), we have

$$e_{k+1}(s,a) \le \biggl[\gamma^k \Bigl(\prod_{s=1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Bigr) e_1 + \sum_{i=1}^{k} \gamma^{k-i} \Bigl(\prod_{s=i+1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Bigr)(I - \gamma \mathbb{P}^{\pi_{\theta_i}})\epsilon_i^{\mathrm{c}}\biggr](s,a), \tag{2.4.9}$$

the proof of which is deferred to Lemma A.3.4 in §A.3.1 of the appendix. By plugging (2.4.9) into (2.4.8), we have

$$\begin{aligned} A_{3,k}(s,a) \le \biggl[\gamma \mathbb{P}^{\pi_{\theta_{k+1}}}(I - \gamma \mathbb{P}^{\pi_{\theta_{k+1}}})^{-1} \biggl(\gamma^k \Bigl(\prod_{s=1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Bigr) e_1 \\ + \sum_{i=1}^{k} \gamma^{k-i} \Bigl(\prod_{s=i+1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Bigr)(I - \gamma \mathbb{P}^{\pi_{\theta_i}})\epsilon_i^{\mathrm{c}}\biggr)\biggr](s,a). \end{aligned} \tag{2.4.10}$$

To better understand (2.4.10) and how it relates to the convergence of $A_{3,k}$, $A_{2,k}$, and $A_{1,k}$ to zero, we discuss in the following two steps.

Step (i). We assume $\epsilon_i^{\mathrm{c}} = 0$, which corresponds to the number of data points $N \to \infty$. Then (2.4.10) yields $A_{3,k} = O(\gamma^k)$, which implies that $A_{3,k}$ defined in (2.4.4) converges to zero driven by the discount factor γ. As discussed above, the convergence of $A_{3,k}$ to zero also implies the contraction between π_{θ_k} and $\pi_{\theta_{k+1}}$ of the actor update and the contraction between Q_{ω_k} and $Q^{\pi_{\theta_k}}$ of the critic update, which illustrates the "double contraction" phenomenon.

Step (ii). The convergence of $A_{3,k}$ to zero further ensures that $A_{2,k}$ converges to zero. To see this, we further assume $A_{3,k} = 0$, which together with the assumption that $\epsilon_{k+1}^{\mathrm{c}} = 0$

implies $Q^{\pi_{\theta_{k+1}}} = \mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k} = Q_{\omega_{k+1}}$ by their definitions in (2.4.4) and (2.4.5), respectively. Then by telescoping the sum of $A_{2,k}$ defined in (2.4.3), which cancels out $Q_{\omega_{k+1}}$ and $Q^{\pi_{\theta_{k+1}}}$, we obtain the convergence of $A_{2,k}$ to zero. Meanwhile, telescoping the sum of $A_{1,k}$ defined in (2.4.2) and the sum of its upper bound in (2.4.8) implies that $A_{1,k}$ converges to zero.

Now, by plugging (2.4.8) and (2.4.10) into (2.4.1), we establish an upper bound of $\sum_{k=0}^{K}(Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a))$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, which is deferred to (A.3.12) in §A.3.1 of the appendix. Hence, we conclude the proof in part 1. See part 1 of §A.3.1 for details.

Part 2. Recall that ρ is a state-action distribution satisfying (ii) of Assumption 2.3.1. In the sequel, we take the expectation over ρ in (A.3.12) and upper bound each term. We first introduce the following lemma, which upper bounds $\epsilon^{\mathrm{c}}_{k+1}$ defined in (2.4.5).

Lemma 2.4.1. Under Assumptions 2.3.2 and 2.3.3, with probability at least $1 - \delta$, it holds for any $k \in \{0, 1, \ldots, K\}$ that

$$\begin{aligned}\mathbb{E}_{\rho_{k+1}}\big[\epsilon^{\mathrm{c}}_{k+1}(s,a)^2\big] &= \mathbb{E}\big[\big(Q_{\omega_{k+1}}(s,a) - [\mathcal{T}^{\pi_{\theta_k}} Q_{\omega_k}](s,a)\big)^2\big] \\ &\le \frac{32(\mathcal{R}_{\max} + R)^2}{N(\sigma^*)^4} \cdot \log^2(NK/p + dK/p),\end{aligned}$$

where the expectation is taken with respect to $(s,a) \sim \rho_{k+1}$.

Proof. See §A.6.1 for a detailed proof. □

On the right-hand side of (A.3.12) in §A.3.1 of the appendix, for the terms not involving $\epsilon^{\mathrm{c}}_{k+1}$, i.e., M_1, M_2, and M_3 in (A.3.13), we take the expectation over ρ and establish their upper bounds in the ℓ_∞-norm over (s,a) in Lemma A.3.5. On the other hand, for the terms involving $\epsilon^{\mathrm{c}}_{k+1}$, i.e., M_4 and M_5 in (A.3.14), we take the expectation over ρ and then change the measure from ρ to ρ_{k+1}. By Assumption 2.3.1 and Lemma 2.4.1, which relies on ρ_{k+1}, we establish the upper bounds in Lemma A.3.6. See part 2 of §A.3.1 for details.

Combining Lemmas A.3.5 and A.3.6 yields Theorem 2.3.4. See §A.3.1 for a detailed proof.

2.5. Conclusion

In this paper, we analyze the actor-critic method in single-timescale setting with function approximation. We theoretically show that the method achieves an $O(K^{1/2})$-regret, which appears to be the first success to provide upper bound of regret of the actor-critic method in the single-timescale setting with function approximation. For future research, we aim to extend our analysis to other variations of actor-critic setting, including mean-field-type control tasks and games (Zhang et al., 2019b), risk-sensitive RL tasks (Prashanth and Ghavamzadeh, 2013), and partially observed MDP (Bhatnagar, 2010).

CHAPTER 3

Actor-Critic Provably Finds Nash Equilibria of Linear-Quadratic Mean-Field Games

3.1. Linear-Quadratic Mean-Field Game

A linear-quadratic mean-field N_{a}-player game involves $N_{\mathrm{a}} \in \mathbb{N}$ agents, whose state transitions are given by

$$x_{t+1}^i = Ax_t^i + Bu_t^i + \overline{A}\overline{x}_t + d^i + \omega_t^i, \qquad \forall t \geq 0,\ i \in [N_{\mathrm{a}}].$$

Here $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times k}$, and $\overline{A} \in \mathbb{R}^{m\times m}$ are matrices, $x_t^i \in \mathbb{R}^m$ and $u_t^i \in \mathbb{R}^k$ are the state and action vectors of agent i, respectively, the vector $d^i \in \mathbb{R}^m$ is a drift term, $\omega_t^i \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $\overline{x}_t = 1/N_{\mathrm{a}} \cdot \sum_{j=1}^{N_{\mathrm{a}}} x_t^j$ is the mean-field state. The agents are coupled through the mean-field state $\overline{x}_t$. In the linear-quadratic mean-field N_{a}-player game, the cost of agent $i \in [N_{\mathrm{a}}]$ at time $t \geq 0$ is given by

$$c_t^i = (x_t^i)^\top Q x_t^i + (u_t^i)^\top R u_t^i + \overline{x}_t^\top \overline{Q}\overline{x}_t,$$

where $Q \in \mathbb{R}^{m\times m}$, $R \in \mathbb{R}^{k\times k}$, and $\overline{Q} \in \mathbb{R}^{m\times m}$ are matrices, and u_t^i is generated by π^i, i.e., the policy of agent i. To measure the performance of agent i following its policy π^i under the influence of the other agents, we define the expected total cost of agent i as

$$J^i(\pi^1, \pi^2, \ldots, \pi^{N_\mathrm{a}}) = \lim_{T\to\infty} \mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T} c_t^i\right).$$

We are interested in finding a Nash equilibrium $(\pi^1, \pi^2, \ldots, \pi^{N_\mathrm{a}})$, which is defined by

$$J^i(\pi^1, \ldots, \pi^{i-1}, \pi^i, \pi^{i+1}, \ldots, \pi^{N_\mathrm{a}}) \leq J^i(\pi^1, \ldots, \pi^{i-1}, \widetilde{\pi}^i, \pi^{i+1}, \ldots, \pi^{N_\mathrm{a}}), \qquad \forall \widetilde{\pi}^i,\ i \in [N_\mathrm{a}].$$

That is, agent i cannot further decrease its expected total cost by unilaterally deviating from its Nash policy.

For the simplicity of discussion, we assume that the drift term d^i is identical for each agent. We consider taking the infinite-population limit $N_\mathrm{a} \to \infty$, where each agent has an infinitesimal contribution to the dynamics of the system. Thus, the joint policy of all the agents except agent i can be modeled as a mean-field policy $\pi^\dagger$, and all the agents following such a mean-field policy $\pi^\dagger$ generate the mean-field state $\mathbb{E}x_t^\dagger$, where $\{x_t^\dagger\}_{t\geq 0}$ is generated following the policy $\pi^\dagger$. By the symmetry of the agents in terms of their state transitions and cost functions, we focus on a fixed agent and drop the superscript i hereafter.

Before we formally present the formulation of linear-quadratic mean-field games, we first introduce the following mean-field LQR (MF-LQR) problem, which aims to find an optimal policy for the fixed agent given the mean-field policy $\pi^\dagger$.

Problem 3.1.1 (MF-LQR)**.** Given the mean-field policy $\pi^\dagger$, we consider the following formulation,

$$x_{t+1} = Ax_t + Bu_t + \overline{A}\mathbb{E}x_t^\dagger + d + \omega_t,$$
$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + (\mathbb{E}x_t^\dagger)^\top \overline{Q}(\mathbb{E}x_t^\dagger),$$
$$J(\pi, \pi^\dagger) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} c(x_t, u_t)\right],$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\{x_t^\dagger\}_{t\geq 0}$ is the trajectory generated by the policy $\pi^\dagger$, $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. Here the expectation $\mathbb{E}x_t^\dagger$ is taken across all the agents. We aim to find π^* such that $J(\pi^*, \pi^\dagger) = \inf_{\pi\in\Pi} J(\pi, \pi^\dagger)$.

Note that a controllable linear system using linear quadratic optimal control is always stable. Further combining the fact that our linear closed-loop dynamics in Problem 3.1.1 is driven by the Gaussian noise term ω_t, we know that the Markov chain of states generated by the policy $\pi^\dagger$ admits a stationary distribution and converges to this stationary distribution. This implies that the mean-field state $\mathbb{E}x_t^\dagger$ converges to a constant vector $\mu^\dagger$ as $t \to \infty$, which serves as a time-invariant mean-field state. As we consider the ergodic setting, it then suffices to study Problem 3.1.1 with t sufficiently large. Therefore, the influence of the mean-field policy $\pi^\dagger$ is captured by the mean-field state $\mu^\dagger$. By re-formulating Problem 3.1.1, with slight abuse of notations, we obtain the following drifted-LQR (D-LQR).

Problem 3.1.2 (D-LQR)**.** Given a mean-field state $\mu \in \mathbb{R}^m$, we consider the following formulation,

$$x_{t+1} = Ax_t + Bu_t + \overline{A}\mu + d + \omega_t,$$

$$c_\mu(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + \mu^\top \overline{Q}\mu,$$

$$J_\mu(\pi) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} c_\mu(x_t, u_t)\right],$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. We aim to find an optimal policy π_μ^* such that $J_\mu(\pi_\mu^*) = \inf_{\pi\in\Pi} J_\mu(\pi)$.

Compared with the most studied LQR problem (Lewis et al., 2012), both the state transition and the cost function in Problem 3.1.2 have drift terms, which act as the mean-field "force" that drives the states away from zero. Such a mean-field "force" introduces additional challenges when solving Problem 3.1.2 in the model-free setting (see §3.2.3 for details). On the other hand, the unique optimal policy π_μ^* of Problem 3.1.2 admits a linear form $\pi_\mu^*(x_t) = -K_{\pi_\mu^*} x_t + b_{\pi_\mu^*}$ under certain regularity conditions (Anderson and Moore, 2007), where the matrix $K_{\pi_\mu^*} \in \mathbb{R}^{k\times m}$ and the vector $b_{\pi_\mu^*} \in \mathbb{R}^k$ are the parameters of π_μ^*. Motivated by such a linear form of the optimal policy, we define the class of linear-Gaussian policies as

$$\Pi = \{\pi(x) = -K_\pi x + b_\pi + \sigma \cdot \eta \colon K_\pi \in \mathbb{R}^{k\times m}, b_\pi \in \mathbb{R}^k\}, \tag{3.1.1}$$

where $\sigma \in \mathbb{R}$ and the standard Gaussian noise term $\eta \in \mathbb{R}^k$ is included to encourage exploration. To solve Problem 3.1.2, it suffices to find the optimal policy π_μ^* within Π. We define $\Lambda_1(\mu) = \pi_\mu^*$ as the optimal policy under the mean-field state μ.

Assume that all the agents follow the linear policy $\pi(x) = -K_\pi x + b_\pi$ under the mean-field state μ. By plugging $u_t = \pi(x_t)$ into the state transition in Problem 3.1.2, as $t \to \infty$, we know that these agents generate a new mean-field state μ_{new} such that

$$\mu_{\text{new}} = (I - A + BK_\pi)^{-1}(Bb_\pi + \overline{A}\mu + d).$$

We define $\Lambda_2(\mu, \pi) = \mu_{\text{new}}$ as such a new mean-field state.

Now, we are ready to present the following linear-quadratic mean-field game (LQ-MFG).

Problem 3.1.3 (LQ-MFG)**.** We consider the following formulation,

$$\begin{aligned}
x_{t+1} &= Ax_t + Bu_t + \overline{A}\mu + d + \omega_t, \\
c(x_t, u_t) &= x_t^\top Q x_t + u_t^\top R u_t + \mu^\top \overline{Q}\mu, \\
J(\pi, \mu) &= \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} c(x_t, u_t)\right],
\end{aligned}$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\mu \in \mathbb{R}^m$ is the mean-field state, $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. We aim to find a pair (μ^*, π^*) such that (i) $J(\pi^*, \mu^*) = \inf_{\pi\in\Pi} J(\pi, \mu^*)$; (ii) $\mathbb{E}x_t^*$ converges to μ^* as $t \to \infty$, where $\{x_t^*\}_{t\geq 0}$ is the Markov chain of states generated by the policy π^*.

The formulation in Problem 3.1.3 is studied by Lasry and Lions (2007); Bensoussan et al. (2016); Saldi et al. (2018a,b). We propose a more general formulation in Problem B.3.2 (see §B.3 of the appendix for details), where an additional interaction term between the state vector x_t and the mean-field state μ is incorporated into the cost function. According to our analysis in §B.3, up to minor modification, the results in the following sections also carry over to Problem B.3.2. Therefore, for the sake of simplicity, we focus on Problem 3.1.3 in the sequel.

In Problem 3.1.3, condition (i) is equivalent to the optimality of the policy π^* under the mean-field state μ^*, namely, $\Lambda_1(\mu^*) = \pi^*$. Meanwhile, condition (ii) is equivalent to the invariance of the mean-field state μ^* given the policy π^*, namely, $\Lambda_2(\mu^*, \pi^*) = \mu^*$. Such equivalence follows from the NCE principle (Huang et al., 2006, 2007), which also motivates the following definition of the Nash equilibrium pair (Saldi et al., 2018a,b).

Definition 3.1.4 (Nash Equilibrium Pair)**.** The pair $(\mu^*, \pi^*) \in \mathbb{R}^m \times \Pi$ constitutes a Nash equilibrium pair of Problem 3.1.3 if it satisfies $\pi^* = \Lambda_1(\mu^*)$ and $\mu^* = \Lambda_2(\mu^*, \pi^*)$. Here μ^* is called the Nash mean-field state and π^* is called the Nash policy.

By Definition 3.1.4, Problem 3.1.3 aims to find a Nash equilibrium pair (μ^*, π^*).

Notations. We denote by $\|M\|_*$ the spectral norm, $\rho(M)$ the spectral radius, $\sigma_{\min}(M)$ the minimum singular value, and $\sigma_{\max}(M)$ the maximum singular value of a matrix M. We use $\|\alpha\|_2$ to represent the ℓ_2-norm of a vector α, and $(\alpha)_i^j$ to denote the sub-vector $(\alpha_i, \alpha_{i+1}, \ldots, \alpha_j)^\top$, where α_k is the k-th entry of the vector α. For scalars $a_1, \ldots, a_n$, we denote by $\mathrm{poly}(a_1, \ldots, a_n)$ the polynomial of $a_1, \ldots, a_n$, and this polynomial may vary from line to line. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$.

3.2. Mean-Field Actor-Critic

We first characterize the existence and uniqueness of the Nash equilibrium pair of Problem 3.1.3 under mild regularity conditions, and then propose a mean-field actor-critic algorithm to obtain such a Nash equilibrium. As a building block of the mean-field actor-critic, we propose the natural actor-critic to solve Problem 3.1.2.

3.2.1. Existence and Uniqueness of Nash Equilibrium Pair

We now establish the existence and uniqueness of the Nash equilibrium pair defined in Definition 3.1.4. We impose the following regularity conditions.

Assumption 3.2.1. We assume that the following statements hold:

(i) The algebraic Riccati equation $X = A^\top XA + Q - A^\top XB(B^\top XB + R)^{-1}B^\top XA$ admits a unique symmetric positive definite solution X^*;

(ii) It holds for $L_0 = L_1L_3 + L_2$ that $L_0 < 1$, where

$$L_1 = \big\| \big[(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top\big]^{-1}\overline{A}\big\|_2 \cdot \big\| \big[K^*Q^{-1}(I-A)^\top - R^{-1}B^\top\big]\big\|_2,$$

$$L_2 = \big[1 - \rho(A - BK^*)\big]^{-1} \cdot \|\overline{A}\|_2, \qquad L_3 = \big[1 - \rho(A - BK^*)\big]^{-1} \cdot \|B\|_2.$$

Here $K^* = -(B^\top X^*B + R)^{-1}B^\top X^*A$.

The first assumption is implied by mild regularity conditions on the matrices A, B, Q, and R, which are (1) the positivity of R; (2) the non-negativity of $Q = C^\top C$; (3) the observability of (A, C); (4) the stability of (A, B). See De Souza et al. (1986); Lewis et al. (2012) for more details. The second assumption is standard in the literature (Bensoussan et al., 2016; Saldi et al., 2018b), which ensures the stability of the LQ-MFG. In the following proposition, we show that Problem 3.1.3 admits a unique Nash equilibrium pair.

Proposition 3.2.2 (Existence and Uniqueness of Nash Equilibrium Pair)**.** Under Assumption 3.2.1, the operator $\Lambda(\cdot) = \Lambda_2(\cdot, \Lambda_1(\cdot))$ is L_0-Lipschitz, where L_0 is given in Assumption 3.2.1. Moreover, there exists a unique Nash equilibrium pair (μ^*, π^*) of Problem 3.1.3.

Proof. See §B.5.1 for a detailed proof. □

3.2.2. Mean-Field Actor-Critic for LQ-MFG

The NCE principle motivates a fixed-point approach to solve Problem 3.1.3, which generates a sequence of policies $\{\pi_s\}_{s\geq 0}$ and mean-field states $\{\mu_s\}_{s\geq 0}$ satisfying the following two properties: (i) Given the mean-field state μ_s, the policy π_s is optimal. (ii) The mean-field state becomes μ_{s+1} as $t \to \infty$, if all the agents follow π_s under the current mean-field state μ_s. Here (i) requires solving Problem 3.1.2 given the mean-field state μ_s, while (ii) requires simulating the agents following the policy π_s given the current mean-field μ_s. Based on such properties, we propose the mean-field actor-critic in Algorithm 2.

Algorithm 2 requires solving Problem 3.1.2 at each iteration to obtain $\pi_s = \Lambda_1(\mu_s)$ and $\mu_{s+1} = \Lambda_2(\mu_s, \pi_s)$. To this end, we introduce the natural actor-critic in §3.2.3 that solves Problem 3.1.2.

3.2.3. Natural Actor-Critic for D-LQR

Now we focus on solving Problem 3.1.2 for a fixed mean-field state μ, we thus drop the subscript μ hereafter. With slight abuse of notations, we write $\pi_{K,b}(x) = -Kx + b + \sigma \cdot \eta$

Algorithm 2 Mean-Field Actor-Critic for solving LQ-MFG.

```
Input:
    • Initial mean-field state $\mu_0$ and Initial policy $\pi_0$ with parameters $K_0$ and $b_0$.
    • Numbers of iterations $S$, $\{N_s\}_{s\in[S]}$, $\{H_s\}_{s\in[S]}$, $\{\widetilde{T}_{s,n}, T_{s,n}\}_{s\in[S],n\in[N_s]}$, $\{\widetilde{T}^b_{s,h}, T^b_{s,h}\}_{s\in[S],h\in[H_s]}$.
    • Stepsizes $\{\gamma_s\}_{s\in[S]}$, $\{\gamma^b_s\}_{s\in[S]}$, $\{\gamma_{s,n,t}\}_{s\in[S],n\in[N_s],t\in[T_{s,n}]}$, $\{\gamma^b_{s,h,t}\}_{s\in[S],h\in[H_s],t\in[T^b_{s,h}]}$.
for $s = 0, 1, 2, \ldots, S-1$ do
    Policy Update: Solve for the optimal policy $\pi_{s+1}$ with parameters $K_{s+1}$ and $b_{s+1}$ of Problem 3.1.2 via Algorithm 3 with $\mu_s$, $\pi_s$, $N_s$, $H_s$, $\{\widetilde{T}_{s,n}, T_{s,n}\}_{n\in[N_s]}$, $\{\widetilde{T}^b_{s,h}, T^b_{s,h}\}_{h\in[H_s]}$, $\gamma_s$, $\gamma^b_s$, $\{\gamma_{s,n,t}\}_{n\in[N_s],t\in[T_{s,n}]}$, and $\{\gamma^b_{s,h,t}\}_{h\in[H_s],t\in[T^b_{s,h}]}$, which gives the estimated mean-field state $\widehat{\mu}_{K_{s+1},b_{s+1}}$.
    Mean-Field State Update: Update the mean-field state via $\mu_{s+1} \leftarrow \widehat{\mu}_{K_{s+1},b_{s+1}}$.
end for
Output: Pair $(\pi_S, \mu_S)$.
```

to emphasize the dependence on K and b, and $J(K, b) = J(\pi_{K,b})$ consequently. Now, we propose the natural actor-critic to solve Problem 3.1.2.

For any policy $\pi_{K,b} \in \Pi$, by the state transition in Problem 3.1.2, we have

$$x_{t+1} = (A - BK)x_t + (Bb + \overline{A}\mu + d) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \Psi_\epsilon), \tag{3.2.1}$$

where $\Psi_\epsilon = \sigma BB^\top + \Psi_\omega$. It is known that if $\rho(A - BK) < 1$, then the Markov chain $\{x_t\}_{t\geq 0}$ induced by (3.2.1) has a unique stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$ (Anderson and Moore, 2007), where the mean-field state $\mu_{K,b}$ and the covariance Φ_K satisfy that

$$\mu_{K,b} = (I - A + BK)^{-1}(Bb + \overline{A}\mu + d), \tag{3.2.2}$$

$$\Phi_K = (A - BK)\Phi_K(A - BK)^\top + \Psi_\epsilon. \tag{3.2.3}$$

Meanwhile, the Bellman equation for Problem 3.1.2 takes the following form

$$P_K = (Q + K^\top RK) + (A - BK)^\top P_K (A - BK). \tag{3.2.4}$$

Then by calculation (see Proposition B.2.2 in §B.2.1 of the appendix for details), it holds that the expected total cost $J(K, b)$ is decomposed as

$$J(K, b) = J_1(K) + J_2(K, b) + \sigma^2 \cdot \operatorname{tr}(R) + \mu^\top \overline{Q} \mu, \tag{3.2.5}$$

where $J_1(K)$ and $J_2(K, b)$ are defined as

$$\begin{aligned} J_1(K) &= \operatorname{tr}\big[(Q + K^\top RK)\Phi_K\big] = \operatorname{tr}(P_K \Psi_\epsilon), \\ J_2(K, b) &= \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}^\top \begin{pmatrix} Q + K^\top RK & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}. \end{aligned} \tag{3.2.6}$$

Here $J_1(K)$ is the expected total cost in the most studied LQR problems (Yang et al., 2019b; Fazel et al., 2018), where the state transition does not have drift terms. Meanwhile, $J_2(K, b)$ corresponds to the expected cost induced by the drift terms. The following two propositions characterize the properties of $J_2(K, b)$.

First, we show that $J_2(K, b)$ is strongly convex in b.

Proposition 3.2.3. Given any K, the function $J_2(K, b)$ is ν_K-strongly convex in b. Here $\nu_K = \sigma_{\min}(Y_{1,K}^\top Y_{1,K} + Y_{2,K}^\top Y_{2,K})$, where $Y_{1,K} = R^{1/2} K (I - A + BK)^{-1} B - R^{1/2}$ and $Y_{2,K} = Q^{1/2}(I - A + BK)^{-1} B$. Also, $J_2(K, b)$ has ι_K-Lipschitz continuous gradient in b, where ι_K is upper bounded as $\iota_K \leq [1 - \rho(A - BK)]^{-2} \cdot (\|B\|_2^2 \cdot \|K\|_2^2 \cdot \|R\|_2 + \|B\|_2^2 \cdot \|Q\|_2)$.

Proof. See §B.5.4 for a detailed proof. □

Second, we show that $\min_b J_2(K, b)$ is independent of K.

Proposition 3.2.4. We define $b^K = \operatorname{argmin}_b J_2(K, b)$, where $J_2(K, b)$ is defined in (3.2.6). It holds that

$$b^K = \big[KQ^{-1}(I-A)^\top - R^{-1}B^\top\big] \cdot \big[(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top\big]^{-1} \cdot (\overline{A}\mu + d).$$

Moreover, $J_2(K, b^K)$ takes the form of

$$J_2(K, b^K) = (\overline{A}\mu + d)^\top \big[(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top\big]^{-1} \cdot (\overline{A}\mu + d),$$

which is independent of K.

Proof. See §B.5.2 for a detailed proof. □

Since $\min_b J_2(K, b)$ is independent of K by Proposition 3.2.4, it holds that the optimal K^* is the same as $\operatorname{argmin}_K J_1(K)$. This motivates us to minimize $J(K, b)$ by first updating K following the gradient direction $\nabla_K J_1(K)$ to the optimal K^*, then updating b following the gradient direction $\nabla_b J_2(K^*, b)$. We now design our algorithm based on this idea.

We define Υ_K, $p_{K,b}$, and $q_{K,b}$ as

$$\Upsilon_K = \begin{pmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{pmatrix} = \begin{pmatrix} \Upsilon_K^{11} & \Upsilon_K^{12} \\ \Upsilon_K^{21} & \Upsilon_K^{22} \end{pmatrix},$$

$$p_{K,b} = A^\top\big[P_K \cdot (\overline{A}\mu + d) + f_{K,b}\big], \qquad q_{K,b} = B^\top\big[P_K \cdot (\overline{A}\mu + d) + f_{K,b}\big], \tag{3.2.7}$$

where $f_{K,b} = (I - A + BK)^{-\top}[(A - BK)^\top P_K(Bb + \overline{A}\mu + d) - K^\top Rb]$. By calculation (see Proposition B.2.3 in §B.2.1 of the appendix for details), the gradients of $J_1(K)$ and

$J_2(K,b)$ take the forms of

$$\nabla_K J_1(K) = 2(\Upsilon_K^{22}K - \Upsilon_K^{21}) \cdot \Phi_K, \qquad \nabla_b J_2(K,b) = \Upsilon_K^{22}(-K\mu_{K,b} + b) + \Upsilon_K^{21}\mu_{K,b} + q_{K,b}.$$

Our algorithm follows the natural actor-critic method (Bhatnagar et al., 2009) and actor-critic method (Konda and Tsitsiklis, 2000). Specifically, (i) To obtain the optimal K^*, in the critic update step, we estimate the matrix Υ_K by $\widehat{\Upsilon}_K$ via a policy evaluation algorithm, e.g., Algorithm 7 or Algorithm 8 (see §B.2.2 and §B.2.3 of the appendix for details); in the actor update step, we update K via $K \leftarrow K - \gamma \cdot (\widehat{\Upsilon}_K^{22}K - \widehat{\Upsilon}_K^{21})$, where the term $\widehat{\Upsilon}_K^{22}K - \widehat{\Upsilon}_K^{21}$ is the estimated natural gradient. (ii) To obtain the optimal b^* given K^*, in the critic update step, we estimate Υ_{K^*}, $q_{K^*,b}$, and $\mu_{K^*,b}$ by $\widehat{\Upsilon}_{K^*}$, $\widehat{q}_{K^*,b}$, and $\widehat{\mu}_{K^*,b}$ via a policy evaluation algorithm; In the actor update step, we update b via $b \leftarrow b - \gamma \cdot \widehat{\nabla}_b J_2(K^*,b)$, where $\widehat{\nabla}_b J_2(K^*,b) = \widehat{\Upsilon}_{K^*}^{22}(-K^*\widehat{\mu}_{K^*,b} + b) + \widehat{\Upsilon}_{K^*}^{21}\widehat{\mu}_{K^*,b} + \widehat{q}_{K^*,b}$ is the estimated gradient. Combining the above procedure, we obtain the natural actor-critic for Problem 3.1.2, which is stated in Algorithm 3.

One may want to apply gradient method to $J(K,b)$ directly in the joint space of K and b. However, the gradient dominance property of $J_1(K)$ in the most studied LQR problem (Yang et al., 2019b) no longer holds for $J(K,b)$. Therefore, the convergence of the gradient method to $J(K,b)$ is not guaranteed in our problem.

3.3. Global Convergence Results

The following theorem establishes the rate of convergence of Algorithm 2 to the Nash equilibrium pair (μ^*, π^*) of Problem 3.1.3.

Algorithm 3 Natural Actor-Critic Algorithm for D-LQR.

```
Input:
    • Mean-field state $\mu$ and initial policy $\pi_{K_0,b_0}$.
    • Numbers of iterations $N$, $H$, $\{\widetilde{T}_n, T_n\}_{n\in[N]}$, $\{\widetilde{T}_h^b, T_h^b\}_{h\in[H]}$.
    • Stepsizes $\gamma$, $\gamma^b$, $\{\gamma_{n,t}\}_{n\in[N],t\in[T_n]}$, $\{\gamma_{h,t}^b\}_{h\in[H],t\in[T_h^b]}$.
for $n = 0, 1, 2, \ldots, N-1$ do
    Critic Update: Compute $\widehat{\Upsilon}_{K_n}$ via Algorithm 7 with $\pi_{K_n,b_0}$, $\mu$, $\widetilde{T}_n$,$T_n$, $\{\gamma_{n,t}\}_{t\in[T_n]}$, $K_0$, and $b_0$ as inputs.
    Actor Update: Update the parameter via
        $$K_{n+1} \leftarrow K_n - \gamma \cdot (\widehat{\Upsilon}_{K_n}^{22} K_n - \widehat{\Upsilon}_{K_n}^{21}).$$
end for
for $h = 0, 1, 2, \ldots, H-1$ do
    Critic Update: Compute $\widehat{\mu}_{K_N,b_h}$, $\widehat{\Upsilon}_{K_N}$, $\widehat{q}_{K_N,b_h}$ via Algorithm 7 with $\pi_{K_N,b_h}$, $\mu$, $\widetilde{T}_h^b$, $T_h^b$, $\{\gamma_{h,t}^b\}_{t\in[T_h^b]}$, $K_0$, and $b_0$.
    Actor Update: Update the parameter via
        $$b_{h+1} \leftarrow b_h - \gamma^b \cdot \big[\widehat{\Upsilon}_{K_N}^{22}(-K_N\widehat{\mu}_{K,b_h} + b_h) + \widehat{\Upsilon}_{K_N}^{21}\widehat{\mu}_{K_N,b_h} + \widehat{q}_{K_N,b_h}\big].$$
end for
Output: Policy $\pi_{K,b} = \pi_{K_N,b_H}$, estimated mean-field state $\widehat{\mu}_{K,b} = \widehat{\mu}_{K_N,b_H}$ .
```

Theorem 3.3.1 (Convergence of Algorithm 2)**.** For a sufficiently small tolerance $\varepsilon > 0$, we set the number of iterations S in Algorithm 2 such that

$$S > \frac{\log\big(\|\mu_0 - \mu^*\|_2 \cdot \varepsilon^{-1}\big)}{\log(1/L_0)}. \tag{3.3.1}$$

For any $s \in [S]$, we define

$$\begin{aligned}\varepsilon_s = \min\Big\{&\big[1-\rho(A-BK^*)\big]^4\big(\|B\|_2 + \|\overline{A}\|_2\big)^{-4}\big(\|\mu_s\|_2^{-2} + \|d\|_2^{-2}\big)\cdot\sigma_{\min}(\Psi_\epsilon)\cdot\sigma_{\min}(R)\cdot\varepsilon^2,\\ &\nu_{K^*}\cdot\big[1-\rho(A-BK^*)\big]^4\cdot\|B\|_2^{-2}\cdot M_b(\mu_s)\cdot\varepsilon^2,\ \varepsilon\Big\}\cdot 2^{-s-10},\end{aligned} \tag{3.3.2}$$

where ν_{K^*} is defined in Proposition 3.2.3 and

$$M_b(\mu_s) = 4\Big\| Q^{-1}(I-A)^\top \cdot \big[(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top\big]^{-1} \cdot (\overline{A}\mu_s + d)\Big\|_2$$

$$\cdot \big[\nu_{K^*}^{-1} + \sigma_{\min}^{-1}(\Psi_\epsilon) \cdot \sigma_{\min}^{-1}(R)\big]^{1/2}. \tag{3.3.3}$$

In the s-th policy update step in Line 3 of Algorithm 2, we set the inputs via Theorem B.2.4 such that $J_{\mu_s}(\pi_{s+1}) - J_{\mu_s}(\pi^*_{\mu_s}) < \varepsilon_s$, where the expected total cost $J_{\mu_s}(\cdot)$ is defined in Problem 3.1.2, and $\pi^*_{\mu_s} = \Lambda_1(\mu_s)$ is the optimal policy under the mean-field state μ_s. Then it holds with probability at least $1-\varepsilon^5$ that

$$\|\mu_S - \mu^*\|_2 \le \varepsilon, \qquad \|K_S - K^*\|_{\mathrm{F}} \le \varepsilon, \qquad \|b_S - b^*\|_2 \le (1+L_1)\cdot\varepsilon.$$

Here μ^* is the Nash mean-field state, K_S and b_S are parameters of the policy π_S, and K^* and b^* are parameters of the Nash policy π^*.

Proof Sketch. The proof of the theorem is based on the convergence of the natural actor-critic algorithm 3 and a contraction argument. First, we prove in Theorem B.2.4 that Algorithm 3 converges linearly to the optimal policy of Problem 3.1.2. By this, in each iteration of Algorithm 2, we control the error between μ_{s+1} and μ^*_{s+1} to be $\varepsilon_s > 0$ with high probability, where μ^*_{s+1} is the mean-field state generated by the optimal policy $\Lambda_1(\mu_s)$; in other words, $\mu^*_{s+1} = \Lambda(\mu_s)$. Combining the fact from Proposition 3.2.2 that $\Lambda(\cdot)$ is a contraction, we deduce that

$$\|\mu_{s+1} - \mu^*\|_2 \le \big\|\Lambda(\mu_s) - \Lambda(\mu^*)\big\|_2 + \widetilde{\varepsilon}_s \le L_0 \cdot \|\mu_s - \mu^*\|_2 + \widetilde{\varepsilon}_s$$

with high probability, where $\widetilde{\varepsilon}_s > 0$ is some error term and is specified in the detailed proof. Moreover, by telescoping sum, and note that the sum $\sum_{s=1}^{S} \widetilde{\varepsilon}_s$ is upper bounded by the desired error ε, we conclude the theorem. See §B.4.1 for a detailed proof. □

We highlight that if the inputs of Algorithm 2 satisfy the conditions stated in Theorem B.2.4, it holds that $J_{\mu_s}(\pi_{s+1}) - J_{\mu_s}(\pi^*_{\mu_s}) < \varepsilon_s$ for any $s \in [S]$. See Theorem B.2.4 in §B.2.1 of the appendix for details. By Theorem 3.3.1, Algorithm 2 converges linearly to the unique Nash equilibrium pair (μ^*, π^*) of Problem 3.1.3. To the best of our knowledge, this theorem is the first successful attempt to establish that reinforcement learning with function approximation finds the Nash equilibrium pairs in mean-field games with theoretical guarantee, which lays the theoretical foundations for applying modern reinforcement learning techniques to general mean-field games.

3.4. Conclusion

For the discrete-time linear-quadratic mean-field games, we provide sufficient conditions for the existence and uniqueness of the Nash equilibrium pair. Moreover, we propose the mean-field actor-critic algorithm with linear function approximation that is shown converges to the Nash equilibrium pair with linear rate of convergence. Our algorithm can be modified to use other parametrized function classes, including deep neural networks, for solving mean-field games. For future research, we aim to extend our algorithm to other variations of mean-field games including risk-sensitive mean-field games (Saldi et al., 2018a; Tembine et al., 2014), robust mean-field games (Bauso et al., 2016), and partially observed mean-field games (Saldi et al., 2019).

CHAPTER 4

Offline Reinforcement Learning with Instrumental Variables in Confounded Markov Decision Processes

4.1. Confounded Markov Decision Processes

In this section, we introduce the framework of confounded Markov decision processes with discrete instrumental variables. We aim to leverage the batch data to find an optimal in-class policy that maximizes the expected total rewards.

Confounded MDPs In a confounded MDP, we observe $\{S_t, A_t, R_t\}_{t\geq 0}$ for each trajectory, where S_t is the observed state, A_t is the action taken after observing S_t, and R_t is the immediate reward received after making an action A_t for $t \geq 0$. We denote by $\mathcal{S}$ and $\mathcal{A}$ the state and action spaces, respectively. Furthermore, we assume that at each decision point $t \geq 0$, there exist some unmeasured state variables $U_t \in \mathcal{U}$, which may confound the effect of action A_t on the rewards and future transitions. Due to such unobserved confounders, the (causal) effect of the action on the immediate and future rewards may not be non-parametrically identified and directly applying standard RL algorithms for MDPs will produce sub-optimal policies.

To address this concern, we study the confounded MDP via the instrumental variable (IV) method (Angrist and Imbens, 1995), which has been widely used in the literature of causal inference (e.g., Pearl (2009); Hernán and Robins (2010)) to identify the causal effect of a treatment under unmeasured confounding. Specifically, at each decision point

t, we further assume that we also observe a time-varying IV $Z_t \in \mathcal{Z}$, which is independent of U_t and does not have a direct effect on the immediate reward R_t and all future states, actions, and rewards. With such an IV, we observe $\{S_t, Z_t, A_t, R_t\}_{t \geq 0}$ for each trajectory in the confounded MDP.

In this work, we consider finite action and IV spaces, i.e., $\mathcal{A} = \{a_j\}_{j \in [K]}$ and $\mathcal{Z} = \{z_j\}_{j \in [K]}$, where $K \geq 2$ is an integer. Furthermore, we consider a simplex encoding for both actions and IVs which enjoy a nice interpretation (Zhang and Liu, 2014). Specifically, for any $j \in [K]$, we let

$$a_j = z_j = \begin{cases} (K-1)^{-1/2}\mathbf{1}_{K-1} & \text{if } j = 1, \\ \frac{(1+\sqrt{K}\mathbf{1}_{K-1})}{(K-1)^{3/2}} + \sqrt{\frac{K}{K-1}} e_{j-1} & \text{if } 2 \leq j \leq K, \end{cases} \tag{4.1.1}$$

where $\mathbf{1}_{K-1} \in \mathbb{R}^{K-1}$ is an all-one vector and $e_j \in \mathbb{R}^{K-1}$ is a vector with all elements 0 except 1 for j-th position. By the simplex encoding in (4.1.1), one can see that $\sum_{j \in [K]} a_j = \sum_{j \in [K]} z_j = 0$ and $a_i^\top a_j = z_i^\top z_j = -\mathbb{1}\{i \neq j\}/(K-1) + \mathbb{1}\{i = j\}$ for any $i, j \in [K]$, where $\mathbb{1}\{\cdot\}$ is an indicator function. We remark that any reasonable encoding mechanisms can be incorporated here and our results can be equally applied.

Value Function and Performance Metric. In the confounded MDP, we aim to find an optimal in-class policy $\pi^* \in \Pi$ such that π^* maximizes the expected total rewards, where Π is a class of time-homogeneous policies mapping from the observed state space $\mathcal{S}$ into the probability distribution over the action space $\mathcal{A}$. In particular, $\pi(a \mid s)$ refers to the probability of choosing action $a \in \mathcal{A}$ given the state value $s \in \mathcal{S}$. Formally, for any

$\pi \in \Pi$, we define the value function V^π and the expected total reward $J(\pi)$ as follows,

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_t \,\middle|\, S_0 = s\right], \qquad J(\pi) = (1-\gamma)\cdot \mathbb{E}_{S_0\sim\nu}\left[V^\pi(S_0)\right], \tag{4.1.2}$$

where the expectation $\mathbb{E}_\pi[\cdot]$ is taken with respect to the distribution such that the action $A_t \sim \pi(\cdot \,|\, S_t)$ for any $t \geq 0$, and ν is a known reference distribution over $\mathcal{S}$. Given the definition of $J(\pi)$ in (4.1.2), our goal is to leverage the batch data to estimate π^*, where

$$\pi^* \in \operatorname*{argmax}_{\pi\in\Pi} J(\pi).$$

Suppose the batch data we have collected consist of N independent and identically distributed copies of $\{S_t, Z_t, A_t, R_t\}_{t\geq 0}$ with total decision points T for each trajectory. Then we can summarize our batch data as $\mathcal{D} = \{\{S_t^i, Z_t^i, A_t^i, R_t^i, S_{t+1}^i\}_{t=0}^{T-1}\}_{i\in[N]}$. In the meanwhile, we define the performance metric as

$$\mathsf{SubOpt}(\pi) = J(\pi^*) - J(\pi),$$

which characterizes the suboptimality of the policy π compared with the optimal in-class policy π^*.

Why is Confounded MDP Challenging? In the standard MDP (Sutton and Barto, 2018), all states are assumed fully observed and the trajectory $\{S_t, A_t, R_t\}_{t\geq 0}$ satisfies the Markovian property. By leveraging the celebrated Bellman equation, under some mild conditions, one can non-parametrically identify $J(\pi)$ for any $\pi \in \Pi$, which serves as a foundation for many existing RL algorithms. However, in the confounded MDP, the Markovian assumption on the trajectory no longer holds and Bellman equation cannot be applied anymore due to the existence of the unmeasured confounders $\{U_t\}_{t\geq 0}$. More seriously, the effect of actions on the rewards and future states cannot be identified even we include all past history information at each decision point $t \geq 0$. Therefore, additional assumptions are needed to identify $J(\pi)$ and in this work, we rely on the IV to deal with such challenges.

Notation. Throughout the paper, we denote by c a positive absolute constant, which may vary from lines to lines. Without further explanation, we denote by $\mathbb{E}_\pi[\cdot]$ the expectation taken with respect to the trajectory generated by the policy π, $\mathbb{E}[\cdot]$ the expectation taken with respect to the trajectory generated by the behavior policy, and $\widehat{\mathbb{E}}[\cdot]$ the empirical average across all N trajectories.

4.2. Assumptions and Identification Results

In this section, we introduce several assumptions to help us identify $J(\pi)$ for any $\pi \in \Pi$ by using an IV. The first assumption is related to the trajectory $\{S_t, U_t, A_t, R_t\}_{t\geq 0}$, where we model it by a time-homogenous MDP.

Assumption 4.2.1. The following statements hold.

(a) For any $t \geq 1$, we have $(S_{t+1}, U_{t+1}) \perp\!\!\!\perp \{S_j, U_j, A_j\}_{0\leq j<t} \,|\, (S_t, U_t, A_t)$ and the transition probability is stationary;

(b) For any $t \geq 0$, we have $R_t = R(U_t, S_t, A_t, S_{t+1}, U_{t+1})$ for some deterministic function $R\colon \mathcal{U} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{U} \to \mathbb{R}$. Also, we assume $|R_t| \leq 1$ almost surely for any $t \geq 0$;

(c) The offline dataset $\mathcal{D}$ is generated by an unknown initial distribution ζ over $\mathcal{S}$ and a stationary policy b, which is a function mapping from $\mathcal{S} \times \mathcal{U} \times \mathcal{Z}$ into a probability distribution over $\mathcal{A}$.

Here b is often called behavior policy in RL literature. Assumption 4.2.1 is standard in the literature of RL, which is mild as $\{U_t\}_{t\geq 0}$ is unobserved. The uniformly bounded assumption on the reward R_t is used to simplify the technical analysis and can be relaxed by imposing some high-order moment condition on R_t instead. Due to the unobserved state variables U_t, we make the following IV assumptions.

Assumption 4.2.2. The following statements hold.

(a) For any $t \geq 0$, we have $(S_{t+1}, U_{t+1}) \perp\!\!\!\perp Z_t \,|\, (S_t, U_t, A_t)$;

(b) For any $a \in \mathcal{A}$ and $t \geq 0$, we have $\mathbb{P}(A_t = a \,|\, S_t, Z_t) \neq \mathbb{P}(A_t = a \,|\, S_t)$;

(c) For any $t \geq 0$, we have $Z_t \perp\!\!\!\perp U_t \,|\, S_t$ and the probability distribution of Z_t given S_t is time-homogeneous.

(d) For any $t \geq 0$, we have the behavior policy satisfy that

$$\begin{aligned} b(A_t = a \,|\, S_t, U_t, Z_t = a) &- \frac{1}{K-1} \sum_{z\in\mathcal{Z}, z\neq a} b(A_t = a \,|\, S_t, U_t, Z_t = z) \\ &= b(A_t = a \,|\, S_t, Z_t = a) - \frac{1}{K-1} \sum_{z\in\mathcal{Z}, z\neq a} b(A_t = a \,|\, S_t, Z_t = z) = \Delta^*(S_t, a), \end{aligned}$$

i.e., the compliance $\Delta^*(S_t, a)$ defined above is independent of the unobserved confounder U_t almost surely.

Assumption 4.2.2(a) states that there is not direct effect of the IV Z_t on the future states and rewards except through the action A_t, which is a typical assumption in the literature of causal inference with IVs (Angrist and Imbens, 1995; Angrist et al., 1996).

Note that by Assumption 4.2.1(b), we have implicitly restricted the effect of Z_t on the reward R_t only through A_t in this assumption. Assumption 4.2.2(b) requires that the IV Z_t will influence the action A_t, which is called IV relevance in the causal inference. Assumption 4.2.2(c), corresponding to IV independence, ensures that the effect of Z_t on futures states and rewards is unconfounded by adjusting the current state S_t. The homogeneous assumption on the conditional distribution of Z_t given S_t is imposed here as our target parameter $J(\pi)$ is defined over the infinite horizon. Define a function

$$\Theta^*(s, z) = \mathbb{P}(Z_t = z \,|\, S_t = s)$$

for every $(s, z) \in \mathcal{S} \times \mathcal{Z}$, which is independent of the decision point due to such time-homogeneity. In addition, Assumption 4.2.2(d) essentially indicates that there is no interaction between U_t and Z_t in affecting whether the action A_t will follow Z_t or not. This so-called independent compliance assumption has been widely adopted in identifying the average treatment effect of binary treatments in causal inference (Wang and Tchetgen Tchetgen, 2018). Here we generalize it to the setting of multiple treatments and instrumental variables, which may be of independent interest. A graphical illustration of Assumptions 4.2.1 and 4.2.2 is presented in Figure 4.1, which also illustrates how the offline data in the confounded MDP are generated. In the following section, we introduce value function (VF)-based identification and marginalized importance sampling (MIS)-based identification, respectively.

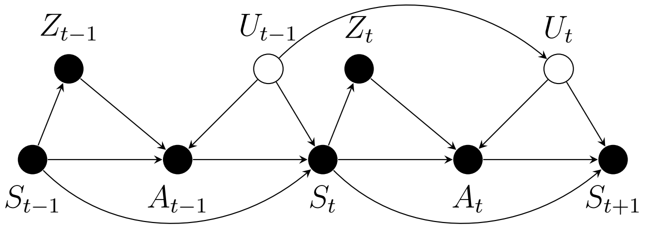

Figure 4.1. A graphical illustration of the confounded MDP.

4.2.1. Value Function-based Identification

In the unconfounded MDP, value function defined in (4.1.2) can be used to identify $J(\pi)$ and itself can be identified via the Bellman equation. However, due to the existence of unobserved confounders, the regular Bellman equation, which relies on the Markovian assumption, does not hold in general and the effect of actions on the reward cannot be identified either. Fortunately, by leveraging the IV, we are able to provide a way to identify $J(\pi)$ via the state-value function V^{π}, which can be identified by an IV-aided Bellman equation. Before stating our result, we make one additional assumption.

Assumption 4.2.3. We have $(Z_t, U_t) \perp\!\!\!\perp (\{S_j, U_j, A_j\}_{0\leq j<t}) \,|\, S_t$ and the probability distribution of (Z_t, U_t) given S_t is time-homogeneous for any $t \geq 0$.

Assumption 4.2.3 ensures that (Z_t, U_t) is “memoryless” and only depends on the current observed state S_t. This essentially ensures that the stochastic process $\{S_t, A_t\}_{t\geq 0}$ satisfies Markov property, which rules out the scenario that U_t affects future rewards and transitions. The memoryless assumption on the unmeasured confounders has been commonly used in the confounded MDP. See Kallus and Zhou (2020); Shi et al. (2022) for more details.

Lemma 4.2.4. Under Assumptions 4.2.1 and 4.2.2, for any $s \in \mathcal{S}$ and $\pi \in \Pi$, we have

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \left(\prod_{j=0}^{t} \frac{Z_j^\top A_j \pi(A_j \mid S_j)}{\Delta^*(S_j, A_j)\Theta^*(S_j, Z_j)}\right) \,\Bigg|\, S_0 = s\right].$$

If additionally Assumption 4.2.3 is satisfied, it holds for any $t \geq 0$ that,

$$V^{\pi}(s) = \mathbb{E}\left[\frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} \cdot (R_t + \gamma V^{\pi}(S_{t+1})) \,\Big|\, S_t = s\right].$$

Then the policy value $J(\pi)$ for $\pi \in \Pi$ can be identified via

$$J(\pi) = (1-\gamma) \cdot \mathbb{E}_{S_0 \sim \nu}\left[V^{\pi}(S_0)\right].$$

Proof. See §C.2.3 for a detailed proof. □

We remark that the Bellman equation in the unconfounded MDP takes the following form,

$$V^{\pi}_{\mathsf{unconf}}(s) = \mathbb{E}\left[\frac{\pi(A_t \mid S_t)}{\mathbb{P}(A_t \mid S_t)} \cdot (R_t + \gamma V^{\pi}_{\mathsf{unconf}}(S_{t+1})) \,\Big|\, S_t = s\right],$$

where $V^{\pi}_{\mathsf{unconf}}$ is the corresponding state-value function in the unconfounded MDP. In comparison, to deal with the unobserved confounders, our identification result in Lemma 4.2.4 incorporates the IVs into the action density ratio. It is also interesting to see that if one can observe the trajectory $\{S_t, Z_t, A_t, R_t\}_{t\geq 0}$ to the infinity, then Assumptions 4.2.1 and 4.2.2 are sufficient to identify $V^{\pi}(s)$ and $J(\pi)$ based on the first statement of Lemma 4.2.4. However, due to the limitation of only observing trajectories up to a finite horizon, we impose Assumption 4.2.3 so that Bellman equation is satisfied and used to break the curse of infinite-horizon. Based on Lemma 4.2.4, we introduce the following VF-based estimating equation, which will be used later in §4.3.1 to construct an estimator of the value function V^π.

Corollary 4.2.5 (VF-based Estimating Equation)**.** Under Assumptions 4.2.1, 4.2.2, and 4.2.3, it holds for any function $g\colon \mathcal{S} \to \mathbb{R}$ that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} \cdot (R_t + \gamma V^\pi(S_{t+1}))\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)V^\pi(S_t)\right].$$

Proof. See §C.2.4 for a detailed proof. □

4.2.2. Marginalized Importance Sampling-based Identification

In this subsection, we propose another way to identify $J(\pi)$ via the marginal importance sampling. We first introduce the following notations. For any $t \geq 0$, we denote by $p_t^\pi(\cdot)$ the marginal distribution of S_t under the known initial observed state distribution ν following the policy π. In the meanwhile, with a slight abuse of notations, we denote by $p_t^b(\cdot)$ the marginal distribution of S_t under the unknown offline data generation distribution ζ following the behavior policy b. In addition, we denote by for every $s \in \mathcal{S}$,

$$d^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_t^\pi(s), \qquad d^b(s) = \frac{1}{T}\sum_{t=0}^{T-1} p_t^b(s), \qquad w^\pi(s) = \frac{d^\pi(s)}{d^b(s)} \tag{4.2.1}$$

the discounted state visitation measure under the policy π, the average state visitation measure under the behavior policy b, and their density ratio, respectively. In order to identify $J(\pi)$ via the marginal importance sampling, i.e., w^π, we make the following assumption.

Assumption 4.2.6. For any $t \geq 0$, $U_t \mid S_t$ is time-homogeneous.

Assumption 4.2.6 is weaker than assuming the unmeasured confounder U_t is memoryless, and *does not* imply that $\{S_t, A_t\}_{t\geq 0}$ is a Markov chain. Such an assumption will be satisfied for example when the trajectory comes from a stationary sequence or unmeasured confounders are independent of observed states and actions. Motivated by the idea of marginalized importance sampling for off-policy evaluation in the standard MDP (Liu et al., 2018), we establish the following novel identification result for the expected total reward $J(\pi)$ in the confounded MDP.

Lemma 4.2.7. Under Assumptions 4.2.1, 4.2.2, and 4.2.6, for any $\pi \in \Pi$, we have

$$J(\pi) = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} \cdot w^\pi(S_t) R_t\right],$$

where w^π is defined in (4.2.1).

Proof. See §C.2.1 for a detailed proof. □

We remark that the expected total reward in the unconfounded MDP takes the following form,

$$J_{\mathsf{unconf}}(\pi) = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \frac{\pi(A_t \,|\, S_t)}{\mathbb{P}(A_t \,|\, S_t)} \cdot w^\pi(S_t) R_t\right],$$

where $J_{\mathsf{unconf}}(\pi)$ is the corresponding expected total reward in the unconfounded MDP, and the expectation $\mathbb{E}[\cdot]$ is taken with respect to the trajectory generated by the behavior policy. In comparison, to deal with the unobserved confounders, our identification result in Lemma 4.2.7 incorporates the IVs into the action density ratio. Based on Lemma 4.2.7, we introduce the following MIS-based estimating equation, which will be used later in §4.3.2 to construct an estimator of the density ratio w^π.

Lemma 4.2.8 (MIS-based Estimating Equation)**.** Under Assumptions 4.2.1, 4.2.2, and 4.2.6, for any $\pi \in \Pi$, it holds for any function $f : \mathcal{S} \to \mathbb{R}$ that

$$(1-\gamma)\mathbb{E}_{S_0 \sim \nu}\left[f(S_0)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} \cdot w^\pi(S_t)\left(f(S_t) - \gamma f(S_{t+1})\right)\right].$$

Proof. See §C.2.2 for a detailed proof. □

Lemma 4.2.8 is somewhat surprising in that with the help of IVs and Assumption 4.2.6, the estimating equation for the density ratio w^π holds even when the Markov condition fails on the observed trajectory, thus allowing the existence of unmeasured confounder U_t that can affect future rewards and transitions. This is different from the existing approaches such as Liu et al. (2018); Zhang et al. (2020) for estimating ratio functions in the standard MDP setting, which relies crucially on the Markovian assumption. Compared with the value function-based identification, marginalized importance sampling-based identification on the $J(\pi)$ requires fewer conditions and can be applied into more general confounded MDP problems.

4.3. Instrumental-Variable-Assisted RL with Pessimism

In this section, we introduce three pessimistic RL methods to estimate π^* in our confounded MDP. Generally, pessimistic RL first employs the offline data to construct a conservative estimate of the values for any policy, then select the policy with the highest conservative estimate of its value. Though the recently proposed pessimistic RL shows promising performance in practice (Kumar et al., 2020; Yu et al., 2020; Kidambi et al., 2020; Deng et al., 2021), its theoretical understanding is far from complete and is only limited to fully observable MDP (Levine et al., 2020). In this section, we adapt the idea of pessimism in our case and present related theoretical results in Section 4.4.

Since both identification results in §4.2.1 and §4.2.2 require estimating the quantities $\Delta^*(s,a)$ and $\Theta^*(s,z)$, we first introduce the estimating procedure for such quantities. We assume that there exists an oracle that gives estimators of $\Delta^*(s,a)$ and $\Theta^*(s,z)$ via two loss functions $\widehat{L}_0(\Delta)$ and $\widehat{L}_1(\Theta)$ as follows,

$$\widehat{\Delta} \in \operatorname*{argmin}_{\Delta \in \mathcal{F}_0} \widehat{L}_0(\Delta), \qquad \widehat{\Theta} \in \operatorname*{argmin}_{\Theta \in \mathcal{F}_1} \widehat{L}_1(\Theta),$$

where $\mathcal{F}_0$ and $\mathcal{F}_1$ are two function classes. We remark that we can use the negative likelihood functions for $\widehat{L}_0$ and $\widehat{L}_1$ (see §4.4 for details). In the meanwhile, we construct two confidence sets for Δ and Θ, respectively, as follows,

(4.3.1)

$$\mathsf{conf}^0_{\alpha_0} = \left\{\Delta \in \mathcal{F}_0 \colon \widehat{L}_0(\Delta) - \widehat{L}_0(\widehat{\Delta}) \le \alpha_0\right\}, \qquad \mathsf{conf}^1_{\alpha_1} = \left\{\Theta \in \mathcal{F}_1 \colon \widehat{L}_1(\Theta) - \widehat{L}_1(\widehat{\Theta}) \le \alpha_1\right\},$$

where (α_0, α_1) are some constants that will be specified later. These two confidence sets are used to construct conservative estimators for $J(\pi)$ via either VF-, MIS-, or doubly robust-based estimation.

4.3.1. VF-based Pessimistic Method

We introduce VF-based pessimistic RL in this section. We first define the following quantity,

$$\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta) = \widehat{\mathbb{E}}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta(S_t, A_t)\Theta(S_t, Z_t)}\left(R_t + \gamma v(S_{t+1})\right) - v(S_t)\right)\right],$$

where $\widehat{\mathbb{E}}[\cdot]$ is the empirical measure defined by the offline data $\mathcal{D}$. In the meanwhile, we define its population counterpart as

$$\Phi^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta) = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^{\top} A_t \pi(A_t \mid S_t)}{\Delta(S_t, A_t)\Theta(S_t, Z_t)}\left(R_t + \gamma v(S_{t+1})\right) - v(S_t)\right)\right]$$

for any (v, g, Δ, Θ), where the expectation $\mathbb{E}[\cdot]$ is taken with respect to the trajectory generated by the behavior policy. Then by the VF-based estimating equation specified in Corollary 4.2.5, it is easy to see that $\Phi^{\pi}_{\mathsf{vf}}(V^{\pi}, g; \Delta^*, \Theta^*) = 0$ for any function $g\colon \mathcal{S} \to \mathbb{R}$.

With the aforementioned notions, for any (Δ, Θ), we construct an estimator of V^{π} via solving the following minimax optimization problem,

$$\widehat{v}^{\pi}_{\Delta,\Theta} \in \operatorname*{argmin}_{v \in \mathcal{V}} \max_{g \in \mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta), \tag{4.3.2}$$

where $\mathcal{V}$ and $\mathcal{W}$ are two sets to be specified later. To obtain an estimation $\widehat{\pi}_{\mathsf{vf}}$ for an optimal in-class policy π^* that maximizes the expected total reward $J(\pi)$ defined in (4.1.2), we formulate the following optimization problem,

(4.3.3)

$$\widehat{\pi}_{\mathsf{vf}} = \operatorname*{argmax}_{\pi \in \Pi} \min_{(\Delta,\Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \min_{v \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi)} (1-\gamma)\mathbb{E}_{S\sim\nu}[v(S)],$$

(4.3.4)

$$\text{where } \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi) = \left\{ v \in \mathcal{V}\colon \max_{g \in \mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta) - \max_{g \in \mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{\Delta,\Theta}, g; \Delta, \Theta) \leq \alpha_{\mathsf{vf}} \right\},$$

where $(\alpha_0, \alpha_1, \alpha_{\mathsf{vf}})$ are constants to be specified, and $\mathsf{conf}^0_{\alpha_0}$ and $\mathsf{conf}^1_{\alpha_1}$ are confidence sets defined in (4.3.1). Intuitively, the policy $\widehat{\pi}_{\mathsf{vf}}$ defined in (4.3.3) aims to maximize the most pessimistic estimator of the expected total reward. As we will see in Theorem 4.4.9, such a pessimistic method provably converges to an optimal policy with data coverage assumption only for the optimal policy and some other mild conditions.

4.3.2. MIS-based Pessimistic Method

We introduce MIS-based pessimistic RL in this section. We first define the following quantity,

$$\begin{aligned}\widehat{\Phi}^{\pi}_{\text{mis}}(w, f; \Delta, \Theta) &= \mathbb{E}_{S_0\sim\nu}\left[(1-\gamma)f(S_0)\right] \\ &\quad - \widehat{\mathbb{E}}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^{\top}A_t\pi(A_t \mid S_t)}{\Delta(S_t, A_t)\Theta(S_t, Z_t)}w(S_t)\left(f(S_t)-\gamma f(S_{t+1})\right)\right].\end{aligned}$$

We define its population counterpart as

$$\begin{aligned}\Phi^{\pi}_{\text{mis}}(w, f; \Delta, \Theta) &= \mathbb{E}_{S_0\sim\nu}\left[(1-\gamma)f(S_0)\right] \\ &\quad - \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^{\top}A_t\pi(A_t \mid S_t)}{\Delta(S_t, A_t)\Theta(S_t, Z_t)}w(S_t)\left(f(S_t)-\gamma f(S_{t+1})\right)\right]\end{aligned}$$

for any (w, f, Δ, Θ). Then by the MIS-based estimating equation specified in Lemma 4.2.8, it can be seen that $\Phi^{\pi}_{\text{mis}}(w^{\pi}, f; \Delta^*, \Theta^*) = 0$ for any function $f\colon \mathcal{S} \to \mathbb{R}$. With the aforementioned notions, for any (Δ, Θ), we construct an estimator of w^{π} via solving the following minimax optimization problem,

$$\widehat{w}^{\pi}_{\Delta,\Theta} \in \operatorname*{argmin}_{w\in\mathcal{W}} \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\text{mis}}(w, f; \Delta, \Theta). \tag{4.3.5}$$

With a slight abuse of notations, here $\mathcal{W}$ and $\mathcal{V}$ are again two sets to be specified later. We aim to obtain an optimal policy that maximizes the expected total reward $J(\pi)$ by utilizing the estimators constructed in (4.3.5). For this, we further define the following estimator of $J(\pi)$ via Lemma 4.2.7,

$$\widehat{L}_{\mathrm{mis}}(w, \pi; \Delta, \Theta) = \widehat{\mathbb{E}}\left[\frac{1}{T}\sum_{t=0}^{T-1} \frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta(S_t, A_t)\Theta(S_t, Z_t)} w(S_t) R_t\right].$$

Then we aim to solve the following optimization problem,

$$\widehat{\pi}_{\mathrm{mis}} \in \operatorname*{argmax}_{\pi \in \Pi} \min_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta,\Theta,\pi)} \widehat{L}_{\mathrm{mis}}(w, \pi; \Delta, \Theta), \tag{4.3.6}$$

$$\text{where } \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi) = \Bigl\{ w \in \mathcal{W} \colon \max_{f \in \mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta) - \max_{f \in \mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{\Delta,\Theta}, f; \Delta, \Theta) < \alpha_{\mathrm{mis}} \Bigr\}, \tag{4.3.7}$$

where $(\alpha_0, \alpha_1, \alpha_{\mathrm{mis}})$ are constants to be specified, and $\mathsf{conf}^0_{\alpha_0}$ and $\mathsf{conf}^1_{\alpha_1}$ are confidence sets defined in (4.3.1). Similarly as in (4.3.3), the policy $\widehat{\pi}_{\mathrm{mis}}$ defined in (4.3.6) aims to maximize the most pessimistic estimator of the expected total reward. As we will see in Theorem 4.4.13, such a pessimistic method provably converges to an optimal policy with realizability assumption only for the optimal policy.

4.3.3. Doubly Robust based Pessimistic Method

As a combination of VF-based and MIS-based policy optimization methods, we introduce a doubly robust (DR)-based pessimistic RL algorithm in this section. We define the following DR estimator with its population counterpart,

$$\begin{aligned}
\widehat{L}_{\mathsf{dr}}(w,v,\pi;\Delta,\Theta) &= \widehat{\mathbb{E}}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)}{\Delta(S_t,A_t)\Theta(S_t,Z_t)}w(S_t)\left(R_t+\gamma v(S_{t+1})-v(S_t)\right)\right] \\
&\quad + (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right], \\
L_{\mathsf{dr}}(w,v,\pi;\Delta,\Theta) &= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)}{\Delta(S_t,A_t)\Theta(S_t,Z_t)}w(S_t)\left(R_t+\gamma v(S_{t+1})-v(S_t)\right)\right] \\
&\quad + (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right].
\end{aligned}$$

Note that $L_{\mathsf{dr}}(w^\pi,v,\pi;\Delta^*,\Theta^*) = L_{\mathsf{dr}}(w,V^\pi,\pi;\Delta^*,\Theta^*) = J(\pi)$ for any $(\pi,w,v)\in\Pi\times\mathcal{W}\times\mathcal{V}$. Thus, the quantity $\widehat{L}_{\mathsf{dr}}$ serves as a valid DR estimator of $J(\pi)$. In the follows, based on such a DR estimator of the expected total reward, we formulate the following optimization problem for estimating the optimal in-class policy,

$$\widehat{\pi}_{\mathsf{dr}} \in \operatorname*{argmax}_{\pi\in\Pi}\ \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\ \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi)} \widehat{L}_{\mathsf{dr}}(w,v,\pi;\Delta,\Theta),$$

$$\text{where } \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi) = \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta,\Theta,\pi)\times\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi), \tag{4.3.8}$$

where $(\alpha_0,\alpha_1,\alpha_{\mathsf{vf}},\alpha_{\mathrm{mis}})$ are constants to be specified, and $\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi)$ and $\mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta,\Theta,\pi)$ are defined in (4.3.4) and (4.3.7), respectively. As we will see in Theorem 4.4.14, such a DR-based pessimistic method provably converges to an optimal policy with realizability assumption only for the optimal policy.

4.4. Theoretical Results

In this section, we investigate theoretical properties of the aforementioned three methods. We aim to derive the finite-sample upper bounds for the sub-optimality of our estimated policies, i.e., $\mathsf{SubOpt}(\widehat{\pi})$, where $\widehat{\pi}$ is either $\widehat{\pi}_{\mathsf{vf}}, \widehat{\pi}_{\mathrm{mis}}$, or $\widehat{\pi}_{\mathsf{dr}}$. To begin with, we first introduce the following definition of covering number, and then impose some assumptions.

Definition 4.4.1 (Covering Number)**.** Let $(\mathcal{C}, \|\cdot\|_\infty)$ be a normed space, and $\mathcal{H} \subseteq \mathcal{C}$. The set $\{x_1, x_2, \ldots, x_n\}$ is a ε-covering over $\mathcal{H}$ if $\mathcal{H} \subseteq \cup_{i=1}^n B(x_i, \varepsilon)$, where $B(x_i, \varepsilon)$ is the L_∞-ball centered at x_i with radius ε. Then the covering number of $\mathcal{H}$ is defined as $N(\varepsilon, \mathcal{H}, \|\cdot\|_\infty) = \min\{n \colon \exists\ \varepsilon\text{-covering over } \mathcal{H} \text{ of size } n\}$.

Assumption 4.4.2. The following statements hold.

(a) For any set $\mathcal{H} \in \{\mathcal{F}_0, \mathcal{F}_1, \mathcal{V}, \mathcal{W}, \Pi\}$, there exists a constant $\mathfrak{C}_{\mathcal{H}}$ such that

$$N(\varepsilon, \mathcal{H}, \|\cdot\|_\infty) \leq c \cdot (1/\varepsilon)^{\mathfrak{C}_{\mathcal{H}}},$$

where $c > 0$ is a constant. Further, we denote by $\mathfrak{C}_{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_k} = \sum_{j \in [k]} \mathfrak{C}_{\mathcal{H}_j}$ for any $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_k\}$.

(b) There exist positive constants C_{Δ^*} and C_{Θ^*} such that $|\Delta^*(s,a)| \geq C_{\Delta^*}^{-1}$ and $\Theta^*(s,z) \geq C_{\Theta^*}^{-1}$ for any $(s,z,a) \in \mathcal{S} \times \mathcal{Z} \times \mathcal{A}$, where $\Theta^*(s,z)$ and $\Delta^*(s,a)$ are defined in Assumption 4.2.2.

(c) We have $|\Delta(s,a)| \geq C_{\Delta^*}^{-1}$ and $\Theta(s,z) \geq C_{\Theta^*}^{-1}$ for any $(\Delta, \Theta, s, a, z) \in \mathcal{F}_0 \times \mathcal{F}_1 \times \mathcal{S} \times \mathcal{A} \times \mathcal{Z}$.

(d) We have $\sup_{s \in \mathcal{S}} |V^{\pi_1}(s) - V^{\pi_2}(s)| \leq L_\Pi \cdot \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\pi_1(a \,|\, s) - \pi_2(a \,|\, s)|$ for any $\pi_1, \pi_2 \in \Pi$, where L_Π is a positive constant.

(e) We have $\|v\|_\infty \leq 1/(1-\gamma)$ and $\|w\|_\infty \leq C_*$ for any $(v, w) \in \mathcal{V} \times \mathcal{W}$, where $C_* > 0$ is a constant.

Assumption 4.4.2(a) states that the function spaces have finite log covering numbers. Assumption 4.4.2(b) states that the conditional probability Θ^* and the compliance Δ^* are uniformly lower bounded. Such an assumption ensures the identifiability of the expected total reward. With Assumption 4.4.2(b), we only need to consider a lower bounded function class to recover Θ^* and Δ^*, which is imposed in Assumption 4.4.2(c). In the meanwhile, the Lipschitz condition imposed in Assumption 4.4.2(d) aims to control the complexity of the value function class induced by Π, i.e., the class $\{V^\pi(\cdot) \colon \pi \in \Pi\}$. Such an assumption is commonly imposed in related literature (Zhou et al., 2017; Liao et al., 2020). Finally, Assumption 4.4.2(e) states that the sets $\mathcal{V}$ and $\mathcal{W}$ are upper bounded.

Assumption 4.4.3. The sequence $\{S_t, Z_t, U_t, A_t\}_{t\geq 0}$ admits a unique stationary distribution G_{stat} over $\mathcal{S} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{A}$ and is geometrically ergodic, i.e., there exists a function $\varphi \colon \mathcal{S} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{A} \to \mathbb{R}^+$ and a constant $\kappa > 0$ such that

$$\|G_{\text{stat}}(\cdot) - G_t(\cdot \,|\, s_0, z_0, u_0, a_0)\|_{\text{TV}} \leq \varphi(s_0, z_0, u_0, a_0) \cdot \exp\left(-2\kappa t\right),$$

where $G_t(\cdot \,|\, s_0, z_0, u_0, a_0)$ is the marginal distribution of (S_t, Z_t, U_t, A_t) given $(S_0, Z_0, U_0, A_0) = (s_0, z_0, u_0, a_0)$ under the behavior policy b. Further, we have $\int \varphi(s, z, u, a) \mathrm{d}\nu(s, z, u, a) \leq c$ and $\int \varphi(s, z, u, a) \mathrm{d}G_{\text{stat}}(s, z, u, a) \leq c$ for some positive absolute constant c.

Assumption 4.4.3 states that the the Markov chain $\{S_t, Z_t, U_t, A_t\}_{t\geq 0}$ mixes geometrically. Such an assumption is widely adopted in the related literature (Van Roy, 1998; Liao et al., 2020, 2021b; Wang et al., 2021a) to deal with dependent data.

To establish the upper bounds for the sub-optimality of the resulting policies, we need to first show that in our proposed algorithms, there exists at least one feasible solution that satisfies the constraints with properly chosen constants. In the following, we focus on $\mathsf{conf}^0_{\alpha_0}$ and $\mathsf{conf}^1_{\alpha_1}$ for Δ^* and Θ^* respectively. Since $\mathsf{conf}^0_{\alpha_0}$ and $\mathsf{conf}^1_{\alpha_1}$ can be constructed by many methods, to keep our theoretical results general, we assume that there exists a proper choice of (α_0, α_1) that ensures $\Delta^* \in \mathsf{conf}^0_{\alpha_0}$ and $\Theta^* \in \mathsf{conf}^1_{\alpha_1}$ and then give a valid example that justifies this assumption.

Assumption 4.4.4. There exists (α_0, α_1) such that with probability at least $1 - \delta$, we have

$$\Delta^* \in \mathsf{conf}^0_{\alpha_0}, \qquad \Theta^* \in \mathsf{conf}^1_{\alpha_1}.$$

Further, with probability at least $1 - \delta$, for any $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \|\Delta^*(S_t, \cdot) - \Delta(S_t, \cdot)\|_1^2\right] \le \xi_0^2 \frac{C_{\Delta^*}}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0} \log \frac{2}{\delta},$$
$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \|\Theta^*(S_t, \cdot) - \Theta(S_t, \cdot)\|_1^2\right] \le \xi_1^2 \frac{C_{\Theta^*}}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_1} \log \frac{2}{\delta}.$$

We now illustrate that Assumption 4.4.4 can be realized via maximum likelihood estimation (MLE) by replacing ξ_0 and ξ_1 with proper quantities. Note that the estimation of Δ^* can be decomposed into the estimation of $\mathbb{P}(A = a \,|\, S = s, Z = z)$ for all $z \in \mathcal{Z}$, which can also be obtained via MLE. This implies that estimating Δ^* is similar to estimating Θ^*. Therefore, we only show how to estimate Θ^* so that Assumption 4.4.4 holds for the simplicity of presentation. By maximum likelihood, we construct the loss function $\widehat{L}_1$ and the estimator $\widehat{\Theta}$ as follows,

$$\widehat{L}_1(\Theta) = -\widehat{\mathbb{E}}\left[\frac{1}{T}\sum_{t=0}^{T-1}\log\Theta(S_t, Z_t)\right] = -\frac{1}{NT}\sum_{i\in[N]}\sum_{t=0}^{T-1}\log\Theta(S_t^i, Z_t^i), \qquad \widehat{\Theta}\in\operatorname*{argmin}_{\Theta\in\mathcal{F}_1}\widehat{L}_1(\Theta),$$

where for the simplicity of notations, we denote by $\widehat{\mathbb{E}}[\cdot]$ the empirical measure generated by the offline data $\mathcal{D}$ hereafter. In addition, we assume that $\mathcal{F}_1$ is a parametric class such that $\mathcal{F}_1 = \{\Theta_\theta\colon \theta\in\mathbb{R}^d, \|\theta\|_2 \le \theta_{\max}\}$. We introduce the following results.

Theorem 4.4.5. Suppose $\mathcal{F}_1 = \{\Theta_\theta\colon \theta\in\mathbb{R}^d \text{ and } \|\theta\|_2 \le \theta_{\max}\}$, and

$$\alpha_1 = c\cdot\frac{C_{\Theta^*}}{NT\kappa}\cdot d\log\frac{\theta_{\max}}{\delta}\log(NT),$$

where $c/(N^2T^2)\cdot\log(NT) \le \delta \le 1$. Then under Assumptions 4.2.2, 4.4.2(b), 4.4.2(c), and 4.4.3, it holds with probability at least $1-\delta$ that $\Theta^*\in\mathsf{conf}^1_{\alpha_1}$. Further, with probability at least $1-\delta$, it holds for any $\Theta\in\mathsf{conf}^1_{\alpha_1}$ that

$$\sqrt{\mathbb{E}\left[\|\Theta(S,\cdot)-\Theta^*(S,\cdot)\|_1^2\right]} \le c\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\cdot d\log\frac{\theta_{\max}}{\delta}}.$$

Proof. See §C.1.1 for a detailed proof. □

Supported by Theorem 4.4.5, we assume Assumption 4.4.4 holds throughout this section.

4.4.1. Theoretical Results for VF-based Pessimistic Method

We first impose the following assumption, which assumes that V^π is realizable in $\mathcal{V}$ for any policy π, and w^{π^*} is realizable in $\mathcal{W}$ only for the optimal policy π^*.

Assumption 4.4.6. We have $V^\pi \in \mathcal{V}$ for any $\pi \in \Pi$ and $w^{\pi^*} \in \mathcal{W}$. Further, we have $-w \in \mathcal{W}$ for any $w \in \mathcal{W}$.

In the following lemma, we show that with a proper choice of α_{vf}, we have $V^\pi \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi)$ with a high probability.

Lemma 4.4.7. Suppose

$$\alpha_{\mathsf{vf}} = c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{W},\mathcal{V},\Pi}}{NT\kappa} \cdot \log\frac{1}{\delta} \log(NT)}$$

and $c/(NT)^2 \leq \delta \leq 1$. Then under Assumptions 4.4.2 and 4.4.6, with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ that $V^\pi \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi)$.

Proof. See §C.3.2 for a detailed proof. □

In the following lemma, we show that for any $v \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi)$, we can upper bound the risk $\max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*)$.

Lemma 4.4.8. Suppose that $(\alpha_0, \alpha_1, \alpha_{\mathsf{vf}})$ is defined in Assumption 4.4.4 and Lemma 4.4.7 and $c/(NT)^2 \leq \delta \leq 1$. Then under Assumptions 4.2.2, 4.2.3, and 4.4.2–4.4.6, with probability at least $1-\delta$, it holds for any policy $\pi \in \Pi$ and $v \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi)$ that

$$\max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) L_\Pi \sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log\frac{1}{\delta} \log(NT)}.$$

Proof. See §C.3.3 for a detailed proof. □

Equipped with the above results, we introduce the following theorem, which characterizes the suboptimality of the learned policy $\widehat{\pi}_{\mathsf{vf}}$ constructed in (4.3.3).

Theorem 4.4.9. Suppose $c/(NT)^2 \leq \delta \leq 1$. Under Assumptions 4.2.2–4.4.6, it holds with probability at least $1-\delta$ that

$$\mathsf{SubOpt}(\widehat{\pi}_{\mathsf{vf}}) \leq c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma}(\xi_0+\xi_1)L_\Pi \sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log\frac{1}{\delta}\log(NT)}.$$

Proof Sketch. In the proof sketch, we assume that we have full knowledge on Δ^* and Θ^*. By the definition of $J(\pi)$ in (4.1.2), we have

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}_{\mathsf{vf}}) &= (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0) - V^{\widehat{\pi}_{\mathsf{vf}}}(S_0)\right] \\ &\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right] - \min_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\widehat{\pi}_{\mathsf{vf}})}(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right] \\ &\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right] - \min_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)}(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right] \\ &\leq (1-\gamma)\cdot \max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)}\left|\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0) - v(S_0)\right]\right|, \end{aligned}$$

where in the first inequality, we use Lemma 4.4.7 that $V^{\widehat{\pi}_{\mathsf{vf}}} \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\widehat{\pi}_{\mathsf{vf}})$ with a high probability; while in the second inequality, we use the optimality of $\widehat{\pi}_{\mathsf{vf}}$. In the meanwhile, by Lemmas 4.2.7 and 4.2.8, we have the following decomposition,

$$(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right] = J(\pi^*) = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} w^{\pi^*}(S_t)\frac{Z_t^\top A_t \pi^*(A_t \mid S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}R_t\right],$$

(4.4.1)

$$(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} w^{\pi^*}(S_t)\frac{Z_t^\top A_t \pi^*(A_t \mid S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}\left(v(S_t) - \gamma v(S_{t+1})\right)\right].$$

Now, by plugging (4.4.1), we have

$$J(\pi^*) - J(\widehat{\pi}_{\mathsf{vf}}) \leq \max_{v \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi^*)} \left| \Phi^{\pi^*}_{\mathsf{vf}}(v, w^{\pi^*}; \Delta^*, \Theta^*) \right| .$$

We then can upper bound the above suboptimality by Lemma 4.4.8, which concludes the proof of the theorem. See §C.3.1 for a detailed proof. □

In Theorem 4.4.9, we impose data coverage and realizability assumptions as in Assumption 4.4.6, which only requires that the offline data covers the trajectory generated by the optimal policy π^* and V^π is realizable in $\mathcal{V}$ for any π.

4.4.2. Theoretical Results for MIS-based Pessimistic Method

We first impose the following assumption, which assumes that w^π is realizable in $\mathcal{W}$ for any policy π, and V^{π^*} is realizable in $\mathcal{V}$ only for the optimal policy π^*.

Assumption 4.4.10. We have $w^\pi \in \mathcal{W}$ for any $\pi \in \Pi$ and $V^{\pi^*} \in \mathcal{V}$. Further, we have $-v \in \mathcal{V}$ for any $v \in \mathcal{V}$.

In the following lemma, we show that with a proper choice of α_{mis}, we have $w^\pi \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi)$ with high probability.

Lemma 4.4.11. Suppose

$$\alpha_{\mathrm{mis}} = c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi} \log \frac{1}{\delta} \log(NT)}$$

and $c/(NT)^2 \leq \delta \leq 1$. Then under Assumptions 4.4.2 and 4.4.10, with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ that $w^\pi \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi)$.

Proof. See §C.4.2 for a detailed proof. □

In the following lemma, we show that for any $w \in \cup_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi)$, we can upper bound the risk $\max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*)$.

Lemma 4.4.12. Suppose that $(\alpha_0, \alpha_1, \alpha_{\mathrm{mis}})$ is defined in Assumption 4.4.4 and Lemma 4.4.11, and $c/(NT)^2 \leq \delta \leq 1$. Then under Assumptions 4.2.2, 4.2.6–4.4.4, and 4.4.10, with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ and $w \in \cup_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi)$ that

$$\max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log \frac{1}{\delta} \log(NT)}.$$

Proof. See §C.4.3 for a detailed proof. □

Equipped with the above results, we introduce the following theorem, which characterizes the suboptimality of the learned policy $\widehat{\pi}_{\mathrm{mis}}$ constructed in (4.3.6).

Theorem 4.4.13. Suppose $c/(NT)^2 \leq \delta \leq 1$. Under Assumptions 4.2.2, 4.2.6–4.4.4, and 4.4.10, it holds with probability at least $1-\delta$ that

$$\mathsf{SubOpt}(\widehat{\pi}_{\mathrm{mis}}) \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log \frac{1}{\delta} \log(NT)}.$$

Proof. See §C.4.1 for a detailed proof. □

In Theorem 4.4.13, we impose data coverage and realizability assumptions as in Assumption 4.4.10, which only requires that V^{π^*} is realizable in $\mathcal{V}$ and the offline data covers the trajectory generated by the policy π for any $\pi \in \Pi$.

4.4.3. Theoretical Results for DR-based Pessimistic Method

Theorem 4.4.14. Suppose that $(\alpha_0, \alpha_1, \alpha_{\text{mis}}, \alpha_{\mathsf{vf}})$ is defined in Assumption 4.4.4, Lemmas 4.4.7, 4.4.11, and at least one of Assumptions 4.4.6 and 4.4.10 hold. Then under Assumptions 4.2.2–4.4.4, it holds with probability at least $1-\delta$ for any $c/(NT)^2 \le \delta \le 1$ that

$$\mathsf{SubOpt}(\widehat{\pi}_{\mathsf{dr}}) \le c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{1}{\delta}\log(NT)},$$

where $\widehat{\pi}_{\mathsf{dr}}$ is defined in (4.3.8).

Proof. See §C.5.1 for a detailed proof. □

Theorem 4.4.14 shows that $\widehat{\pi}_{\mathsf{dr}}$ is a doubly robust estimator of the optimal policy in the sense that either Assumption 4.4.6 or Assumption 4.4.10 ensures the convergence of $\widehat{\pi}_{\mathsf{dr}}$.

Our results in §4.3.1–§4.3.3 hinge on the data coverage and realizability assumptions. Can we obtain a similar upper bound if such assumptions are violated? In this section, we answer this question affirmatively. First, we introduce the following assumption. We denote by

$$(4.4.2)\qquad \widetilde{v}^{\pi} \in \operatorname*{argmin}_{v\in\mathcal{V}}\max_{w\in\mathcal{W}}\Phi_{\mathsf{vf}}^{\pi}(v,w;\Delta^*,\Theta^*), \qquad \widetilde{w}^{\pi} \in \operatorname*{argmin}_{w\in\mathcal{W}}\max_{v\in\mathcal{V}}\Phi_{\text{mis}}^{\pi}(w,v;\Delta^*,\Theta^*).$$

Assumption 4.4.15 (Model Misspecification)**.** The following statements hold.

(a) We have $\|V^{\pi}-\widetilde{v}^{\pi}\|_\infty \le \varepsilon_{\mathsf{vf}}^{\mathcal{V}}$ for any $\pi\in\Pi$ and $\|w^{\pi^*}-\widetilde{w}^{\pi^*}\|_\infty \le \varepsilon_{\mathsf{vf}}^{\mathcal{W}}$.

(b) We have $\|w^{\pi}-\widetilde{w}^{\pi}\|_\infty \le \varepsilon_{\text{mis}}^{\mathcal{W}}$ for any $\pi\in\Pi$ and $\|V^{\pi^*}-\widetilde{v}^{\pi^*}\|_\infty \le \varepsilon_{\text{mis}}^{\mathcal{V}}$.

Though Assumption 4.4.15 requires that (a) and (b) hold simultaneously, we remark that previous assumptions imposed in VF-, MIS-, and DR-based pessimism can be recovered by such an assumption. Specifically, Assumptions 4.4.6 and 4.4.10 can be recovered by taking $(\varepsilon_{\mathsf{vf}}^{\mathcal{V}}, \varepsilon_{\mathsf{vf}}^{\mathcal{W}}, \varepsilon_{\mathrm{mis}}^{\mathcal{V}}, \varepsilon_{\mathrm{mis}}^{\mathcal{W}}) = (0, 0, \infty, \infty)$ and $(\varepsilon_{\mathsf{vf}}^{\mathcal{V}}, \varepsilon_{\mathsf{vf}}^{\mathcal{W}}, \varepsilon_{\mathrm{mis}}^{\mathcal{V}}, \varepsilon_{\mathrm{mis}}^{\mathcal{W}}) = (\infty, \infty, 0, 0)$, respectively, in Assumption 4.4.15. Similarly, the data coverage and realizability assumptions in Theorem 4.4.14 can also be recovered by either taking $(\varepsilon_{\mathsf{vf}}^{\mathcal{V}}, \varepsilon_{\mathsf{vf}}^{\mathcal{W}}, \varepsilon_{\mathrm{mis}}^{\mathcal{V}}, \varepsilon_{\mathrm{mis}}^{\mathcal{W}}) = (0, 0, \infty, \infty)$ or taking $(\varepsilon_{\mathsf{vf}}^{\mathcal{V}}, \varepsilon_{\mathsf{vf}}^{\mathcal{W}}, \varepsilon_{\mathrm{mis}}^{\mathcal{V}}, \varepsilon_{\mathrm{mis}}^{\mathcal{W}}) = (\infty, \infty, 0, 0)$.

Theorem 4.4.16. Suppose that $(\alpha_0, \alpha_1, \alpha_{\mathrm{mis}}, \alpha_{\mathsf{vf}})$ is defined in Assumption 4.4.4, Lemma 4.4.7, and Lemma 4.4.11. Then under Assumptions 4.2.2–4.4.4, and 4.4.15, it holds with probability at least $1 - \delta$ for any $c/(NT)^2 \leq \delta \leq 1$ that

$$\begin{aligned}\mathsf{SubOpt}(\widehat{\pi}_{\mathsf{dr}}) \leq c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{NT}{\delta}} \\ + 3C_{\Delta^*}C_{\Theta^*}\min\left\{C_*\varepsilon_{\mathsf{vf}}^{\mathcal{V}} + \varepsilon_{\mathsf{vf}}^{\mathcal{W}}/(1-\gamma),\ C_*\varepsilon_{\mathrm{mis}}^{\mathcal{V}} + \varepsilon_{\mathrm{mis}}^{\mathcal{W}}/(1-\gamma)\right\},\end{aligned}$$

where $\widehat{\pi}_{\mathsf{dr}}$ is defined in (4.3.8).

Proof. See §C.5.2 for a detailed proof. □

Due to the model misspecification, compared with Theorem 4.4.14, there is an additional bias term on the upper bound of the suboptimality in Theorem 4.4.16. We remark that either $(\varepsilon_{\mathsf{vf}}^{\mathcal{V}}, \varepsilon_{\mathsf{vf}}^{\mathcal{W}}) = (0, 0)$ or $(\varepsilon_{\mathrm{mis}}^{\mathcal{V}}, \varepsilon_{\mathrm{mis}}^{\mathcal{W}}) = (0, 0)$ ensures zero bias in Theorem 4.4.16.

4.5. Dual Formulation

To improve the computational efficiency of estimating the optimal in-class policy due to the confidence sets, we propose a dual formulation of the aforementioned pessimistic methods. For illustration purpose, we only consider the dual formulation of the VF-based pessimistic method proposed in §4.3.1. Similar formulations for MIS-based and DR-based methods can also be derived accordingly.

For the ease of presentation, we assume that there exists an oracle that gives us Δ^* and Θ^*. Without the existence of such an oracle, we only need to employ two additional dual variables to consider the uncertainty induced by estimating Δ^* and Θ^*.

We consider the following dual form of (4.3.3),

$$\begin{aligned} \widehat{\pi}^{\dagger}_{\mathsf{vf}} &= \operatorname*{argmax}_{\pi\in\Pi} \max_{\lambda\geq 0} \min_{v\in\mathcal{V}} \, (1-\gamma)\mathbb{E}_{S\sim\nu}[v(S)] + \lambda\cdot\Big(\widehat{M}^{\pi}_{\mathsf{vf}}(v) - \alpha_{\mathsf{vf}}\Big), \\ \text{s.t. } \widehat{M}^{\pi}_{\mathsf{vf}}(v) &= \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{\Delta^*,\Theta^*}, g; \Delta^*, \Theta^*), \end{aligned} \tag{4.5.1}$$

where $\widehat{v}^{\pi}_{\Delta^*,\Theta^*} = \operatorname{argmin}_{v\in\mathcal{V}} \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*)$ and λ is the dual variable that corresponds to the constraint $v \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi)$. In comparison to the constrained optimization problem in (4.3.3), the problem in (4.5.1) can be solved efficiently using gradient methods.

In the following theorem, we characterize the suboptimality of $\widehat{\pi}^{\dagger}_{\mathsf{vf}}$.

Theorem 4.5.1. Suppose that $\mathcal{V}$ is convex, $c/(NT)^2 \leq \delta \leq 1$, and α_{vf} is defined in Lemma 4.4.8. Under Assumptions 4.2.2–4.4.6, it holds with probability at least $1-\delta$ that

$$\mathsf{SubOpt}(\widehat{\pi}^{\dagger}_{\mathsf{vf}}) \leq c\cdot \frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)L_{\Pi}\sqrt{\frac{1}{NT\kappa}\cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\cdot \log\frac{1}{\delta}\log(NT)}.$$

Proof. See §C.6.1 for a detailed proof. □

In Theorem 4.5.1, we show a similar suboptimality holds as in Theorem 4.4.9. Thus, to avoid computation challenges induced by the confidence sets, we only need to solve

(4.5.1) to obtain an optimal policy. We remark similar dual formulations for MIS-based and DR-based methods also hold, as well as their theoretical properties.

4.6. Identifiability

We discuss identifiability in this section. First, we use tabular MDP as an example to illustrate non-identifiability. Then we discuss the identifiability in our methods.

Non-Identifiability in Tabular MDP. We consider a tabular MDP with states $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$, where the behavior policy b used to generate offline data only covers states $\{s_2, \ldots, s_{|\mathcal{S}|}\}$. We assume that the expected total reward under such a tabular MDP is $J(\pi)$ for any policy π. Since the offline data generated following b never covers the state s_1, we cannot infer any information of the reward received at the state s_1. Thus, for any policy π that arrives the state s_1 with a nonzero probability, we cannot identify the value $J(\pi)$ uniquely. In the meanwhile, the state-value function $V^\pi \colon \mathcal{S} \to \mathbb{R}$ is not uniquely identifiable for any policy π (even for π^*), since the value $V^\pi(s_1)$ is not identifiable.

Identifiability in Our Methods. In §4.2 and §4.3, we do not explicitly impose any identifiability assumptions. But certain identifiability assumptions are implied by the data coverage assumptions as follows.

- VF-based pessimism. As imposed in Assumption 4.4.6, we require that w^{π^*} is upper bounded. Thus, we know that the trajectory generated by the optimal policy π^* is covered by the offline data, which implies that $J(\pi^*)$ is identifiable.
- MIS-based pessimism. As imposed in Assumption 4.4.10, we require that w^π is upper bounded for any policy π. Thus, we know that the trajectory generated by

any policy π is covered by the offline data, which implies that $J(\pi)$ is identifiable for any π.

- DR-based pessimism. Since either Assumption 4.4.6 or Assumption 4.4.10 hold, we require that $J(\pi^*)$ is identifiable or $J(\pi)$ is identifiable for any π.

It is worth noting that though we impose realizability assumptions on the state-value function V^π in Assumptions 4.4.6 and 4.4.10, we do not require that V^π is identifiable for any π (even for π^*). Such non-identifiability implies that the IV-aided Bellman equation in Lemma 4.2.4 may have multiple fixed point solutions.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C. and Weisz, G. (2019a). Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*.

Abbasi-Yadkori, Y., Lazic, N., Szepesvari, C. and Weisz, G. (2019b). Exploration-enhanced politex. *arXiv preprint arXiv:1908.10479*.

Agarwal, A., Kakade, S. M., Lee, J. D. and Mahajan, G. (2019). Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*.

Agostinelli, F., McAleer, S., Shmakov, A. and Baldi, P. (2019). Solving the Rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence*, **1** 356–363.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R. et al. (2019). Solving Raubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*.

Alizadeh, F., Haeberly, J.-P. A. and Overton, M. L. (1998). Primal-dual interior-point methods for semidefinite programming: convergence rates, stability and numerical results. *SIAM Journal on Optimization*, **8** 746–768.

Allen-Zhu, Z., Li, Y. and Liang, Y. (2018a). Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*.

Allen-Zhu, Z., Li, Y. and Song, Z. (2018b). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, **10** 251–276.

Anderson, B. D. and Moore, J. B. (2007). *Optimal control: linear quadratic methods*. Courier Corporation.

Angrist, J. and Imbens, G. (1995). Identification and estimation of local average treatment effects.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, **91** 444–455.

Antos, A., Szepesvári, C. and Munos, R. (2007). Value-iteration based fitted policy iteration: learning with a single trajectory. In *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. IEEE.

Antos, A., Szepesvári, C. and Munos, R. (2008a). Fitted Q-iteration in continuous action-space mdps. In *Advances in neural information processing systems*.

Antos, A., Szepesvári, C. and Munos, R. (2008b). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, **71** 89–129.

Araki, B., Strang, J., Pohorecky, S., Qiu, C., Naegeli, T. and Rus, D. (2017). Multi-robot path planning for a swarm of robots that can both fly and drive. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

Arora, S., Du, S. S., Hu, W., Li, Z. and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*.

Ash, C. (2000). Social-self-interest. *Annals of public and cooperative economics*, **71** 261–284.

Axtell, R. L. (2002). Non-cooperative dynamics of multi-agent teams. In *Autonomous Agents and Multiagent Systems*.

Bardi, M. (2011). Explicit solutions of some linear-quadratic mean field games. *Networks and heterogeneous media*, **7** 243–261.

Bardi, M. and Priuli, F. S. (2014). Linear-quadratic n-person and mean-field games with ergodic cost. *SIAM Journal on Control and Optimization*, **52** 3022–3052.

Barrera, D. and Gobet, E. (2021). Generalization bounds for nonparametric regression with β-mixing samples. *arXiv preprint arXiv:2108.00997*.

Bauso, D., Tembine, H. and Başar, T. (2016). Robust mean field games. *Dynamic games and applications*, **6** 277–303.

Bellemare, M. G., Naddaf, Y., Veness, J. and Bowling, M. (2013). The Arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, **47** 253–279.

Bennett, A., Kallus, N., Li, L. and Mousavi, A. (2021). Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Bensoussan, A., Chau, M., Lai, Y. and Yam, S. C. P. (2017). Linear-quadratic mean field Stackelberg games with state and control delays. *SIAM Journal on Control and Optimization*, **55** 2748–2781.

Bensoussan, A., Frehse, J. and Yam, P. (2013). *Mean field games and mean field type control theory*. Springer.

Bensoussan, A., Sung, K., Yam, S. C. P. and Yung, S.-P. (2016). Linear-quadratic mean field games. *Journal of Optimization Theory and Applications*, **169** 496–529.

Berbee, H. C. P. (1979). Random walks with stationary increments and renewal theory.

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

Bhandari, J. and Russo, D. (2020). A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*.

Bhandari, J., Russo, D. and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*.

Bhatnagar, S. (2010). An actor-critic algorithm with function approximation for discounted cost constrained Markov Decision Processes. *Systems & Control Letters*, **59** 760–766.

Bhatnagar, S., Ghavamzadeh, M., Lee, M. and Sutton, R. S. (2008). Incremental natural actor-critic algorithms. In *Advances in Neural Information Processing Systems*.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M. and Lee, M. (2009). Natural actor–critic algorithms. *Automatica*, **45** 2471–2482.

Biswas, A. (2015). Mean field games with ergodic cost for discrete time markov processes. *arXiv preprint arXiv:1510.08968*.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint.* Cambridge University Press.

Borkar, V. S. and Konda, V. R. (1997). The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, **22** 525–543.

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, **38** 447–469.

Bowling, M. (2001). Rational and convergent learning in stochastic games. In *International Conference on Artificial Intelligence.*

Bowling, M. and Veloso, M. (2000). An analysis of stochastic game theory for multiagent reinforcement learning. Tech. rep., Carnegie Mellon University.

Bradtke, S. J. (1993). Reinforcement learning applied to linear quadratic regulation. In *Advances in Neural Information Processing Systems.*

Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine learning*, **22** 33–57.

Bradtke, S. J., Ydstie, B. E. and Barto, A. G. (1994). Adaptive linear quadratic control using policy iteration. In *American Control Conference*, vol. 3. IEEE.

Briani, A. and Cardaliaguet, P. (2018). Stable solutions in potential mean field game systems. *Nonlinear Differential Equations and Applications*, **25** 1.

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J. and Schneeweiss, S. (2010). Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, **48** S114.

Bu, J., Mesbahi, A., Fazel, M. and Mesbahi, M. (2019). LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921.*

Busoniu, L., Babuska, R. and De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, **38** 156–172.

Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2019). Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830* 1283–1294.

Caines, P. E. and Kizilkale, A. C. (2017). ϵ-nash equilibria for partially observed LQG mean field games with a major player. *IEEE Transactions on Automatic Control*, **62** 3225–3234.

Calderone, D. J. (2017). *Models of Competition for Intelligent Transportation Infrastructure: Parking, Ridesharing, and External Factors in Routing Decisions.* University of California, Berkeley.

Cao, Y. and Gu, Q. (2019a). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210.*

Cao, Y. and Gu, Q. (2019b). A generalization theory of gradient descent for learning over-parameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384.*

Carmona, R. and Delarue, F. (2013). Probabilistic analysis of mean-field games. *SIAM Journal on Control and Optimization*, **51** 2705–2734.

Carmona, R. and Delarue, F. (2018). *Probabilistic Theory of Mean Field Games with Applications I-II.* Springer.

Casgrain, P., Ning, B. and Jaimungal, S. (2019). Deep Q-learning for Nash equilibria: Nash-DQN. *arXiv preprint arXiv:1904.10554.*

Castro, D. D. and Meir, R. (2010). A convergent online single time scale actor critic algorithm. *Journal of Machine Learning Research*, **11** 367–410.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*.

Chizat, L. and Bach, F. (2018). A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*.

Conitzer, V. and Sandholm, T. (2007). AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, **67** 23–43.

Daniely, A. (2017). SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*.

de Cote, E. M., Lazaric, A. and Restelli, M. (2006). Learning to cooperate in multi-agent social dilemmas. In *International Conference on Autonomous Agents and Multiagent Systems*. ACM.

De Souza, C., Gevers, M. and Goodwin, G. (1986). Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices. *IEEE Transactions on Automatic Control*, **31** 831–838.

Dean, S., Mania, H., Matni, N., Recht, B. and Tu, S. (2017). On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*.

Dean, S., Mania, H., Matni, N., Recht, B. and Tu, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*.

Deng, Z., Fu, Z., Wang, L., Yang, Z., Bai, C., Wang, Z. and Jiang, J. (2021). Score: Spurious correlation reduction for offline reinforcement learning. *arXiv preprint arXiv:2110.12468*.

Doerr, B., Linares, R., Zhu, P. and Ferrari, S. (2018). Random finite set theory and optimal control for large spacecraft swarms. *arXiv preprint arXiv:1810.00696*.

Du, S. S., Chen, J., Li, L., Xiao, L. and Zhou, D. (2017). Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Du, S. S., Lee, J. D., Li, H., Wang, L. and Zhai, X. (2018). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*.

Ernst, D., Geurts, P. and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, **6** 503–556.

Fang, J. (2014). The LQR controller design of two-wheeled self-balancing robot based on the particle swarm optimization algorithm. *Mathematical Problems in Engineering*, **2014**.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C. and Mannor, S. (2016). Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, **17** 4809–4874.

Farahmand, A.-m., Szepesvári, C. and Munos, R. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*.

Fazel, M., Ge, R., Kakade, S. M. and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*.

Fu, Z., Yang, Z., Chen, Y. and Wang, Z. (2019). Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. *arXiv preprint arXiv:1910.07498*.

Ganzfried, S. and Sandholm, T. (2009). Computing equilibria in multiplayer stochastic games of imperfect information. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Gao, R., Cai, T., Li, H., Wang, L., Hsieh, C.-J. and Lee, J. D. (2019). Convergence of adversarial training in overparametrized networks. *arXiv preprint arXiv:1906.07916*.

Geer, S. A., van de Geer, S. and Williams, D. (2000). *Empirical Processes in M-estimation*, vol. 6. Cambridge university press.

Geist, M., Scherrer, B. and Pietquin, O. (2019). A theory of regularized markov decision processes. *arXiv preprint arXiv:1901.11275*.

Gomes, D. A., Mohr, J. and Souza, R. R. (2010). Discrete time, finite state space mean field games. *Journal de mathématiques pures et appliquées*, **93** 308–328.

Gomes, D. A. et al. (2014). Mean field games models—a brief survey. *Dynamic Games and Applications*, **4** 110–154.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F. and Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature medicine*, **25** 16–18.

Guéant, O., Lasry, J.-M. and Lions, P.-L. (2011). Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*. Springer, 205–266.

Guo, X., Hu, A., Xu, R. and Zhang, J. (2019). Learning mean-field games. *arXiv preprint arXiv:1901.09585*.

Haarnoja, T., Zhou, A., Abbeel, P. and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.

Hao, B., Lazic, N., Abbasi-Yadkori, Y., Joulani, P. and Szepesvari, C. (2020). Provably efficient adaptive approximate policy iteration. *arXiv preprint arXiv:2002.03069*.

Hardt, M., Ma, T. and Recht, B. (2016). Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*.

Heinrich, J. and Silver, D. (2016). Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*.

Hernán, M. A. and Robins, J. M. (2010). Causal inference: What if.

Hong, M., Wai, H.-T., Wang, Z. and Yang, Z. (2020). A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*.

Hu, J. and Wellman, M. P. (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. In *International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc.

Hu, J. and Wellman, M. P. (2003). Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, **4** 1039–1069.

Huang, J. and Huang, M. (2017). Robust mean field linear-quadratic-gaussian games with unknown L^2-disturbance. *SIAM Journal on Control and Optimization*, **55** 2811–2840.

Huang, J. and Li, N. (2018). Linear–quadratic mean-field game for stochastic delayed systems. *IEEE Transactions on Automatic Control*, **63** 2722–2729.

Huang, J., Li, X. and Wang, T. (2016a). Mean-field linear-quadratic-Gaussian (LQG) games for stochastic integral systems. *IEEE Transactions on Automatic Control*, **61** 2670–2675.

Huang, J., Wang, S. and Wu, Z. (2016b). Backward mean-field linear-quadratic-Gaussian (LQG) games: Full and partial information. *IEEE Transactions on Automatic Control*, **61** 3784–3796.

Huang, M., Caines, P. E. and Malhamé, R. P. (2003). Individual and mass behaviour in large population stochastic wireless power control problems: centralized and nash equilibrium solutions. In *Conference on Decision and Control*. IEEE.

Huang, M., Caines, P. E. and Malhamé, R. P. (2007). Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ε-Nash equilibria. *IEEE transactions on automatic control*, **52** 1560–1571.

Huang, M., Malhamé, R. P., Caines, P. E. et al. (2006). Large population stochastic dynamic games: closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, **6** 221–252.

Huang, M. and Zhou, M. (2019). Linear quadratic mean field games: Asymptotic solvability and relation to the fixed point approach. *arXiv preprint arXiv:1903.08776*.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K., Koster, R. et al. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

Jayakumar, S. and Aditya, M. (2019). Reinforcement learning in stationary mean-field games. In *International Conference on Autonomous Agents and Multiagent Systems*.

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, **33** 2747–2758.

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

Kakade, S. M. (2002). A natural policy gradient. In *Advances in neural information processing systems*.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V. et al. (2018). Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*. PMLR.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *J. Mach. Learn. Res.*, **21** 1–63.

Kallus, N. and Uehara, M. (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*.

Kallus, N. and Zhou, A. (2020). Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, **33** 22293–22304.

Kidambi, R., Rajeswaran, A., Netrapalli, P. and Joachims, T. (2020). MOReL: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, **24** 1716–1720.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*.

Korda, N. and La, P. (2015). On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International Conference on Machine Learning*.

Kumar, A., Zhou, A., Tucker, G. and Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, vol. 35. Springer Science & Business Media.

Lagoudakis, M. G. and Parr, R. (2002). Value function approximation in zero-sum Markov games. In *Uncertainty in Artificial Intelligence*.

Lasry, J.-M. and Lions, P.-L. (2006a). Jeux à champ moyen. I–le cas stationnaire. *Comptes Rendus Mathématique*, **343** 619–625.

Lasry, J.-M. and Lions, P.-L. (2006b). Jeux à champ moyen. II–horizon fini et contrôle optimal. *Comptes Rendus Mathématique*, **343** 679–684.

Lasry, J.-M. and Lions, P.-L. (2007). Mean field games. *Japanese journal of mathematics*, **2** 229–260.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521** 436–444.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J. and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J. and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems.

Levine, S., Kumar, A., Tucker, G. and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Lewis, F. L., Vrabie, D. and Syrmos, V. L. (2012). *Optimal control*. John Wiley & Sons.

Li, S., Zhang, W. and Zhao, L. (2017). Connections between mean-field game and social welfare optimization. *arXiv preprint arXiv:1703.10211*.

Li, T. and Zhang, J.-F. (2008). Asymptotically optimal decentralized control for large population stochastic multiagent systems. *IEEE Transactions on Automatic Control*, **53** 1643–1660.

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Li, Y. (2018). Deep reinforcement learning. *arXiv preprint arXiv:1810.06339*.

Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*.

Liao, L., Fu, Z., Yang, Z., Wang, Y., Kolar, M. and Wang, Z. (2021a). Instrumental variable value iteration for causal offline reinforcement learning. *arXiv preprint* *arXiv:2102.09907*.

Liao, P., Klasnja, P. and Murphy, S. (2021b). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, **116** 382–391.

Liao, P., Qi, Z., Klasnja, P. and Murphy, S. (2020). Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*. Elsevier, 157–163.

Littman, M. L. (2001). Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Liu, B., Cai, Q., Yang, Z. and Wang, Z. (2019). Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306* 10564–10575.

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S. and Petrik, M. (2015). Finite-sample analysis of proximal gradient TD algorithms. In *Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Liu, Q., Li, L., Tang, Z. and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, vol. 31.

Luenberger, D. G. (1997). *Optimization by vector space methods*. John Wiley & Sons.

Maei, H. R. (2018). Convergent actor-critic algorithms under off-policy training and function approximation. *arXiv preprint arXiv:1802.07842*.

Magnus, J. R. (1979). The expectation of products of quadratic forms in normal variables: the practice. *Statistica Neerlandica*, **33** 131–136.

Magnus, J. R. et al. (1978). *The moments of products of quadratic forms in normal variables*. Univ., Instituut voor Actuariaat en Econometrie.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P. L. and Wainwright, M. J. (2018). Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *arXiv preprint arXiv:1812.08305*.

Mei, J., Xiao, C., Szepesvari, C. and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392*.

Meitz, M. and Saikkonen, P. (2021). Subgeometric ergodicity and β-mixing. *Journal of Applied Probability*, **58** 594–608.

Mguni, D., Jennings, J. and de Cote, E. M. (2018). Decentralised learning in systems with many, many strategic agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Minciardi, R. and Sacile, R. (2011). Optimal control in a cooperative network of smart power grids. *IEEE Systems Journal*, **6** 126–133.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*.

Moon, J. and Başar, T. (2014). Discrete-time LQG mean field games with unreliable communication. In *Conference on Decision and Control*. IEEE.

Moon, J. and Başar, T. (2018). Linear quadratic mean field stackelberg differential games. *Automatica*, **97** 200–213.

Moravčík, M., Schmid, M., Burch, N., Lisỳ, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M. and Bowling, M. (2017). Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, **356** 508–513.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, **9** 815–857.

Nachum, O., Chow, Y., Dai, B. and Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, vol. 32.

Nachum, O. and Dai, B. (2020). Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*.

Namkoong, H., Keramati, R., Yadlowsky, S. and Brunskill, E. (2020). Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, **33** 18819–18831.

Nash, J. (1951). Non-cooperative games. *Annals of mathematics* 286–295.

OpenAI (2018). Openai five. `https://blog.openai.com/openai-five/`.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pérolat, J., Piot, B., Geist, M., Scherrer, B. and Pietquin, O. (2016a). Softened approximate policy iteration for markov games. In *International Conference on Machine Learning*.

Pérolat, J., Piot, B. and Pietquin, O. (2018). Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*.

Pérolat, J., Piot, B., Scherrer, B. and Pietquin, O. (2016b). On the use of non-stationary strategies for solving two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*.

Perolat, J., Scherrer, B., Piot, B. and Pietquin, O. (2015). Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning (ICML 2015)*.

Peters, J., Janzing, D. and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

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

Peters, J. and Schaal, S. (2008b). Reinforcement learning of motor skills with policy gradients. *Neural Networks*, **21** 682–697.

Prashanth, L. and Ghavamzadeh, M. (2013). Actor-critic algorithms for risk-sensitive mdps. In *Advances in neural information processing systems*.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series* 80.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P. and Ghassemi, M. (2017). Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*. PMLR.

Rudelson, M., Vershynin, R. et al. (2013). Hanson-wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, **18** 82–90.

Saldi, N., Basar, T. and Raginsky, M. (2018a). Discrete-time risk-sensitive mean-field games. *arXiv preprint arXiv:1808.03929*.

Saldi, N., Basar, T. and Raginsky, M. (2018b). Markov–Nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, **56** 4256–4287.

Saldi, N., Basar, T. and Raginsky, M. (2019). Approximate Nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research*.

Sallab, A. E., Abdou, M., Perot, E. and Yogamani, S. (2017). Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, **2017** 70–76.

Sandholm, W. H. (2010). *Population Games and Evolutionary Dynamics*. MIT Press.

Scherrer, B. (2013). On the performance bounds of some policy search dynamic programming algorithms. *arXiv preprint arXiv:1306.0539*.

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B. and Geist, M. (2015). Approximate modified policy iteration and its application to the game of tetris. *Journal of Machine Learning Research*, **16** 1629–1676.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. and Moritz, P. (2015). Trust region policy optimization. In *International Conference on Machine Learning*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shalev-Shwartz, S., Shammah, S. and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Shi, C., Zhang, S., Lu, W. and Song, R. (2020). Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*.

Shi, C., Zhu, J., Shen, Y., Luo, S., Zhu, H. and Song, R. (2022). Off-policy confidence interval estimation with confounded markov decision process. *arXiv preprint arXiv:2202.10589*.

Shoham, Y., Powers, R. and Grenager, T. (2003). Multi-agent reinforcement learning: a critical survey. *Technical Report*.

Shoham, Y., Powers, R. and Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, **171** 365–377.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, **529** 484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. et al. (2017). Mastering the game of Go without human knowledge. *Nature*, **550** 354.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, **3** 9–44.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Sutton, R. S., Barto, A. G. et al. (1998). *Introduction to Reinforcement Learning*, vol. 135. MIT press.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C. and Wiewiora, E. (2009a). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM.

Sutton, R. S., Maei, H. R. and Szepesvári, C. (2009b). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*.

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

Szepesvári, C. and Munos, R. (2005). Finite time bounds for sampling based fitted value iteration. In *International Conference on Machine Learning*. ACM.

Sznitman, A.-S. (1991). Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*. Springer, 165–251.

Tang, Z., Feng, Y., Li, L., Zhou, D. and Liu, Q. (2019). Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*.

Tembine, H. and Huang, M. (2011). Mean field difference games: Mckean-Vlasov dynamics. In *Conference on Decision and Control and European Control Conference*. IEEE.

Tembine, H., Zhu, Q. and Başar, T. (2014). Risk-sensitive mean-field games. *IEEE Transactions on Automatic Control*, **59** 835–850.

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

Todorov, E., Erez, T. and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE.

Tosatto, S., Pirotta, M., D'Eramo, C. and Restelli, M. (2017). Boosted fitted Q-iteration. In *International Conference on Machine Learning*. JMLR. org.

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

Tu, S. and Recht, B. (2017). Least-squares temporal difference learning for the linear quadratic regulator. *arXiv preprint arXiv:1712.08642*.

Tu, S. and Recht, B. (2018). The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *arXiv preprint arXiv:1812.03565* 3036–3083.

Uehara, M., Huang, J. and Jiang, N. (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*. PMLR.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W. and Xie, T. (2021). Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*.

Uehara, M. and Jiang, N. (2019). Minimax weight and Q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809* 9659–9668.

Van Roy, B. (1998). *Learning and value function approximation in complex decision processes*. Ph.D. thesis, Massachusetts Institute of Technology.

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W., Dudzik, A., Huang, A., Georgiev, P., Powell, R. et al. (2019). Alphastar: Mastering the real-time strategy game starcraft ii. https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii/.

Wai, H.-T., Yang, Z., Wang, P. Z. and Hong, M. (2018). Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Advances in Neural Information Processing Systems*.

Wang, B.-C. and Zhang, J.-F. (2012). Mean field games for large-population multiagent systems with Markov jump parameters. *SIAM Journal on Control and Optimization*, **50** 2308–2334.

Wang, J., Qi, Z. and Wong, R. K. (2021a). Projected state-action balancing weights for offline reinforcement learning. *arXiv preprint arXiv:2109.04640*.

Wang, J., Zhang, W., Yuan, S. et al. (2017a). Display advertising with real-time bidding (RTB) and behavioural targeting. *Foundations and Trends® in Information Retrieval*, **11** 297–435.

Wang, L., Cai, Q., Yang, Z. and Wang, Z. (2019). Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*.

Wang, L. and Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 531–550.

Wang, L., Yang, Z. and Wang, Z. (2021b). Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, **34**.

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

Wang, R., Wu, Y., Salakhutdinov, R. and Kakade, S. (2021c). Instabilities of offline rl with pre-trained neural representation. In *International Conference on Machine Learning*. PMLR.

Wang, Y., Chen, W., Liu, Y., Ma, Z.-M. and Liu, T.-Y. (2017b). Finite sample analysis of the GTD policy evaluation algorithms in Markov setting. In *Advances in Neural Information Processing Systems*.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, **8** 279–292.

Wei, C.-Y., Hong, Y.-T. and Lu, C.-J. (2017). Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*.

Wu, L., Ma, C. and Weinan, E. (2018). How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*.

Wu, Y., Zhang, W., Xu, P. and Gu, Q. (2020). A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P. and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, **34**.

Xie, T., Ma, Y. and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*.

Xu, T., Wang, Z. and Liang, Y. (2020). Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*.

Xu, T., Zou, S. and Liang, Y. (2019). Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*.

Yang, E. and Gu, D. (2004). Multiagent reinforcement learning for multi-robot systems: A survey. *Manuscript*.

Yang, J., Ye, X., Trivedi, R., Xu, H. and Zha, H. (2018a). Deep mean field games for learning optimal behavior policy of large populations. In *International Conference on Learning Representations*.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W. and Wang, J. (2018b). Mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1802.05438*.

Yang, Z., Chen, Y., Hong, M. and Wang, Z. (2019a). On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*.

Yang, Z., Chen, Y., Hong, M. and Wang, Z. (2019b). On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*.

Yang, Z., Xie, Y. and Wang, Z. (2019c). A theoretical analysis of deep Q-learning. *arXiv preprint arXiv:1901.00137*.

Yu, H. (2010). Convergence of least squares temporal difference methods under general conditions. In *International Conference on Machine Learning*.

Yu, H. (2017). On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv preprint arXiv:1712.09652*.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C. and Ma, T. (2020). MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

Zhan, W., Huang, B., Huang, A., Jiang, N. and Lee, J. D. (2022). Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*.

Zhang, C. and Liu, Y. (2014). Multicategory angle-based large-margin classification. *Biometrika*, **101** 625–640.

Zhang, J. and Bareinboim, E. (2019). Near-optimal reinforcement learning in dynamic treatment regimes. *Advances in Neural Information Processing Systems*, **32**.

Zhang, K., Koppel, A., Zhu, H. and Başar, T. (2019a). Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*.

Zhang, K., Yang, Z. and Başar, T. (2019b). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*.

Zhang, K., Yang, Z., Liu, H., Zhang, T. and Başar, T. (2018). Finite-sample analyses for fully decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:1812.02783*.

Zhang, R., Dai, B., Li, L. and Schuurmans, D. (2020). Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*.

Zhou, X., Mayer-Hamblett, N., Khan, U. and Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, **112** 169–187.

Zhou, X. Y. and Li, D. (2000). Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics and Optimization*, **42** 19–33.

Ziebart, B. D., Maas, A. L., Bagnell, J. A. and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, vol. 3.

Zou, D., Cao, Y., Zhou, D. and Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*.

Zou, S., Xu, T. and Liang, Y. (2019). Finite-sample analysis for SARSA and Q-learning with linear function approximation. *arXiv preprint arXiv:1902.02234* 8665–8675.

APPENDIX A

Supplemental Materials in Chapter 2

A.1. Details of Algorithms

We introduce the actor-critic method with DNN approximation in Algorithm 4, which relies on Algorithms 5 and 6 for the actor and critic updates.

Algorithm 4 Deep Neural Actor-Critic Method

```
Input: Number of iterations $K, N_a, N_c$, stepsizes $\alpha, \eta$, and temperature parameter $\beta$.
Initialization: Set $\tau_0 \leftarrow \infty$ and initialize DNNs $f_{\theta_0}$ and $Q_{\omega_0}$ as specified in §2.2.2.
for $k = 0, 1, 2, \ldots, K$ do
    Actor Update: Update $\theta_{k+1}$ via Algorithm 5 with input $\pi_{\theta_k}$, $\theta_0$, $Q_{\omega_k}$, $\alpha$, $\beta$, $\tau_{k+1} = (k+1)^{-1} \cdot \beta$, and $N_a$.
    Critic Update: Update $\omega_{k+1}$ via Algorithm 6 with input $\pi_{\theta_{k+1}}$, $Q_{\omega_k}$, $\omega_0$, $\eta$, and $N_c$.
end for
Output: $\{\pi_{\theta_k}\}_{k \in [K+1]}$, where $\pi_{\theta_k} \propto \exp(\tau_k^{-1} f_{\theta_k})$.
```

Algorithm 5 Actor Update for Deep Neural Actor-Critic Method

```
Input: Policy $\pi_\theta \propto \exp(\tau^{-1} f_\theta)$, initial actor parameter $\theta_0$, action-value function $Q_\omega$, stepsize $\alpha$, temperature parameter $\beta$, temperature $\widetilde{\tau}$, and number of iterations $N_a$.
Initialization: Set $\theta(0) \leftarrow \theta_0$.
for $n = 0, 1, 2, \ldots, N_a - 1$ do
    Sample $(s, a)$ as specified in §2.2.2.
    Set $\theta(n+1) \leftarrow \Gamma_{\mathcal{B}(\theta_0, R_a)}(\theta(n) - \alpha \cdot (f_{\theta(n)}(s, a) - \widetilde{\tau} \cdot (\beta^{-1} Q_\omega(s, a) + \tau^{-1} f_\theta(s, a))) \cdot \nabla_\theta f_{\theta(n)}(s, a))$.
end for
Output: $\overline{\theta} = 1/N_a \cdot \sum_{n=1}^{N_a} \theta(n)$.
```

Algorithm 6 Critic Update for Deep Neural Actor-Critic Method

```
Input: Policy $\pi_\theta$, action-value function $Q_\omega$, initial critic parameter $\omega_0$, stepsize $\eta$, and number of iterations $N_{\rm c}$.
Initialization: Set $\omega(0) \leftarrow \omega_0$.
for $n = 0, 1, 2, \ldots, N_{\rm c} - 1$ do
    Sample $(s, a, r, s', a')$ as specified in §2.2.2.
    Set $\delta(n) \leftarrow Q_{\omega(n)}(s, a) - (1 - \gamma) \cdot r - \gamma \cdot Q_\omega(s', a')$.
    Set $\omega(n + 1) \leftarrow \Gamma_{\mathcal{B}(\omega_0, R_{\rm c})}(\omega(n) - \eta \cdot \delta(n) \cdot \nabla_\omega Q_{\omega(n)}(s, a))$.
end for
Output: $\overline{\omega} = 1/N_{\rm c} \cdot \sum_{n=1}^{N_{\rm c}} \omega(n)$.
```

A.2. Convergence Results of Algorithm 4

In this section, we upper bound the regret of the deep neural actor-critic method. Hereafter we assume that $|\mathcal{R}(s, a)| \le \mathcal{R}_{\max}$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $\mathcal{R}_{\max}$ is a positive absolute constant. First, we impose the following assumptions in parallel to Assumption 2.3.1. Recall that ρ^* is the stationary state-action distribution of π^*, while ρ_k is the stationary state-action distribution of π_{θ_k}.

Assumption A.2.1 (Concentrability Coefficient)**.** The following statements hold.

(i) There exists a positive absolute constant ϕ^* such that $\phi_k^* \le \phi^*$ for any $k \ge 1$, where $\phi_k^* = \|{\rm d}\rho^*/{\rm d}\rho_k\|_{\rho_k, 2}$.

(ii) For the state-action distribution ρ used to define the regret in (2.3.1), we assume that for any $k \ge 1$ and a sequence of policies $\{\pi_i\}_{i \ge 1}$, the k-step future-state-action distribution $\rho \mathbb{P}^{\pi_1} \cdots \mathbb{P}^{\pi_k}$ is absolutely continuous with respect to ρ^*. Also, it holds that

$$C_{\rho,\rho^*} = (1-\gamma)^2 \sum_{k=1}^{\infty} k^3 \gamma^k \cdot c(k) < \infty,$$

where $c(k) = \sup_{\{\pi_i\}_{i\in[k]}} \|\mathrm{d}(\rho\mathbb{P}^{\pi_1}\cdots\mathbb{P}^{\pi_k})/\mathrm{d}\rho^*\|_{\rho^*,\infty}$.

Meanwhile, we impose the following assumption in parallel to Assumption 2.3.2.

Assumption A.2.2 (Zero Approximation Error)**.** For any $Q_\omega \in \mathcal{U}(m_{\mathrm{c}}, H_{\mathrm{c}}, R_{\mathrm{c}})$ and policy π, it holds that $\mathcal{T}^\pi Q_\omega \in \mathcal{U}(m_{\mathrm{c}}, H_{\mathrm{c}}, R_{\mathrm{c}})$, where $\mathcal{T}^\pi$ is defined in (2.1.4).

Assumption A.2.2 states that $\mathcal{U}(m_{\mathrm{c}}, H_{\mathrm{c}}, R_{\mathrm{c}})$ is closed under the Bellman evaluation operator $\mathcal{T}^\pi$, which is commonly imposed in the literature (Munos and Szepesvári, 2008; Antos et al., 2008a; Farahmand et al., 2010, 2016; Tosatto et al., 2017; Yang et al., 2019c; Liu et al., 2019).

We upper bound the regret of the deep neural actor-critic method in Algorithm 4 in the sequel. To establish such an upper bound, we first establish the rates of convergence of Algorithms 5 and 6 as follows.

Proposition A.2.3. For any sufficiently large $N_{\mathrm{a}} > 0$, let $m_{\mathrm{a}} = \Omega(d^{3/2} R_{\mathrm{a}}^{-1} H_{\mathrm{a}}^{-3/2} \log(m_{\mathrm{a}}^{1/2}/R_{\mathrm{a}})^{3/2})$, $H_{\mathrm{a}} = O(N_{\mathrm{a}}^{1/4})$, and $R_{\mathrm{a}} = O(m_{\mathrm{a}}^{1/2} H_{\mathrm{a}}^{-6} (\log m_{\mathrm{a}})^{-3})$. We denote by $\overline{\theta}$ the output of Algorithm 5 with input $\pi_\theta \propto \exp(\tau^{-1} f_\theta)$, θ_0, Q_ω, α, β, $\widetilde{\tau} = (\tau^{-1} + \beta^{-1})^{-1}$, and N_{a}. Also, let $\widetilde{f} = \widetilde{\tau} \cdot (\beta^{-1} Q_\omega + \tau^{-1} f_\theta)$. With probability at least $1 - \exp(-\Omega(R_{\mathrm{a}}^{2/3} m_{\mathrm{a}}^{2/3} H_{\mathrm{a}}))$ over the random initialization θ_0, we have

$$\mathbb{E}\big[\big(f_{\overline{\theta}}(s,a) - \widetilde{f}(s,a)\big)^2\big] = O(R_{\mathrm{a}}^2 N_{\mathrm{a}}^{-1/2} + R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/6} H_{\mathrm{a}}^7 \log m_{\mathrm{a}}).$$

Here the expectation is taken over the randomness of $\overline{\theta}$ conditioning on the initialization θ_0 and $(s, a) \sim \rho_{\pi_\theta}$, where ρ_{π_θ} is the stationary state-action distribution of π_θ.

Proof. See §A.5.2 for a detailed proof. □

Proposition A.2.4. For any sufficiently large $N_{\mathrm{c}} > 0$, let $m_{\mathrm{c}} = \Omega(d^{3/2} R_{\mathrm{c}}^{-1} H_{\mathrm{c}}^{-3/2} \log(m_{\mathrm{c}}^{1/2}/R_{\mathrm{c}})^{3/2})$, $H_{\mathrm{c}} = O(N_{\mathrm{c}}^{1/4})$, and $R_{\mathrm{c}} = O(m_{\mathrm{c}}^{1/2} H_{\mathrm{c}}^{-6} (\log m_{\mathrm{c}})^{-3})$. We denote by $\overline{\omega}$ the output of Algorithm 6 with input π_θ, Q_ω, ω_0, η, and N_{c}. Also, let $\widetilde{Q} = (1-\gamma) \cdot \mathcal{R} + \gamma \cdot \mathbb{P}^{\pi_\theta} Q_\omega$. With probability at least $1 - \exp(-\Omega(R_{\mathrm{c}}^{2/3} m_{\mathrm{c}}^{2/3} H_{\mathrm{c}}))$ over the random initialization ω_0, we have

$$\mathbb{E}\big[\big(Q_{\bar{\omega}}(s,a) - \widetilde{Q}(s,a)\big)^2\big] = O(R_{\mathrm{c}}^2 N_{\mathrm{c}}^{-1/2} + R_{\mathrm{c}}^{8/3} m_{\mathrm{c}}^{-1/6} H_{\mathrm{c}}^7 \log m_{\mathrm{c}}).$$

Here the expectation is taken over the randomness of $\overline{\omega}$ conditioning on the initialization ω_0 and $(s, a) \sim \rho_{\pi_\theta}$, where ρ_{π_θ} is the stationary state-action distribution of π_θ.

Proof. See §A.5.3 for a detailed proof. □

Propositions A.2.3 and A.2.4 characterize the errors that arise from the actor and critic updates in Algorithm 4, respectively. In particular, if the widths m_{a} and m_{c} of the DNNs f_θ and Q_ω are sufficiently large, the errors characterized in Propositions A.2.3 and A.2.4 decay to zero at the rates of $O(N_{\mathrm{a}}^{-1/2})$ and $O(N_{\mathrm{c}}^{-1/2})$, respectively. Propositions A.2.3 and A.2.4 act as the key ingredients to upper bounding the regret of the deep neural actor-critic method.

Based on Propositions A.2.3 and A.2.4, we upper bound the regret of Algorithm 4 in the following theorem, which is in parallel to Theorem 2.3.4.

Theorem A.2.5. We assume that Assumptions A.2.1 and A.2.2 hold. Let ρ be a state-action distribution satisfying (ii) of Assumption A.2.1. Also, for any sufficiently large

$K > 0$, let $N_{\mathrm{a}} = \Omega(K^6 C^4_{\rho,\rho^*}(\phi^* + \psi^* + 1)^4 R_{\mathrm{a}}^4)$, $N_{\mathrm{c}} = \Omega(K^6 C^4_{\rho,\rho^*}\phi^{*4} R_{\mathrm{c}}^4)$, $H_{\mathrm{a}} = H_{\mathrm{c}} = O(N_{\mathrm{c}}^{1/4})$, $R_{\mathrm{a}} = R_{\mathrm{c}} = O(m_{\mathrm{c}}^{1/2} H_{\mathrm{c}}^{-6}(\log m_{\mathrm{c}})^{-3})$, $m_{\mathrm{a}} = m_{\mathrm{c}} = \Omega(d^{3/2} K^6 C^{12}_{\rho,\rho^*}(\phi^* + \psi^* + 1)^{12} R_{\mathrm{c}}^{16} H_{\mathrm{c}}^{42} \log(m_{\mathrm{c}}^{1/2}/R_{\mathrm{c}})^{3/2})$, $\beta = K^{1/2}$, and the sequence $\{\theta_k\}_{k\in[K]}$ be generated by Algorithm 4. With probability at least $1 - 1/K$ over the random initialization θ_0 and ω_0, it holds that

$$\mathbb{E}\Big[\sum_{k=0}^{K} Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a)\Big] \le \big(2(1-\gamma)^{-3}\log|\mathcal{A}| + O(1)\big)\cdot K^{1/2},$$

where the expectation is taken over the randomness of $(s,a) \sim \rho$ and $\{\theta_{k+1}\}_{k\in[K]}$ conditioning on the initialization θ_0 and ω_0.

Proof. See §A.3.2 for a detailed proof. □

When the architecture of the actor and critic neural networks are properly chosen, Theorem A.2.5 establishes an $O(K^{1/2})$ regret of Algorithm 4, where K is the total number of iterations. Specifically speaking, to establish such a regret upper bound, we need the widths m_{a} and m_{c} of the DNNs f_θ and Q_ω to be sufficiently large. Meanwhile, to control the errors of actor update and critic update in Algorithm 4, we also run sufficiently large numbers of iterations in Algorithms 5 and 6.

In terms of the total sample complexity, to simplify our discussion, we omit constant and logarithmic terms here. To obtain an ε-globally optimal policy, it suffices to set $K \asymp \varepsilon^{-2}$ in Algorithm 4. By plugging such a K into $N_{\mathrm{a}} = \Omega(K^6 C^4_{\rho,\rho^*}(\phi^* + \psi^* + 1)^4 R_{\mathrm{a}}^4)$ and $N_{\mathrm{c}} = \Omega(K^6 C^4_{\rho,\rho^*}\phi^{*4} R_{\mathrm{c}}^4)$ as required in Theorem A.2.5, we have $N_{\mathrm{a}} = \widetilde{O}(\varepsilon^{-12})$ and $N_{\mathrm{c}} = \widetilde{O}(\varepsilon^{-12})$. Thus, to achieve an ε-globally optimal policy, the total sample complexity of Algorithm 4 is $\widetilde{O}(\varepsilon^{-14})$. With the modification to off-policy setting as in §2.2.1, the total sample complexity of Algorithm 4 is $\widetilde{O}(\varepsilon^{-12})$.

To the best of our knowledge, we establish the rate of convergence and global optimality of the actor-critic method under single-timescale setting with DNN approximation for the first time.

A.3. Proofs of Theorems

A.3.1. Proof of Theorem 2.3.4

Recall that ρ is a state-action distribution satisfying (ii) of Assumption 2.3.1. We first upper bound $\sum_{k=0}^{K}(Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a))$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ in part 1. Then by further taking the expectation over ρ and invoking Lemma 2.4.1 in part 2, we conclude the proof of Theorem 2.3.4.

Part 1. In the sequel, we upper bound $\sum_{k=0}^{K}(Q^*(s,a)-Q^{\pi_{\theta_{k+1}}}(s,a))$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$. By the definition of Q^* in (2.1.2), it holds for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ that

$$
\begin{aligned}
&[Q^* - Q^{\pi_{\theta_{k+1}}}](s,a) \\
&\quad = \sum_{\ell=0}^{\infty} \big[(1-\gamma)\cdot(\gamma\mathbb{P}^{\pi^*})^{\ell}\mathcal{R}\big](s,a) - Q^{\pi_{\theta_{k+1}}}(s,a) \\
&\quad = \sum_{\ell=0}^{\infty} \big[(1-\gamma)\cdot(\gamma\mathbb{P}^{\pi^*})^{\ell}\mathcal{R} + (\gamma\mathbb{P}^{\pi^*})^{\ell+1}Q^{\pi_{\theta_{k+1}}} - (\gamma\mathbb{P}^{\pi^*})^{\ell+1}Q^{\pi_{\theta_{k+1}}}\big](s,a) - Q^{\pi_{\theta_{k+1}}}(s,a) \\
&\quad = \sum_{\ell=0}^{\infty} \big[(1-\gamma)\cdot(\gamma\mathbb{P}^{\pi^*})^{\ell}\mathcal{R} + (\gamma\mathbb{P}^{\pi^*})^{\ell+1}Q^{\pi_{\theta_{k+1}}} - (\gamma\mathbb{P}^{\pi^*})^{\ell}Q^{\pi_{\theta_{k+1}}}\big](s,a) \\
&\quad = \sum_{\ell=0}^{\infty} \Big[(\gamma\mathbb{P}^{\pi^*})^{\ell}\big((1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi^*}Q^{\pi_{\theta_{k+1}}} - Q^{\pi_{\theta_{k+1}}}\big)\Big](s,a),
\end{aligned}
\tag{A.3.1}
$$

where $\mathbb{P}^{\pi^*}$ is defined in (2.1.3). We upper bound $[(1-\gamma)\cdot\mathcal{R}+\gamma\cdot\mathbb{P}^{\pi^*}Q^{\pi_{\theta_{k+1}}} - Q^{\pi_{\theta_{k+1}}}](s,a)$ on the RHS of (A.3.1) in the sequel. By calculation, we have

$$
\begin{aligned}
&\big[(1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi^*}Q^{\pi_{\theta_{k+1}}} - Q^{\pi_{\theta_{k+1}}}\big](s,a) \\
&\quad = \Big[\big((1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi^*}Q^{\pi_{\theta_{k+1}}}\big) - \big((1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi^*}Q_{\omega_k}\big)\Big](s,a) \\
&\qquad + \Big[\big((1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi^*}Q_{\omega_k}\big) - \big((1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi_{\theta_{k+1}}}Q_{\omega_k}\big)\Big](s,a) \\
&\qquad + \Big[\big((1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi_{\theta_{k+1}}}Q_{\omega_k}\big) - Q^{\pi_{\theta_{k+1}}}\Big](s,a) \\
&\quad = A_{1,k}(s,a) + A_{2,k}(s,a) + A_{3,k}(s,a),
\end{aligned}
\tag{A.3.2}
$$

where $A_{1,k}$, $A_{2,k}$, and $A_{3,k}$ are defined as follows,

$$A_{1,k}(s,a) = \big[\gamma(\mathbb{P}^{\pi^*} - \mathbb{P}^{\pi_{\theta_{k+1}}})Q_{\omega_k}\big](s,a),$$

$$A_{2,k}(s,a) = \big[\gamma\mathbb{P}^{\pi^*}(Q^{\pi_{\theta_{k+1}}} - Q_{\omega_k})\big](s,a),$$

$$A_{3,k}(s,a) = [\mathcal{T}^{\pi_{\theta_{k+1}}}Q_{\omega_k} - Q^{\pi_{\theta_{k+1}}}](s,a). \tag{A.3.3}$$

Here $\mathcal{T}^{\pi_{\theta_{k+1}}}$ is defined in (2.1.4). By the following three lemmas, we upper bound $A_{1,k}$, $A_{2,k}$, and $A_{3,k}$ on the RHS of (A.3.2), respectively.

Lemma A.3.1. It holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$A_{1,k}(s,a) = \big[\gamma(\mathbb{P}^{\pi^*} - \mathbb{P}^{\pi_{\theta_{k+1}}})Q_{\omega_k}\big](s,a) \leq \big[\gamma\beta \cdot \mathbb{P}(\vartheta_k + \epsilon_{k+1}^{\mathrm{a}})\big](s,a),$$

where ϑ_k and $\epsilon_{k+1}^{\mathrm{a}}$ are defined as follows,

$$\vartheta_k(s) = \mathrm{KL}\big(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s)\big) - \mathrm{KL}\big(\pi^*(\cdot\,|\,s)\,\|\,\pi_{\theta_{k+1}}(\cdot\,|\,s)\big), \tag{A.3.4}$$

$$\epsilon_{k+1}^{\mathrm{a}}(s) = \big\langle \log\big(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)\big) - \beta^{-1}\cdot Q_{\omega_k}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\big\rangle. \tag{A.3.5}$$

Proof. See §A.6.2 for a detailed proof. □

We remark that $\epsilon_{k+1}^{\mathrm{a}} = 0$ for any k in the linear actor-critic method. Meanwhile, such a term is included in Lemma A.3.1 only aiming to generalize to the deep neural actor-critic method.

Lemma A.3.2. It holds for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ that

$$A_{2,k}(s,a) \leq \big[(\gamma\mathbb{P}^{\pi^*})^{k+1}(Q^* - Q_{\omega_0})\big](s,a) + \gamma\beta \cdot \sum_{i=0}^{k-1}\big[(\gamma\mathbb{P}^{\pi^*})^{k-i}\mathbb{P}(\vartheta_i + \epsilon^{\rm a}_{i+1})\big](s,a) + \sum_{i=0}^{k-1}\big[(\gamma\mathbb{P}^{\pi^*})^{k-i}\epsilon^{\rm c}_{i+1}\big](s,a),$$

where ϑ_i is defined in (A.3.4) of Lemma A.3.1, $\epsilon^{\rm a}_{i+1}$ is defined in (A.3.5) of Lemma A.3.1, and $\epsilon^{\rm c}_{i+1}$ is defined as follows,

$$\epsilon^{\rm c}_{i+1}(s,a) = [\mathcal{T}^{\pi_{\theta_{i+1}}}Q_{\omega_i} - Q_{\omega_{i+1}}](s,a). \tag{A.3.6}$$

Proof. See §A.6.3 for a detailed proof. □

We remark that $\epsilon^{\rm a}_{k+1} = 0$ for any k in the linear actor-critic method. Meanwhile, such a term is included in Lemma A.3.2 only aiming to generalize to the deep neural actor-critic method.

Lemma A.3.3. It holds for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ that

$$A_{3,k}(s,a) = \big[\gamma\mathbb{P}^{\pi_{\theta_{k+1}}}(I - \gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}e_{k+1}\big](s,a),$$

where e_{k+1} is defined as follows,

$$e_{k+1}(s,a) = [Q_{\omega_k} - \mathcal{T}^{\pi_{\theta_{k+1}}}Q_{\omega_k}](s,a). \tag{A.3.7}$$

Proof. See §A.6.4 for a detailed proof. □

We upper bound e_{k+1} in (A.3.7) of Lemma A.3.3 using Lemma A.3.4 as follows.

Lemma A.3.4. It holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$e_{k+1}(s,a) \leq \biggl[\gamma^k \Bigl(\prod_{s=1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Bigr) e_1 + \sum_{i=1}^{k} \gamma^{k-i} \Bigl(\prod_{s=i+1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Bigr) \bigl(\gamma\beta\mathbb{P}\epsilon_{i+1}^{\mathrm{b}} + (I - \gamma\mathbb{P}^{\pi_{\theta_i}})\epsilon_i^{\mathrm{c}}\bigr)\biggr](s,a).$$

where $\epsilon_i^{\mathrm{c}}(s,a)$ is defined in (A.3.6) of Lemma A.3.2 and $\epsilon_{i+1}^{\mathrm{b}}(s)$ is defined as follows,

$$\epsilon_{i+1}^{\mathrm{b}}(s) = \bigl\langle \log\bigl(\pi_{\theta_{i+1}}(\cdot \,|\, s)/\pi_{\theta_i}(\cdot \,|\, s)\bigr) - \beta^{-1} \cdot Q_{\omega_i}(s,\cdot), \pi_{\theta_i}(\cdot \,|\, s) - \pi_{\theta_{i+1}}(\cdot \,|\, s)\bigr\rangle. \tag{A.3.8}$$

Proof. See §A.6.5 for a detailed proof. □

We remark that $\epsilon_{i+1}^{\mathrm{b}} = 0$ for any i in the linear actor-critic method. Meanwhile, such a term is included in Lemma A.3.4 only aiming to generalize to the deep neural actor-critic method.

Combining Lemmas A.3.3 and A.3.4, we obtain the following upper bound of $A_{3,k}$,

$$\begin{aligned} A_{3,k}(s,a) &= \bigl[\gamma\mathbb{P}^{\pi_{\theta_{k+1}}}(I - \gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1} e_{k+1}\bigr](s,a) \\ &\leq \biggl[\gamma\mathbb{P}^{\pi_{\theta_{k+1}}}(I - \gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1} \biggl(\gamma^k \Bigl(\prod_{s=1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Bigr) e_1 \\ &\qquad + \sum_{i=1}^{k} \gamma^{k-i} \Bigl(\prod_{s=i+1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Bigr) \bigl(\beta\gamma\mathbb{P}\epsilon_{i+1}^{\mathrm{b}} + (I - \gamma\mathbb{P}^{\pi_{\theta_i}})\epsilon_i^{\mathrm{c}}\bigr)\biggr)\biggr](s,a). \end{aligned} \tag{A.3.9}$$

Combining (A.3.1), (A.3.2), Lemma A.3.1 and Lemma A.3.2, it holds for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ that

$$
\begin{aligned}
&\sum_{k=0}^{K} [Q^* - Q^{\pi_{\theta_{k+1}}}](s, a) \\
&\quad \le \sum_{k=0}^{K} \Big[(I - \gamma \mathbb{P}^{\pi^*})^{-1} \big((\gamma \mathbb{P}^{\pi^*})^{k+1} (Q^* - Q_{\omega_0}) + \sum_{i=0}^{k} (\gamma \mathbb{P}^{\pi^*})^{k-i} \gamma \beta \mathbb{P} (\vartheta_i + \epsilon_{i+1}^{\mathrm{a}}) \\
&\qquad\qquad + \sum_{i=0}^{k-1} (\gamma \mathbb{P}^{\pi^*})^{k-i} \epsilon_{i+1}^{\mathrm{c}} + A_{3,k} \big) \Big] (s, a)
\end{aligned}
$$

(A.3.10)

$$
\begin{aligned}
&= \Big[(I - \gamma \mathbb{P}^{\pi^*})^{-1} \Big(\sum_{k=0}^{K} (\gamma \mathbb{P}^{\pi^*})^{k+1} (Q^* - Q_{\omega_0}) + \sum_{k=0}^{K} \sum_{i=0}^{k} (\gamma \mathbb{P}^{\pi^*})^{k-i} \gamma \beta \mathbb{P} \epsilon_{i+1}^{\mathrm{a}} \\
&\qquad + \sum_{k=0}^{K} \sum_{i=0}^{k-1} (\gamma \mathbb{P}^{\pi^*})^{k-i} \epsilon_{i+1}^{\mathrm{c}} + \sum_{k=0}^{K} A_{3,k} + \sum_{k=0}^{K} \sum_{i=0}^{k} (\gamma \mathbb{P}^{\pi^*})^{k-i} \gamma \beta \mathbb{P} \vartheta_i \Big) \Big] (s, a),
\end{aligned}
$$

where ϑ_i, $\epsilon_{i+1}^{\mathrm{a}}$, $\epsilon_{i+1}^{\mathrm{c}}$, and e_{k+1} are defined in (A.3.4) of Lemma A.3.1, (A.3.5) of Lemma A.3.1, (A.3.6) of Lemma A.3.2, and (A.3.7) of Lemma A.3.3, respectively. We upper bound the last term as follows,

$$\begin{aligned}
\biggl[\sum_{k=0}^{K}\sum_{i=0}^{k}(\gamma\mathbb{P}^{\pi^*})^{k-i}\gamma\beta\mathbb{P}\vartheta_i\biggr](s,a) &= \biggl[\sum_{k=0}^{K}\sum_{i=0}^{k}\gamma\beta(\gamma\mathbb{P}^{\pi^*})^{i}\mathbb{P}\vartheta_{k-i}\biggr](s,a) \\
&= \biggl[\sum_{i=0}^{K}\gamma\beta(\gamma\mathbb{P}^{\pi^*})^{i}\mathbb{P}\sum_{k=i}^{K}\vartheta_{k-i}\biggr](s,a) \\
&= \biggl[\sum_{i=0}^{K}\gamma\beta(\gamma\mathbb{P}^{\pi^*})^{i}\mathbb{P}\sum_{k=i}^{K}\Bigl(\mathrm{KL}\bigl(\pi^*\,\|\,\pi_{\theta_{k-i}}\bigr)-\mathrm{KL}\bigl(\pi^*\,\|\,\pi_{\theta_{k-i+1}}\bigr)\Bigr)\biggr](s,a) \\
&= \biggl[\sum_{i=0}^{K}\gamma\beta(\gamma\mathbb{P}^{\pi^*})^{i}\mathbb{P}\bigl(\mathrm{KL}(\pi^*\,\|\,\pi_{\theta_0})-\mathrm{KL}(\pi^*\,\|\,\pi_{\theta_{K-i+1}})\bigr)\biggr](s,a) \\
&\le \biggl[\sum_{i=0}^{K}\gamma\beta(\gamma\mathbb{P}^{\pi^*})^{i}\mathbb{P}\mathrm{KL}(\pi^*\,\|\,\pi_{\theta_0})\biggr](s,a), \qquad \text{(A.3.11)}
\end{aligned}$$

where we use the definition of ϑ_{k-i} in (A.3.4) of Lemma A.3.1 and the non-negativity of the KL divergence in the second equality and the last inequality, respectively. By plugging

(A.3.9) and (A.3.11) into (A.3.10), we have

$$
\begin{aligned}
&\sum_{k=0}^{K}[Q^* - Q^{\pi_{\theta_{k+1}}}](s,a) \\
&\text{(A.3.12)} \\
&\quad \le \biggl[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\biggl(\sum_{k=0}^{K}(\gamma\mathbb{P}^{\pi^*})^{k+1}(Q^* - Q_{\omega_0}) + \sum_{k=0}^{K}\sum_{i=0}^{k}(\gamma\mathbb{P}^{\pi^*})^{k-i}\gamma\beta\mathbb{P}\epsilon^{\mathrm{a}}_{i+1} \\
&\quad + \sum_{k=0}^{K}\sum_{i=0}^{k-1}(\gamma\mathbb{P}^{\pi^*})^{k-i}\epsilon^{\mathrm{c}}_{i+1} + \sum_{k=0}^{K}\gamma^{k+1}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\Bigl(\prod_{s=1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Bigr)e_1 \\
&\quad + \sum_{k=0}^{K}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\sum_{\ell=1}^{k}\gamma^{k-\ell+1}\Bigl(\prod_{s=\ell+1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Bigr)\bigl(\gamma\beta\mathbb{P}\epsilon^{\mathrm{b}}_{\ell+1} + (I-\gamma\mathbb{P}^{\pi_{\theta_\ell}})\epsilon^{\mathrm{c}}_{\ell}\bigr)\biggr)\biggr](s,a). \\
&\quad + \sum_{i=0}^{K}(\gamma\mathbb{P}^{\pi^*})^{i}\gamma\beta\mathbb{P}\mathrm{KL}(\pi^*\,\|\,\pi_{\theta_0})
\end{aligned}
$$

We remark that $\epsilon^{\mathrm{a}}_{i+1} = \epsilon^{\mathrm{b}}_{i+1} = 0$ for any i in the linear actor-critic method. Meanwhile, such terms is included in (A.3.12) only aiming to generalize to the deep neural actor-critic method. This concludes the proof in part 1.

Part 2. Recall that ρ is a state-action distribution satisfying (ii) of Assumption 2.3.1. In the sequel, we take the expectation over ρ in (A.3.12) and upper bound each term. Recall that $\epsilon^{\mathrm{a}}_{i+1} = \epsilon^{\mathrm{b}}_{i+1} = 0$ for any i in the linear actor-critic method. Hence, we only need to consider terms in (A.3.12) that do not involve $\epsilon^{\mathrm{a}}_{i+1}$ or $\epsilon^{\mathrm{b}}_{i+1}$. We first upper bound terms on the RHS of (A.3.12) that do not involve $\epsilon^{\mathrm{c}}_{i+1}$. More specifically, for any measure ρ

satisfying satisfying (ii) of Assumption 2.3.1, we upper bound the following three terms,

$$
\begin{aligned}
M_1 &= \mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}(\gamma\mathbb{P}^{\pi^*})^{k+1}(Q^*-Q_{\omega_0})\Big], \\
M_2 &= \mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\gamma^{k+1}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\Big(\prod_{s=1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Big)e_1\Big], \\
M_3 &= \mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{i=0}^{K}(\gamma\mathbb{P}^{\pi^*})^{i}\gamma\beta\mathbb{P}\mathrm{KL}(\pi^*\,\|\,\pi_{\theta_0})\Big].
\end{aligned}
\tag{A.3.13}
$$

We upper bound M_1, M_2, and M_3 in the following lemma.

Lemma A.3.5. It holds that

$$
\begin{aligned}
&|M_1| \le 4(1-\gamma)^{-2}\cdot(\mathcal{R}_{\max}+R), \qquad |M_2| \le (1-\gamma)^{-3}\cdot(2R+\mathcal{R}_{\max}), \\
&|M_3| \le (1-\gamma)^{-2}\cdot\log|\mathcal{A}|\cdot K^{1/2},
\end{aligned}
$$

where M_1, M_2, and M_3 are defined in (A.3.13).

Proof. See §A.6.6 for a detailed proof. □

Now, we upper bound terms on the RHS of (A.3.12) that involve $\epsilon_{i+1}^{\mathrm{c}}$. More specifically, for any measure ρ satisfying (ii) of Assumption 2.3.1, we upper bound the following two terms,

$$
\begin{aligned}
M_4 &= \mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\sum_{i=0}^{k}(\gamma\mathbb{P}^{\pi^*})^{k-i}\epsilon_{i+1}^{\mathrm{c}}\Big], \\
M_5 &= \mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\sum_{\ell=1}^{k}\gamma^{k-\ell+1}\Big(\prod_{s=\ell+1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Big)(I-\gamma\mathbb{P}^{\pi_{\theta_\ell}})\epsilon_\ell^{\mathrm{c}}\Big].
\end{aligned}
\tag{A.3.14}
$$

We upper bound M_4 and M_5 in the following lemma.

Lemma A.3.6. It holds that

$$|M_4| \leq 3KC_{\rho,\rho^*} \cdot \varepsilon_Q, \qquad |M_5| \leq KC_{\rho,\rho^*} \cdot \varepsilon_Q.$$

where M_4 and M_5 are defined in (A.3.14).

Proof. See §A.6.7 for a detailed proof. □

Now, by plugging Lemmas A.3.5 and A.3.6 into (A.3.12), we have

$$\mathbb{E}_\rho\Big[\sum_{k=0}^{K} Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a)\Big]$$
$$\leq 2(1-\gamma)^{-3} \cdot \log|\mathcal{A}| \cdot K^{1/2} + 4KC_{\rho,\rho^*} \cdot \varepsilon_Q + O(1). \tag{A.3.15}$$

Meanwhile, by changing measure from ρ^* to ρ_{k+1}, it holds for any k that

$$\mathbb{E}_{\rho^*}[|\epsilon^{\mathrm{c}}_{k+1}|] \leq \sqrt{\mathbb{E}_{\rho_{k+1}}\big[(\epsilon^{\mathrm{c}}_{k+1}(s,a))^2\big]} \cdot \phi^*_{k+1}, \tag{A.3.16}$$

where ϕ^*_{k+1} is defined in Assumption 2.3.1. Also, by Lemma 2.4.1, with probability at least $1-\delta$, it holds for any $k \in \{0,1,\ldots,K\}$ that

$$\sqrt{\mathbb{E}_{\rho_{k+1}}\big[(\epsilon^{\mathrm{c}}_{k+1}(s,a))^2\big]} = O\big(1/(\sqrt{N}\sigma^*) \cdot \log(KN/\delta)\big). \tag{A.3.17}$$

Now, by plugging (A.3.17) into (A.3.16), combining the definition of $\varepsilon_Q = \max_k \mathbb{E}_{\rho^*}[|\epsilon^{\mathrm{c}}_{k+1}(s,a)|]$, it holds with probability at least $1-\delta$ that

$$\varepsilon_Q = O\big(\phi^*/(\sqrt{N}\sigma^*) \cdot \log(KN/\delta)\big). \tag{A.3.18}$$

Combining (A.3.15), (A.3.18), and the choices of parameters stated in the theorem that

$$N = \Omega\big(KC_{\rho,\rho^*}^2(\phi^*/\sigma^*)^2 \cdot \log^2(KN/\delta)\big),$$

we have

$$\mathbb{E}_\rho\Big[\sum_{k=0}^K Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a)\Big] \le \big(2(1-\gamma)^{-3}\log|\mathcal{A}| + O(1)\big)\cdot K^{1/2},$$

which concludes the proof of Theorem 2.3.4.

A.3.2. Proof of Theorem A.2.5

We follow the proof of Theorem 2.3.4 in §A.3.1. Following similar arguments when deriving (A.3.12) in §A.3.1, we have

$$\sum_{k=0}^K [Q^* - Q^{\pi_{\theta_{k+1}}}](s,a)$$

(A.3.19)

$$\begin{aligned}
&\le \bigg[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\cdot\bigg(\sum_{k=0}^K(\gamma\mathbb{P}^{\pi^*})^{k+1}(Q^*-Q_{\omega_0}) + \sum_{k=0}^K\sum_{i=0}^k(\gamma\mathbb{P}^{\pi^*})^{k-i}\cdot\gamma\beta\mathbb{P}\epsilon_{i+1}^{\mathrm{a}} \\
&\quad + \sum_{k=0}^K\sum_{i=0}^{k-1}(\gamma\mathbb{P}^{\pi^*})^{k-i}\epsilon_{i+1}^{\mathrm{c}} + \sum_{i=0}^K(\gamma\mathbb{P}^{\pi^*})^i\cdot\gamma\beta\mathbb{P}\cdot\mathrm{KL}(\pi^*\,\|\,\pi_{\theta_0}) \\
&\quad + \sum_{k=0}^K\gamma^{k+1}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\Big(\prod_{s=1}^k\mathbb{P}^{\pi_{\theta_s}}\Big)e_1 \\
&\quad + \sum_{k=0}^K\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\sum_{\ell=1}^k\gamma^{k-\ell+1}\Big(\prod_{s=\ell+1}^k\mathbb{P}^{\pi_{\theta_s}}\Big)\big(\beta\gamma\mathbb{P}\epsilon_{\ell+1}^{\mathrm{b}} - (I-\gamma\mathbb{P}^{\pi_{\theta_\ell}})\epsilon_\ell^{\mathrm{c}}\big)\bigg)\bigg](s,a),
\end{aligned}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here $\epsilon^{\mathrm{a}}_{i+1}$, $\epsilon^{\mathrm{b}}_{\ell+1}$, $\epsilon^{\mathrm{c}}_{i+1}$, and e_1 are defined in (A.3.5), (A.3.8), (A.3.6), and (A.3.7), respectively.

Now, it remains to upper bound each term on the RHS of (A.3.19). We introduce the following error propagation lemma.

Lemma A.3.7. Suppose that

$$\mathbb{E}_{\rho_k}\Big[\big(f_{\theta_{k+1}}(s,a) - \tau_{k+1} \cdot (\beta^{-1} Q_{\omega_k}(s,a) - \tau_k^{-1} f_{\theta_k}(s,a))\big)^2\Big]^{1/2} \le \varepsilon_{k+1,f}. \tag{A.3.20}$$

Then, we have

$$\mathbb{E}_{\nu^*}\big[|\epsilon^{\mathrm{a}}_{k+1}(s)|\big] \le \sqrt{2}\tau_{k+1}^{-1} \cdot \varepsilon_{k+1,f} \cdot (\phi^*_k + \psi^*_k), \quad \mathbb{E}_{\nu^*}\big[|\epsilon^{\mathrm{b}}_{k+1}(s)|\big] \le \sqrt{2}\tau_{k+1}^{-1} \cdot \varepsilon_{k+1,f} \cdot (1 + \psi^*_k),$$

where $\epsilon^{\mathrm{a}}_{k+1}$ and $\epsilon^{\mathrm{b}}_{k+1}$ are defined in (A.3.5) and (A.3.8), respectively, ϕ^*_k and ψ^*_k are defined in Assumption A.2.1.

Proof. See §A.6.8 for a detailed proof. □

Following from Lemma A.4.4, with probability at least $1 - O(H_{\mathrm{c}}) \exp(-\Omega(H_{\mathrm{c}}^{-1} m_{\mathrm{c}}))$, we have $|Q_{\omega_0}| \le 2$. Also, from the fact that $|\mathcal{R}(s,a)| \le \mathcal{R}_{\max}$, we know that $|Q^*| \le \mathcal{R}_{\max}$. Therefore, for any measure ρ, we have

$$\begin{aligned}
&\Big|\mathbb{E}_{\rho}\Big[(I - \gamma \mathbb{P}^{\pi^*})^{-1} \sum_{k=0}^{K} (\gamma \mathbb{P}^{\pi^*})^{k+1} (Q^* - Q_{\omega_0})\Big]\Big| \\
&\qquad \le \mathbb{E}_{\rho}\Big[(I - \gamma \mathbb{P}^{\pi^*})^{-1} \sum_{k=0}^{K} (\gamma \mathbb{P}^{\pi^*})^{k+1} |Q^* - Q_{\omega_0}|\Big] \\
&\qquad \le \mathcal{R}_{\max}(1-\gamma)^{-1} \sum_{k=0}^{K} \gamma^{k+1} \le \mathcal{R}_{\max}(1-\gamma)^{-2}.
\end{aligned} \tag{A.3.21}$$

Also, by changing the index of summation, we have

$$
\begin{aligned}
&\Big|\mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\sum_{i=0}^{k}(\gamma\mathbb{P}^{\pi^*})^{k-i}\gamma\beta\mathbb{P}\epsilon_{i+1}^{\mathrm{a}}\Big]\Big| \\
&\quad = \Big|\mathbb{E}_\rho\Big[\sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{j=0}^{\infty}(\gamma\mathbb{P}^{\pi^*})^{k-i+j}\gamma\beta\mathbb{P}\epsilon_{i+1}^{\mathrm{a}}\Big]\Big| \\
&\quad = \Big|\mathbb{E}_\rho\Big[\sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{t=k-i}^{\infty}(\gamma\mathbb{P}^{\pi^*})^{t}\gamma\beta\mathbb{P}\epsilon_{i+1}^{\mathrm{a}}\Big]\Big| \\
&\quad \le \sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{t=k-i}^{\infty}\big|\mathbb{E}_\rho\big[(\gamma\mathbb{P}^{\pi^*})^{t}\gamma\beta\mathbb{P}\epsilon_{i+1}^{\mathrm{a}}\big]\big|, \qquad \text{(A.3.22)}
\end{aligned}
$$

where we expand $(I-\gamma\mathbb{P}^{\pi^*})^{-1}$ into an infinite sum in the first equality. Further, by changing the measure of the expectation on the RHS of (A.3.22), we have

$$
\sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{t=k-i}^{\infty}\big|\mathbb{E}_\rho\big[(\gamma\mathbb{P}^{\pi^*})^{t}\gamma\beta\mathbb{P}\epsilon_{i+1}^{\mathrm{a}}\big]\big| \le \sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{t=k-i}^{\infty}\beta\gamma^{t+1}c(t)\cdot\mathbb{E}_{\nu^*}[|\epsilon_{i+1}^{\mathrm{A}}|], \qquad \text{(A.3.23)}
$$

where $c(t)$ is defined in Assumption A.2.1. Further, by Lemma A.3.7 and interchanging the summation on the RHS of (A.3.23), we have

$$
\begin{aligned}
&\Big|\mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\sum_{i=0}^{k}(\gamma\mathbb{P}^{\pi^*})^{k-i}\gamma\beta\mathbb{P}\epsilon_{i+1}^{\mathrm{a}}\Big]\Big| \\
&\quad \le 2\sum_{k=0}^{K}\sum_{t=0}^{\infty}\sum_{i=\max\{0,k-t\}}^{k}\beta\gamma^{t+1}c(t)\cdot\tau_{i+1}^{-1}\varepsilon_f(\phi_i^*+\psi_i^*) \\
&\quad \le \sum_{k=0}^{K}\sum_{t=0}^{\infty}4kt\gamma^{t+1}c(t)\cdot\varepsilon_f(\phi^*+\psi^*) \\
&\quad \le \gamma\sum_{k=0}^{K}4C_{\rho,\rho^*}\cdot\varepsilon_f(\phi^*+\psi^*) \le 2\gamma K^2 C_{\rho,\rho^*}(\phi^*+\psi^*)\cdot\varepsilon_f, \qquad \text{(A.3.24)}
\end{aligned}
$$

where $\varepsilon_f = \max_i \mathbb{E}_{\rho_i}[(f_{\theta_{i+1}}(s,a) - \tau_{i+1}\cdot(\beta^{-1}Q_{\omega_i}(s,a) - \tau_i^{-1} f_{\theta_i}(s,a)))^2]^{1/2}$, and C_{ρ,ρ^*} is defined in Assumption A.2.1. Here in the second inequality, we use the fact that $\tau_{i+1}^{-1} = (i+1)\cdot\beta^{-1}$, and $\phi_i^* \le \phi^*$ and $\psi_i^* \le \psi^*$ by Assumption A.2.1.

By similar arguments in the derivation of (A.3.24), we have

$$\begin{aligned}
&\Big|\mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\sum_{i=0}^{k-1}(\gamma\mathbb{P}^{\pi^*})^{k-i}\epsilon_{i+1}^{\mathrm{c}}\Big]\Big| \le 2(K+1)C_{\rho,\rho^*}\phi^*\cdot\varepsilon_Q,\\
&\Big|\mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{i=0}^{K}(\gamma\mathbb{P}^{\pi^*})^{i}\gamma\beta\mathbb{P}\mathrm{KL}(\pi^*\,\|\,\pi_{\theta_0})\Big]\Big| \le \log|\mathcal{A}|\cdot K^{1/2}(1-\gamma)^{-2},\\
&\mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\gamma^{k+1}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\Big(\prod_{s=1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Big)e_1\Big] \le (2+\mathcal{R}_{\max})\cdot(1-\gamma)^{-3},
\end{aligned}\tag{A.3.25}$$

where $\varepsilon_Q = \max_i \mathbb{E}_{\rho^*}[|\epsilon_{i+1}^{\mathrm{c}}|]$. And we use the fact that $\beta = K^{1/2}$.

Now, it remains to upper bound the last term on the RHS of (A.3.19). We first consider the terms involving $\epsilon_{\ell+1}^{\mathrm{b}}$. We have

$$\begin{aligned}
&\mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\sum_{\ell=1}^{k}\gamma^{k-\ell+1}\Big(\prod_{s=\ell+1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Big)\beta\gamma\mathbb{P}\epsilon_{\ell+1}^{\mathrm{b}}\Big]\\
&\quad=\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\sum_{k=0}^{K}\sum_{\ell=1}^{k}\mathbb{E}_\rho\Big[(\gamma\mathbb{P}^{\pi^*})^{j}(\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{i+1}\gamma^{k-\ell}\Big(\prod_{s=\ell+1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Big)\beta\gamma\mathbb{P}\epsilon_{\ell+1}^{\mathrm{b}}\Big]\\
&\quad\le\beta\gamma\sum_{k=0}^{K}\sum_{\ell=1}^{k}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\gamma^{i+j+k-\ell+1}\cdot\mathbb{E}_{\rho^*}[|\mathbb{P}\epsilon_{\ell+1}^{\mathrm{b}}|]\cdot c(i+j+k-\ell+1)\\
&\quad\le 2\gamma\sum_{k=0}^{K}\sum_{\ell=1}^{k}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\gamma^{i+j+k-\ell+1}\cdot(\ell+1)\varepsilon_f\cdot(1+\psi_\ell^*)\cdot c(i+j+k-\ell+1),
\end{aligned}\tag{A.3.26}$$

where we expand $(I - \gamma \mathbb{P}^{\pi^*})^{-1}$ and $(I - \gamma \mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}$ to infinite sums in the first equality, change the measure of the expectation in the first inequality, and use Lemma A.3.7 in the last inequality. Now, by changing the index of the summation, we have

$$\begin{aligned}
&\gamma \sum_{k=0}^{K} \sum_{\ell=1}^{k} \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \gamma^{i+j+k-\ell+1} \cdot (\ell+1)\varepsilon_f \cdot (1+\psi_\ell^*) \cdot c(i+j+k-\ell+1) \\
&\qquad = \gamma \sum_{k=0}^{K} \sum_{\ell=1}^{k} \sum_{j=0}^{\infty} \sum_{t=j+k-\ell+1}^{\infty} \gamma^t \cdot (\ell+1)\varepsilon_f \cdot (1+\psi_\ell^*) \cdot c(t) \\
&\qquad \leq \gamma \sum_{k=0}^{K} \sum_{j=0}^{\infty} \sum_{t=j+1}^{\infty} \sum_{\ell=\max\{0,j+k-t+1\}}^{k} \gamma^t \cdot (\ell+1)\varepsilon_f \cdot (1+\psi^*) \cdot c(t), \qquad \text{(A.3.27)}
\end{aligned}$$

where we use the fact that $\psi_\ell^* \leq \psi^*$ from Assumption A.2.1 in the last inequality. By further manipulating the order of summations of the RHS of (A.3.27), we have

$$\begin{aligned}
&\gamma \sum_{k=0}^{K} \sum_{j=0}^{\infty} \sum_{t=j+1}^{\infty} \sum_{\ell=\max\{0,j+k-t+1\}}^{k} \gamma^t \cdot (\ell+1)\varepsilon_f(1+\psi^*) \cdot c(t) \\
&\quad \leq \gamma \sum_{k=0}^{K} \sum_{j=0}^{\infty} \Big(\sum_{t=j+1}^{j+k+1} (t-j)(2k+j-k+1) \cdot \gamma^t c(t) + \sum_{t=j+k+2}^{\infty} k^2 \cdot \gamma^t c(t) \Big) \cdot \varepsilon_f(1+\psi^*) \\
&\quad = \gamma \sum_{k=0}^{K} \Big(\sum_{t=1}^{\infty} \sum_{j=\max\{0,t-k-1\}}^{t-1} (t-j)(2k+j-k+1) \cdot \gamma^t c(t) \\
&\qquad\qquad + \sum_{t=k+2}^{\infty} \sum_{j=1}^{t-k-2} k^2 \cdot \gamma^t c(t) \Big) \cdot \varepsilon_f(1+\psi^*) \\
&\quad \leq 20\gamma \sum_{k=0}^{K} \Big(\sum_{t=1}^{\infty} k^2 \cdot t\gamma^t c(t) + \sum_{t=1}^{\infty} k^2 \cdot t\gamma^t c(t) \Big) \cdot \varepsilon_f(1+\psi^*)
\end{aligned}$$

(A.3.28)

$$\leq 20\gamma K \cdot C_{\rho,\rho^*} \cdot \varepsilon_f(1+\psi^*),$$

where we use the definition of C_{ρ,ρ^*} from Assumption A.2.1 in the last inequality. Now, combining (A.3.26), (A.3.27), and (A.3.28), we have

$$\mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\sum_{\ell=1}^{k}\gamma^{k-\ell+1}\Big(\prod_{s=\ell+1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Big)\beta\gamma\mathbb{P}\epsilon_{\ell+1}^{\mathrm{b}}\Big]$$

$$\leq 20\gamma K\cdot C_{\rho,\rho^*}\cdot\varepsilon_f\cdot(1+\psi^*). \tag{A.3.29}$$

Following from similar arguments when deriving (A.3.29), we have

$$\mathbb{E}_\rho\Big[(I-\gamma\mathbb{P}^{\pi^*})^{-1}\sum_{k=0}^{K}\mathbb{P}^{\pi_{\theta_{k+1}}}(I-\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1}\sum_{\ell=1}^{k}\gamma^{k-\ell+1}\Big(\prod_{s=\ell+1}^{k}\mathbb{P}^{\pi_{\theta_s}}\Big)(I-\gamma\mathbb{P}^{\pi_{\theta_\ell}})\epsilon_{\ell}^{\mathrm{c}}\Big]$$

(A.3.30)

$$\leq 20K\cdot C_{\rho,\rho^*}\phi^*\cdot\varepsilon_Q,$$

Now, by plugging (A.3.21), (A.3.24), (A.3.25), (A.3.29), and (A.3.30) into (A.3.19), with probability at least $1-O(H_{\mathrm{c}})\exp(-\Omega(H_{\mathrm{c}}^{-1}m_{\mathrm{c}}))$, we have

(A.3.31)

$$\mathbb{E}_\rho\Big[\sum_{k=0}^{K}Q^*(s,a)-Q^{\pi_{\theta_{k+1}}}(s,a)\Big]$$

$$\leq 2\log|\mathcal{A}|\cdot K^{1/2}(1-\gamma)^{-3}+60K^2C_{\rho,\rho^*}(\phi^*+\psi^*+1)\cdot\varepsilon_f+50KC_{\rho,\rho^*}\phi^*\cdot\varepsilon_Q.$$

Meanwhile, following from Propositions A.2.3 and A.2.4, it holds with probability at least $1 - 1/K$ that

$$\varepsilon_f = O\big(R_{\mathrm{a}} N_{\mathrm{a}}^{-1/4} + R_{\mathrm{a}}^{4/3} m_{\mathrm{a}}^{-1/12} H_{\mathrm{a}}^{7/2} (\log m_{\mathrm{a}})^{1/2}\big),$$

$$\varepsilon_Q = O\big(R_{\mathrm{c}} N_{\mathrm{c}}^{-1/4} + R_{\mathrm{c}}^{4/3} m_{\mathrm{c}}^{-1/12} H_{\mathrm{c}}^{7/2} (\log m_{\mathrm{c}})^{1/2}\big). \tag{A.3.32}$$

Combining (A.3.31), (A.3.32), and the choices of parameters stated in the theorem, it holds with probability at least $1 - 1/K$ that

$$\mathbb{E}_\rho\Big[\sum_{k=0}^{K} Q^*(s,a) - Q^{\pi_{\theta_{k+1}}}(s,a)\Big] \le \big(2(1-\gamma)^{-3}\log|\mathcal{A}| + O(1)\big)\cdot K^{1/2},$$

which concludes the proof of Theorem A.2.5.

A.4. Supporting Results

In this section, we provide some supporting results in the proof of Theorems 2.3.4 and A.2.5. We introduce Lemma A.4.1, which applies to both Algorithms 1 and 4. To introduce Lemma A.4.1, for any policy π and action-value function Q, we define $\widetilde{\pi}(a \,|\, s) \propto \exp(\beta^{-1}Q(s,a)) \cdot \pi(a \,|\, s)$.

Lemma A.4.1. For any $s \in \mathcal{S}$ and $\pi^\dagger$, we have

$$\begin{aligned}\beta^{-1}\cdot\langle Q(s,\cdot), \pi^\dagger(\cdot\,|\,s) - \widetilde{\pi}(\cdot\,|\,s)\rangle \le \mathrm{KL}\big(\pi^\dagger(\cdot\,|\,s)\,\|\,\pi(\cdot\,|\,s)\big) - \mathrm{KL}\big(\pi^\dagger(\cdot\,|\,s)\,\|\,\widetilde{\pi}(\cdot\,|\,s)\big)\\ + \big\langle \log\big(\widetilde{\pi}(\cdot\,|\,s)/\pi(\cdot\,|\,s)\big) - \beta^{-1}\cdot Q(s,\cdot), \pi^\dagger(\cdot\,|\,s) - \widetilde{\pi}(\cdot\,|\,s)\big\rangle.\end{aligned}$$

Proof. By calculation, it suffices to show that

$$\begin{aligned}&\bigl\langle \log(\widetilde{\pi}(\cdot\,|\,s)/\pi(\cdot\,|\,s)), \pi^\dagger(\cdot\,|\,s) - \widetilde{\pi}(\cdot\,|\,s)\bigr\rangle \\ &\qquad \le \mathrm{KL}(\pi^\dagger(\cdot\,|\,s)\,\|\,\pi(\cdot\,|\,s)) - \mathrm{KL}(\pi^\dagger(\cdot\,|\,s)\,\|\,\widetilde{\pi}(\cdot\,|\,s)).\end{aligned}$$

By the definition of the KL divergence, it holds for any $s \in \mathcal{S}$ that

$$\begin{aligned}&\mathrm{KL}(\pi^\dagger(\cdot\,|\,s)\,\|\,\pi(\cdot\,|\,s)) - \mathrm{KL}(\pi^\dagger(\cdot\,|\,s)\,\|\,\widetilde{\pi}(\cdot\,|\,s)) \\ &\qquad = \bigl\langle \log(\widetilde{\pi}(\cdot\,|\,s)/\pi(\cdot\,|\,s)), \pi^\dagger(\cdot\,|\,s)\bigr\rangle. \qquad \text{(A.4.1)}\end{aligned}$$

Meanwhile, for the term on the RHS of (A.4.1), we have

$$\begin{aligned}&\bigl\langle \log(\widetilde{\pi}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)), \pi^\dagger(\cdot\,|\,s)\bigr\rangle \\ &\quad = \bigl\langle \log(\widetilde{\pi}(\cdot\,|\,s)/\pi(\cdot\,|\,s)), \pi^\dagger(\cdot\,|\,s) - \widetilde{\pi}(\cdot\,|\,s)\bigr\rangle \\ &\qquad + \bigl\langle \log(\widetilde{\pi}(\cdot\,|\,s)/\pi(\cdot\,|\,s)), \widetilde{\pi}(\cdot\,|\,s)\bigr\rangle \\ &\quad = \bigl\langle \log(\widetilde{\pi}(\cdot\,|\,s)/\pi(\cdot\,|\,s)), \pi^\dagger(\cdot\,|\,s) - \widetilde{\pi}(\cdot\,|\,s)\bigr\rangle + \mathrm{KL}(\widetilde{\pi}(\cdot\,|\,s)\,\|\,\pi(\cdot\,|\,s)) \\ &\quad \ge \bigl\langle \log(\widetilde{\pi}(\cdot\,|\,s)/\pi(\cdot\,|\,s)), \pi^\dagger(\cdot\,|\,s) - \widetilde{\pi}(\cdot\,|\,s)\bigr\rangle. \qquad \text{(A.4.2)}\end{aligned}$$

Combining (A.4.1) and (A.4.2), we obtain that

$$\begin{aligned}&\bigl\langle \log(\widetilde{\pi}(\cdot\,|\,s)/\pi(\cdot\,|\,s)), \pi^\dagger(\cdot\,|\,s) - \widetilde{\pi}(\cdot\,|\,s)\bigr\rangle \\ &\qquad \le \mathrm{KL}(\pi^\dagger(\cdot\,|\,s)\,\|\,\pi(\cdot\,|\,s)) - \mathrm{KL}(\pi^\dagger(\cdot\,|\,s)\,\|\,\widetilde{\pi}(\cdot\,|\,s)),\end{aligned}$$

which concludes the proof of Lemma A.4.1. □

A.4.1. Local Linearization of DNNs

In the proofs of Propositions A.2.3 and A.2.4 in §A.5.2 and §A.5.3, respectively, we utilize the linearization of DNNs. We introduce some related auxiliary results here. First, we define the linearization $\bar{u}_\theta$ of the DNN $u_\theta \in \mathcal{U}(w, H, R)$ as follows,

$$\bar{u}_\theta(\cdot) = u_{\theta_0}(\cdot) + (\theta - \theta_0)^\top \nabla_{\theta_0} u_\theta(\cdot),$$

where θ_0 is the initialization of u_θ. The following lemmas characterize the linearization error.

Lemma A.4.2. Suppose that $H = O(m^{1/12} R^{-1/6} (\log m)^{-1/2})$ and $m = \Omega(d^{3/2} R^{-1} H^{-3/2} \cdot \log(m^{1/2}/R)^{3/2})$. Then with probability at least $1 - \exp(-\Omega(R^{2/3} m^{2/3} H))$ over the random initialization θ_0, it holds for any $\theta \in \mathcal{B}(\theta_0, R)$ and any $(s, a) \in \mathcal{S} \times \mathcal{A}$ that

$$\|\nabla_\theta u_\theta(s, a) - \nabla_\theta u_{\theta_0}(s, a)\|_2 = O\big(R^{1/3} m^{-1/6} H^{5/2} (\log m)^{1/2}\big)$$

and

$$\|\nabla_\theta u_\theta(s, a)\|_2 = O(H).$$

Proof. See the proof of Lemma A.5 in Gao et al. (2019) for a detailed proof. □

Lemma A.4.3. Suppose that $H = O(m^{1/12} R^{-1/6} (\log m)^{-1/2})$ and $m = \Omega(d^{3/2} R^{-1} H^{-3/2} \cdot \log(m^{1/2}/R)^{3/2})$. Then with probability at least $1 - \exp(-\Omega(R^{2/3} m^{2/3} H))$ over the random initialization θ_0, it holds for any $\theta \in \mathcal{B}(\theta_0, R)$ and any $(s, a) \in \mathcal{S} \times \mathcal{A}$ that

$$|u_\theta(s, a) - \bar{u}_\theta(s, a)| = O\big(R^{4/3} m^{-1/6} H^{5/2} (\log m)^{1/2}\big).$$

Proof. Recall that

$$\bar{u}_\theta(s,a) = u_{\theta_0}(s,a) + (\theta - \theta_0)^\top \nabla_\theta u_{\theta_0}(s,a).$$

By mean value theorem, there exists $t \in [0,1]$, which depends on θ and (s,a), such that

$$u_\theta(s,a) - \bar{u}_\theta(s,a) = (\theta - \theta_0)^\top \big(\nabla_\theta u_{\theta_0 + t(\theta-\theta_0)}(s,a) - \nabla_\theta u_{\theta_0}(s,a)\big).$$

Further by Lemma A.4.2, we have

$$\begin{aligned} |u_\theta(s,a) - \bar{u}_\theta(s,a)| &\leq \|\theta - \theta_0\|_2 \cdot \big\|\nabla_\theta u_{\theta_0 + t\cdot(\theta-\theta_0)}(s,a) - \nabla_\theta u_{\theta_0}(s,a)\big\|_2 \\ &= O\big(R^{4/3} m^{-1/6} H^{5/2} (\log m)^{1/2}\big), \end{aligned}$$

where we use Cauchy-Schwarz inequality in the first inequality. This concludes the proof of Lemma A.4.3. □

We denote by $x^{(h)}$ the output of the h-th layer of the DNN $u_\theta \in \mathcal{U}(m,H,R)$, and $x^{(h),0}$ the output of the h-th layer of the DNN $u_{\theta_0} \in \mathcal{U}(m,H,R)$. The following lemma upper bounds the distance between $x^{(h)}$ and $x^{(h),0}$.

Lemma A.4.4. With probability at least $1 - \exp(-\Omega(R^{2/3} m^{2/3} H))$ over the random initialization θ_0, for any $\theta \in \mathcal{B}(\theta_0, R)$ and any $h \in [H]$, we have

$$\|x^{(h)} - x^{(h),0}\|_2 = O\big(R H^{5/2} m^{-1/2} (\log m)^{1/2}\big).$$

Also, with probability at least $1 - O(H)\exp(-\Omega(H^{-1}m))$ over the random initialization θ_0, for any $\theta \in \mathcal{B}(\theta_0, R)$ and any $h \in [H]$, it holds that

$$2/3 \leq \|x^{(h)}\|_2 \leq 4/3.$$

Proof. The first inequality follows from Lemma A.5 in Gao et al. (2019), and the second inequality follows from Lemma 7.1 in Allen-Zhu et al. (2018b). □

A.5. Proofs of Propositions

A.5.1. Proof of Proposition 2.2.1

The proof follows the proof of Proposition 3.1 in Liu et al. (2019). First, we write the update $\widetilde{\pi}_{k+1} \leftarrow \mathrm{argmax}_\pi \, \mathbb{E}_{\nu_k}[\langle Q_{\omega_k}(s,\cdot), \pi(\cdot \,|\, s)\rangle - \beta \cdot \mathrm{KL}(\pi(\cdot \,|\, s) \,\|\, \pi_{\theta_k}(\cdot \,|\, s))]$ as a constrained optimization problem in the following way,

$$\begin{aligned} &\max_\pi \; \mathbb{E}_{\nu_k}\big[\langle \pi(\cdot \,|\, s), Q_{\omega_k}(s,\cdot)\rangle - \beta \cdot \mathrm{KL}(\pi(\cdot \,|\, s) \,\|\, \pi_{\theta_k}(\cdot \,|\, s))\big] \\ &\text{s.t.} \sum_{a \in \mathcal{A}} \pi(a \,|\, s) = 1, \qquad \text{for any } s \in \mathcal{S}. \end{aligned}$$

We consider the Lagrangian of the above programming problem,

$$\begin{aligned} &\int_{s \in \mathcal{S}} \Big(\langle \pi(\cdot \,|\, s), Q_{\omega_k}(s,\cdot)\rangle - \beta \cdot \mathrm{KL}\big(\pi(\cdot \,|\, s) \,\|\, \pi_{\theta_k}(\cdot \,|\, s)\big)\Big) \mathrm{d}\nu_k(s) \\ &\qquad + \int_{s \in \mathcal{S}} \Big(\sum_{a \in \mathcal{A}} \pi(a \,|\, s) - 1\Big) \mathrm{d}\lambda(s), \end{aligned}$$

where $\lambda(\cdot)$ is the dual parameter, which is a function on $\mathcal{S}$. Now, by plugging in

$$\pi_{\theta_k}(a \,|\, s) = \frac{\exp(\tau_k^{-1} f_{\theta_k}(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\tau_k^{-1} f_{\theta_k}(s,a'))},$$

we have the following optimality condition,

$$Q_{\omega_k}(s,a) + \beta \tau_k^{-1} f_{\theta_k}(s,a) - \beta \cdot \Big(\log\Big(\sum_{a' \in \mathcal{A}} \exp(\tau_k^{-1} f_{\theta_k}(s,a')) \Big) + \log \pi(a \,|s) + 1 \Big) + \frac{\lambda(s)}{\nu_k(s)}$$
$$= 0,$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Note that $\log(\sum_{a' \in \mathcal{A}} \exp(\tau_k^{-1} f_{\theta_k}(s,a')))$ is only a function of s. Thus, we have

$$\widehat{\pi}_{k+1}(a \,|\, s) \propto \exp(\beta^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a))$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, which concludes the proof of Proposition 2.2.1.

A.5.2. Proof of Proposition A.2.3

We define the local linearization of f_θ as follows,

$$\bar{f}_\theta = f_{\theta_0} + (\theta - \theta_0)^\top \nabla_{\theta_0} f_\theta. \qquad \text{(A.5.1)}$$

Meanwhile, we denote by

$$g_n = \big(f_{\theta(n)} - \widetilde{\tau}\cdot(\beta^{-1}Q_\omega + \tau^{-1}f_\theta)\big)\cdot\nabla_\theta f_{\theta(n)}, \qquad g_n^e = \mathbb{E}_{\rho_{\pi_\theta}}[g_n],$$

$$\bar{g}_n = \big(\bar{f}_{\theta(n)} - \widetilde{\tau}\cdot(\beta^{-1}Q_\omega + \tau^{-1}f_\theta)\big)\cdot\nabla_\theta f_{\theta_0}, \qquad \bar{g}_n^e = \mathbb{E}_{\rho_{\pi_\theta}}[\bar{g}_n],$$

$$g_* = \big(f_{\theta_*} - \widetilde{\tau}\cdot(\beta^{-1}Q_\omega + \tau^{-1}f_\theta)\big)\cdot\nabla_\theta f_{\theta_*}, \qquad g_*^e = \mathbb{E}_{\rho_{\pi_\theta}}[g_*],$$

$$\bar{g}_* = \big(\bar{f}_{\theta_*} - \widetilde{\tau}\cdot(\beta^{-1}Q_\omega + \tau^{-1}f_\theta)\big)\cdot\nabla_\theta f_{\theta_0}, \qquad \bar{g}_*^e = \mathbb{E}_{\rho_{\pi_\theta}}[\bar{g}_*], \tag{A.5.2}$$

where θ_* satisfies that

$$\theta_* = \Gamma_{\mathcal{B}(\theta_0,R_{\mathrm{a}})}(\theta_* - \alpha\cdot\bar{g}_*^e). \tag{A.5.3}$$

By Algorithm 5, we know that

$$\theta(n+1) = \Gamma_{\mathcal{B}(\theta_0,R_{\mathrm{a}})}(\theta(n) - \alpha\cdot g_n). \tag{A.5.4}$$

By (A.5.3) and (A.5.4), we have

$$\begin{aligned}
&\mathbb{E}_{\rho_{\pi_\theta}}\big[\|\theta(n+1)-\theta_*\|_2^2 \,|\, \theta(n)\big]\\
&\quad= \mathbb{E}_{\rho_{\pi_\theta}}\big[\|\Gamma_{\mathcal{B}(\theta_0,R_{\mathrm{a}})}(\theta(n)-\alpha\cdot g_n) - \Gamma_{\mathcal{B}(\theta_0,R_{\mathrm{a}})}(\theta_*-\alpha\cdot\bar{g}_*^e)\|_2^2 \,|\, \theta(n)\big]\\
&\quad\le \mathbb{E}_{\rho_{\pi_\theta}}\big[\|(\theta(n)-\alpha\cdot g_n) - (\theta_*-\alpha\cdot\bar{g}_*^e)\|_2^2 \,|\, \theta(n)\big]\\
&\quad= \|\theta(n)-\theta_*\|_2^2 + 2\alpha\cdot\underbrace{\langle\theta_*-\theta(n), g_n^e-\bar{g}_*^e\rangle}_{\text{(i)}} + \alpha^2\cdot\underbrace{\mathbb{E}_{\rho_{\pi_\theta}}\big[\|g_n-\bar{g}_*^e\|_2^2 \,|\, \theta(n)\big]}_{\text{(ii)}},
\end{aligned} \tag{A.5.5}$$

where we use the fact that $\Gamma_{\mathcal{B}(\theta_0,R_{\mathrm{a}})}$ is a contraction mapping in the first inequality. We upper bound term (i) and term (ii) on the RHS of (A.5.5) in the sequel.

Upper Bound of Term (i). By Cauchy–Schwarz inequality, it holds that

$$
\begin{aligned}
\langle \theta_* - \theta(n), g_n^e - \bar{g}_*^e \rangle &= \langle \theta_* - \theta(n), g_n^e - \bar{g}_n^e \rangle + \langle \theta_* - \theta(n), \bar{g}_n^e - \bar{g}_*^e \rangle \\
&\le \|\theta_* - \theta(n)\|_2 \cdot \|g_n^e - \bar{g}_n^e\|_2 + \langle \theta_* - \theta(n), \bar{g}_n^e - \bar{g}_*^e \rangle \\
&\le 2R_{\mathrm{a}} \cdot \|g_n^e - \bar{g}_n^e\|_2 + \langle \theta_* - \theta(n), \bar{g}_n^e - \bar{g}_*^e \rangle, \qquad \text{(A.5.6)}
\end{aligned}
$$

where we use the fact that $\theta(n), \theta_* \in \mathcal{B}(\theta_0, R_{\mathrm{a}})$ in the last inequality. Further, by the definitions in (A.5.2), it holds that

$$
\begin{aligned}
\langle \theta_* - \theta(n), \bar{g}_n^e - \bar{g}_*^e \rangle &= \mathbb{E}_{\rho_{\pi_\theta}} \big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*}) \cdot \langle \theta_* - \theta(n), \nabla_\theta f_{\theta_0} \rangle \big] \\
&= \mathbb{E}_{\rho_{\pi_\theta}} \big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*}) \cdot (\bar{f}_{\theta_*} - \bar{f}_{\theta(n)}) \big] \\
&= -\mathbb{E}_{\rho_{\pi_\theta}} \big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2 \big], \qquad \text{(A.5.7)}
\end{aligned}
$$

where we use (A.5.1) in the second equality. Combining (A.5.6) and (A.5.7), we obtain the following upper bound of term (i),

$$
\langle \theta_* - \theta(n), g_n^e - \bar{g}_*^e \rangle \le 2R_{\mathrm{a}} \cdot \|g_n^e - \bar{g}_n^e\|_2 - \mathbb{E}_{\rho_{\pi_\theta}} \big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2 \big]. \qquad \text{(A.5.8)}
$$

Upper Bound of Term (ii). We now upper bound term (ii) on the RHS of (A.5.5). It holds by Cauchy-Schwarz inequality that

$$
\begin{aligned}
\mathbb{E}_{\rho_{\pi_\theta}} \big[\|g_n - \bar{g}_*^e\|_2^2 \,|\, \theta(n) \big] &\le 2\mathbb{E}_{\rho_{\pi_\theta}} \big[\|g_n - g_n^e\|_2^2 \,|\, \theta(n) \big] + 2\|g_n^e - \bar{g}_*^e\|_2^2 \\
&\le 2 \underbrace{\mathbb{E}_{\rho_{\pi_\theta}} \big[\|g_n - g_n^e\|_2^2 \,|\, \theta(n) \big]}_{\text{(ii).a}} + 4 \underbrace{\|g_n^e - \bar{g}_n^e\|_2^2}_{\text{(ii).b}} + 4 \underbrace{\|\bar{g}_n^e - \bar{g}_*^e\|_2^2}_{\text{(ii).c}}. \qquad \text{(A.5.9)}
\end{aligned}
$$

We upper bound term (ii).a, term (ii).b, and term (ii).c in the sequel.

Upper Bound of Term (ii).a. Note that

$$\mathbb{E}_{\rho_{\pi_\theta}}\big[\|g_n - g_n^e\|_2^2 \,|\, \theta(n)\big] = \mathbb{E}_{\rho_{\pi_\theta}}\big[\|g_n\|_2^2 - \|g_n^e\|_2^2 \,|\, \theta(n)\big] \le \mathbb{E}_{\rho_{\pi_\theta}}\big[\|g_n\|_2^2 \,|\, \theta(n)\big]. \tag{A.5.10}$$

Meanwhile, by the definition of g_n in (A.5.2), it holds that

$$\|g_n\|_2^2 = \big(f_{\theta(n)} - \widetilde{\tau} \cdot (\beta^{-1} Q_\omega + \tau^{-1} f_\theta)\big)^2 \cdot \|\nabla_\theta f_{\theta(n)}\|_2^2. \tag{A.5.11}$$

We first upper bound f_θ as follows,

$$f_\theta^2 = x^{(H_\mathrm{a})\top} b b^\top x^{(H_\mathrm{a})} = x^{(H_\mathrm{a})\top} x^{(H_\mathrm{a})} = \|x^{(H_\mathrm{a})}\|_2^2,$$

where $x^{(H_\mathrm{a})}$ is the output of the H_a-th layer of the DNN f_θ. Further combining Lemma A.4.4, it holds with probability at least $1 - O(H_\mathrm{a}) \exp(-\Omega(H_\mathrm{a}^{-1} m_\mathrm{a}))$ that

$$|f_\theta| \le 2. \tag{A.5.12}$$

Following from similar arguments, with probability at least $1 - O(H_\mathrm{a}) \exp(-\Omega(H_\mathrm{a}^{-1} m_\mathrm{a}))$, we have

$$|Q_\omega| \le 2, \qquad |f_{\theta(n)}| \le 2. \tag{A.5.13}$$

Combining Lemma A.4.2, (A.5.10), (A.5.11), (A.5.12), and (A.5.13), it holds with probability at least $1 - \exp(-\Omega(R_\mathrm{a}^{2/3} m_\mathrm{a}^{2/3} H_\mathrm{a}))$ that

$$\mathbb{E}_{\rho_{\pi_\theta}}\big[\|g_n - g_n^e\|_2^2 \,|\, \theta(n)\big] = O(H_\mathrm{a}^2), \tag{A.5.14}$$

which establishes an upper bound of term (ii).a.

Upper Bound of Term (ii).b. It holds that

$$\begin{aligned}
\|g_n^e - \bar{g}_n^e\|_2 &= \Big\|\mathbb{E}_{\rho_{\pi_\theta}}\big[\big(f_{\theta(n)} - \widetilde{\tau}\cdot(\beta^{-1}Q_\omega + \tau^{-1}f_\theta)\big)\cdot\nabla_\theta f_{\theta(n)} \\
&\qquad\qquad - \big(\bar{f}_{\theta(n)} - \widetilde{\tau}\cdot(\beta^{-1}Q_\omega + \tau^{-1}f_\theta)\big)\cdot\nabla_\theta f_{\theta_0}\big]\Big\|_2 \\
&\le \mathbb{E}_{\rho_{\pi_\theta}}\big[\|f_{\theta(n)}\nabla_\theta f_{\theta(n)} - \bar{f}_{\theta(n)}\nabla_\theta f_{\theta_0}\|_2\big] \\
&\quad + \widetilde{\tau}\cdot\mathbb{E}_{\rho_{\pi_\theta}}\big[\|(\beta^{-1}Q_\omega + \tau^{-1}f_\theta)\cdot(\nabla_\theta f_{\theta_0} - \nabla_\theta f_{\theta(n)})\|_2\big] \\
&\le \mathbb{E}_{\rho_{\pi_\theta}}\big[\|f_{\theta(n)}\nabla_\theta f_{\theta_0} - \bar{f}_{\theta(n)}\nabla_\theta f_{\theta_0}\|_2\big] + \mathbb{E}_{\rho_{\pi_\theta}}\big[\|f_{\theta(n)}\nabla_\theta f_{\theta(n)} - f_{\theta(n)}\nabla_\theta f_{\theta_0}\|_2\big] \\
&\quad + \mathbb{E}_{\rho_{\pi_\theta}}\big[\|\widetilde{\tau}\cdot(\beta^{-1}Q_\omega + \tau^{-1}f_\theta)\cdot(\nabla_\theta f_{\theta_0} - \nabla_\theta f_{\theta(n)})\|_2\big].
\end{aligned} \tag{A.5.15}$$

We upper bound the three terms on the RHS of (A.5.15) in the sequel, respectively.

For the term $\|f_{\theta(n)}\nabla_\theta f_{\theta_0} - \bar{f}_{\theta(n)}\nabla_\theta f_{\theta_0}\|_2$ on the RHS of (A.5.15), following from Lemmas A.4.2 and A.4.3, it holds with probability at least $1 - \exp(-\Omega(R_\mathrm{a}^{2/3}m_\mathrm{a}^{2/3}H_\mathrm{a}))$ that

$$\|f_{\theta(n)}\nabla_\theta f_{\theta_0} - \bar{f}_{\theta(n)}\nabla_\theta f_{\theta_0}\|_2 = O\big(R_\mathrm{a}^{4/3}m_\mathrm{a}^{-1/6}H_\mathrm{a}^{7/2}(\log m_\mathrm{a})^{1/2}\big). \tag{A.5.16}$$

For the term $\|f_{\theta(n)}\nabla_\theta f_{\theta(n)} - f_{\theta(n)}\nabla_\theta f_{\theta_0}\|_2$ on the RHS of (A.5.15), following from (A.5.13) and Lemma A.4.2, with probability at least $1 - \exp(-\Omega(R_\mathrm{a}^{2/3}m_\mathrm{a}^{2/3}H_\mathrm{a}))$, we have

$$\|f_{\theta(n)}\nabla_\theta f_{\theta(n)} - f_{\theta(n)}\nabla_\theta f_{\theta_0}\|_2 = O\big(R_\mathrm{a}^{1/3}m_\mathrm{a}^{-1/6}H_\mathrm{a}^{5/2}(\log m_\mathrm{a})^{1/2}\big). \tag{A.5.17}$$

For the term $\|\widetilde{\tau} \cdot (\beta^{-1}Q_\omega + \tau^{-1}f_\theta) \cdot (\nabla_\theta f_{\theta_0} - \nabla_\theta f_{\theta(n)})\|_2$ on the RHS of (A.5.15), we first upper bound $\widetilde{\tau} \cdot (\beta^{-1}Q_\omega + \tau^{-1}f_\theta)$ as follows,

$$|\widetilde{\tau} \cdot (\beta^{-1}Q_\omega + \tau^{-1}f_\theta)| \leq 2,$$

where we use (A.5.12), (A.5.13), and the fact that $\widetilde{\tau}^{-1} = \beta^{-1} + \tau^{-1}$. Further combining Lemma A.4.2, it holds with probability at least $1 - \exp(-\Omega(R_{\mathrm{a}}^{2/3} m_{\mathrm{a}}^{2/3} H_{\mathrm{a}}))$ that

$$\|\widetilde{\tau} \cdot (\beta^{-1}Q_\omega + \tau^{-1}f_\theta) \cdot (\nabla_\theta f_{\theta_0} - \nabla_\theta f_{\theta(n)})\|_2 = O\big(R_{\mathrm{a}}^{1/3} m_{\mathrm{a}}^{-1/6} H_{\mathrm{a}}^{5/2} (\log m_{\mathrm{a}})^{1/2}\big). \tag{A.5.18}$$

Now, combining (A.5.15), (A.5.16), (A.5.17), and (A.5.18), it holds with probability at least $1 - \exp(-\Omega(R_{\mathrm{a}}^{2/3} m_{\mathrm{a}}^{2/3} H_{\mathrm{a}}))$ that

$$\|g_n^e - \bar{g}_n^e\|_2^2 = O\big(R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/3} H_{\mathrm{a}}^7 \log m_{\mathrm{a}}\big), \tag{A.5.19}$$

which establishes an upper bound of term (ii).b.

Upper Bound of Term (ii).c. It holds that

$$\|\bar{g}_n^e - \bar{g}_*^e\|_2^2 = \big\|\mathbb{E}_{\rho_{\pi_\theta}}[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})\nabla_\theta f_{\theta_0}]\big\|_2^2 \leq \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2 \cdot \|\nabla_\theta f_{\theta_0}\|_2^2\big].$$

Further combining Lemma A.4.2, it holds with probability at least $1 - \exp(-\Omega(R_{\mathrm{a}}^{2/3} m_{\mathrm{a}}^{2/3} H_{\mathrm{a}}))$ that

$$\|\bar{g}_n^e - \bar{g}_*^e\|_2^2 \leq O(H_{\mathrm{a}}^2) \cdot \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2\big], \tag{A.5.20}$$

which establishes an upper bound of term (ii).c.

Now, combining (A.5.9), (A.5.14), (A.5.19), and (A.5.20), we have

(A.5.21)

$$\mathbb{E}_{\rho_{\pi_\theta}}\big[\|g_n - \bar{g}_*^e\|_2^2 \,|\, \theta(n)\big] \le O\big(R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/3} H_{\mathrm{a}}^7 \log m_{\mathrm{a}}\big) + O(H_{\mathrm{a}}^2) \cdot \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2\big],$$

which is an upper bound of term (ii) on the RHS of (A.5.5).

By plugging the upper bound of term (i) in (A.5.8) and the upper bound of term (ii) in (A.5.21) into (A.5.5), combining (A.5.19), with probability at least $1-\exp(-\Omega(R_{\mathrm{a}}^{2/3} m_{\mathrm{a}}^{2/3} H_{\mathrm{a}}))$, we have

$$\mathbb{E}_{\rho_{\pi_\theta}}\big[\|\theta(n+1) - \theta_*\|_2^2 \,|\, \theta(n)\big]$$

(A.5.22)

$$\le \|\theta(n) - \theta_*\|_2^2 + 2\alpha \cdot \Big(O\big(R_{\mathrm{a}}^{7/3} m_{\mathrm{a}}^{-1/6} H_{\mathrm{a}}^{7/2} (\log m_{\mathrm{a}})^{1/2}\big) - \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2\big]\Big)$$
$$+ \alpha^2 \cdot \Big(O\big(R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/3} H_{\mathrm{a}}^7 \log m_{\mathrm{a}}\big) + O(H_{\mathrm{a}}^2) \cdot \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2\big]\Big).$$

Rearranging terms in (A.5.22), it holds with probability at least $1-\exp(-\Omega(R_{\mathrm{a}}^{2/3} m_{\mathrm{a}}^{2/3} H_{\mathrm{a}}))$ that

$$(2\alpha - \alpha^2 \cdot O(H_{\mathrm{a}}^2)) \cdot \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2\big]$$

(A.5.23)

$$\le \|\theta(n) - \theta_*\|_2^2 - \mathbb{E}_{\rho_{\pi_\theta}}\big[\|\theta(n+1) - \theta_*\|_2^2 \,|\, \theta(n)\big] + \alpha \cdot O\big(R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/6} H_{\mathrm{a}}^7 \log m_{\mathrm{a}}\big).$$

By telescoping the sum and using Jensen's inequality in (A.5.23), we have

$$
\begin{aligned}
\mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{f}_{\bar{\theta}} - \bar{f}_{\theta_*})^2\big] &\leq \frac{1}{N_{\mathrm{a}}} \cdot \sum_{n=0}^{N_{\mathrm{a}}-1} \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{f}_{\theta(n)} - \bar{f}_{\theta_*})^2\big] \\
&\leq 1/N_{\mathrm{a}} \cdot \big(2\alpha - \alpha^2 \cdot O(H_{\mathrm{a}}^2)\big)^{-1} \cdot \big(\|\theta_0 - \theta_*\|_2^2 + \alpha N_{\mathrm{a}} \cdot O(R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/6} H_{\mathrm{a}}^7 \log m_{\mathrm{a}})\big) \\
&\leq N_{\mathrm{a}}^{-1/2} \cdot \|\theta_0 - \theta_*\|_2^2 + O(R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/6} H_{\mathrm{a}}^7 \log m_{\mathrm{a}}),
\end{aligned}
$$

where the last line comes from the choices that $\alpha = N_{\mathrm{a}}^{-1/2}$ and $H_{\mathrm{a}} = O(N_{\mathrm{a}}^{1/4})$. Further combining Lemma A.4.3 and using triangle inequality, we have

$$
\mathbb{E}_{\rho_{\pi_\theta}}\big[(f_{\bar{\theta}} - \bar{f}_{\theta_*})^2\big] = O(R_{\mathrm{a}}^2 N_{\mathrm{a}}^{-1/2} + R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/6} H_{\mathrm{a}}^7 \log m_{\mathrm{a}}). \tag{A.5.24}
$$

By the definition of θ_* in (A.5.3), we know that

$$
\langle \bar{g}_*^e, \theta - \theta_* \rangle \geq 0, \qquad \text{for any } \theta \in \mathcal{B}(\theta_0, R_{\mathrm{a}}). \tag{A.5.25}
$$

By plugging the definition of $\bar{g}_*^e$ into (A.5.25), we have

$$
\mathbb{E}_{\rho_{\pi_\theta}}\big[\langle \bar{f}_{\theta_*} - \widetilde{\tau} \cdot (\beta^{-1} Q_\omega + \tau^{-1} f_\theta), \bar{f}_{\theta^\dagger} - \bar{f}_{\theta_*} \rangle\big] \geq 0, \qquad \text{for any } \theta^\dagger \in \mathcal{B}(\theta_0, R_{\mathrm{a}}),
$$

which is equivalent to

$$
\theta_* = \operatorname*{argmin}_{\theta^\dagger \in \mathcal{B}(\theta_0, R_{\mathrm{a}})} \mathbb{E}_{\rho_{\pi_\theta}}\big[\big(\bar{f}_{\theta^\dagger} - \widetilde{\tau} \cdot (\beta^{-1} Q_\omega + \tau^{-1} f_\theta)\big)^2\big]. \tag{A.5.26}
$$

Meanwhile, by the fact that $\theta_0 = \omega_0$, we have

$$\begin{aligned}
\widetilde{\tau} \cdot (\beta^{-1}\bar{Q}_\omega + \tau^{-1}\bar{f}_\theta) &= \widetilde{\tau} \cdot \big(\beta^{-1} \cdot (Q_{\omega_0} + (\omega - \omega_0)^\top \nabla_\omega Q_{\omega_0}) + \tau^{-1} \cdot (f_{\theta_0} + (\theta - \theta_0)^\top \nabla_\theta f_{\theta_0})\big) \\
&= f_{\theta_0} + \big(\widetilde{\tau} \cdot (\beta^{-1}\omega + \tau^{-1}\theta) - \theta_0\big)^\top \nabla_\theta f_{\theta_0},
\end{aligned}$$

where the second line comes from $\widetilde{\tau}^{-1} = \beta^{-1} + \tau^{-1}$. Note that $\theta \in \mathcal{B}(\theta_0, R_{\mathrm{a}})$, $\omega \in \mathcal{B}(\omega_0, R_{\mathrm{c}})$, $\theta_0 = \omega_0$, and $R_{\mathrm{a}} = R_{\mathrm{c}}$, we know that $\widetilde{\tau} \cdot (\beta^{-1}\omega + \tau^{-1}\theta) \in \mathcal{B}(\theta_0, R_{\mathrm{a}})$. Therefore, with probability at least $1 - \exp(-\Omega(R_{\mathrm{a}}^{2/3} m_{\mathrm{a}}^{2/3} H_{\mathrm{a}}))$ we have

$$\begin{aligned}
&\mathbb{E}_{\rho_{\pi_\theta}}\big[\big(\bar{f}_{\theta_*} - \widetilde{\tau} \cdot (\beta^{-1} Q_\omega + \tau^{-1} f_\theta)\big)^2\big] \\
&\quad \le \mathbb{E}_{\rho_{\pi_\theta}}\big[\big(\widetilde{\tau} \cdot (\beta^{-1}\bar{Q}_\omega + \tau^{-1}\bar{f}_\theta) - \widetilde{\tau} \cdot (\beta^{-1} Q_\omega + \tau^{-1} f_\theta)\big)^2\big] \\
&\quad \le \widetilde{\tau}^2 \cdot \beta^{-2} \cdot \mathbb{E}_{\rho_{\pi_\theta}}[(\bar{Q}_\omega - Q_\omega)^2] + \widetilde{\tau}^2 \cdot \tau^{-2} \cdot \mathbb{E}_{\rho_{\pi_\theta}}[(\bar{f}_\theta - f_\theta)^2] \\
&\quad = O(R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/3} H_{\mathrm{a}}^5 \log m_{\mathrm{a}}), \qquad \text{(A.5.27)}
\end{aligned}$$

where the first inequality comes from (A.5.26), and the last inequality comes from Lemma A.4.3 and the fact that $R_{\mathrm{c}} = R_{\mathrm{a}}$, $m_{\mathrm{c}} = m_{\mathrm{a}}$, and $H_{\mathrm{c}} = H_{\mathrm{a}}$. Combining (A.5.24) and (A.5.27), by triangle inequality, we have

$$\mathbb{E}_{\rho_{\pi_\theta}}\big[\big(f_{\overline{\theta}}(s,a) - \widetilde{\tau} \cdot (\beta^{-1} Q_\omega(s,a) + \tau^{-1} f_\theta(s,a))\big)^2\big] = O(R_{\mathrm{a}}^2 N_{\mathrm{a}}^{-1/2} + R_{\mathrm{a}}^{8/3} m_{\mathrm{a}}^{-1/6} H_{\mathrm{a}}^7 \log m_{\mathrm{a}}),$$

which finishes the proof of Proposition A.2.3.

A.5.3. Proof of Proposition A.2.4

The proof is similar to that of Proposition A.2.3 in §A.5.2. For the completeness of the paper, we present it here. We define the local linearization of Q_ω as follows,

$$\bar{Q}_\omega = Q_{\omega_0} + (\omega - \omega_0)^\top \nabla_{\omega_0} Q_\omega. \tag{A.5.28}$$

We denote by

$$\begin{aligned}
g_n &= \big(Q_{\omega(n)}(s_0, a_0) - \gamma \cdot Q_\omega(s_1, a_1) - (1-\gamma) \cdot r_0\big) \cdot \nabla_\omega Q_{\omega(n)}(s_0, a_0), \quad & g_n^e &= \mathbb{E}_{\pi_\theta}[g_n],\\
\bar{g}_n &= \big(\bar{Q}_{\omega(n)}(s_0, a_0) - \gamma \cdot Q_\omega(s_1, a_1) - (1-\gamma) \cdot r_0\big) \cdot \nabla_\omega Q_{\omega_0}(s_0, a_0), \quad & \bar{g}_n^e &= \mathbb{E}_{\pi_\theta}[\bar{g}_n],\\
g_* &= \big(Q_{\omega_*}(s_0, a_0) - \gamma \cdot Q_\omega(s_1, a_1) - (1-\gamma) \cdot r_0\big) \cdot \nabla_\omega Q_{\omega_*}(s_0, a_0), \quad & g_*^e &= \mathbb{E}_{\pi_\theta}[g_*],\\
\bar{g}_* &= \big(\bar{Q}_{\omega_*}(s_0, a_0) - \gamma \cdot Q_\omega(s_1, a_1) - (1-\gamma) \cdot r_0\big) \cdot \nabla_\omega Q_{\omega_0}(s_0, a_0), \quad & \bar{g}_*^e &= \mathbb{E}_{\pi_\theta}[\bar{g}_*],
\end{aligned} \tag{A.5.29}$$

where ω_* satisfies that

$$\omega_* = \Gamma_{\mathcal{B}(\omega_0, R_c)}(\omega_* - \alpha \cdot \bar{g}_*^e). \tag{A.5.30}$$

Here the expectation $\mathbb{E}_{\pi_\theta}[\cdot]$ is taken following $(s_0, a_0) \sim \rho_{\pi_\theta}(\cdot)$, $s_1 \sim P(\cdot \,|\, s_0, a_0)$, $a_1 \sim \pi_\theta(\cdot \,|\, s_1)$, and $r_0 = \mathcal{R}(s_0, a_0)$. By Algorithm 6, we know that

$$\omega(n+1) = \Gamma_{\mathcal{B}(\omega_0, R_c)}(\omega(n) - \eta \cdot g_n).$$

Note that

$$
\begin{aligned}
&\mathbb{E}_{\pi_\theta}\big[\|\omega(n+1)-\omega_*\|_2^2 \,|\, \omega(n)\big] \\
&\quad = \mathbb{E}_{\pi_\theta}\big[\|\Gamma_{\mathcal{B}(\omega_0,R_{\rm c})}(\omega(n)-\eta\cdot g_n)-\Gamma_{\mathcal{B}(\omega_0,R_{\rm c})}(\omega_*-\eta\cdot \bar{g}_*^e)\|_2^2 \,|\, \omega(n)\big] \\
&\quad \le \mathbb{E}_{\pi_\theta}\big[\|(\omega(n)-\eta\cdot g_n)-(\omega_*-\eta\cdot \bar{g}_*^e)\|_2^2 \,|\, \omega(n)\big] \\
&\quad = \|\omega(n)-\omega_*\|_2^2 + 2\eta\cdot \underbrace{\langle \omega_*-\omega(n), g_n^e-\bar{g}_*^e\rangle}_{\text{(iii)}} + \eta^2\cdot \underbrace{\mathbb{E}_{\pi_\theta}\big[\|g_n-\bar{g}_*^e\|_2^2 \,|\, \omega(n)\big]}_{\text{(iv)}}. \qquad \text{(A.5.31)}
\end{aligned}
$$

We upper bound term (iii) and term (iv) on the RHS of (A.5.31) in the sequel.

Upper Bound of Term (iii). By Hölder's inequality, it holds that

$$
\begin{aligned}
&\langle \omega_*-\omega(n), g_n^e-\bar{g}_*^e\rangle \\
&\quad = \langle \omega_*-\omega(n), g_n^e-\bar{g}_n^e\rangle + \langle \omega_*-\omega(n), \bar{g}_n^e-\bar{g}_*^e\rangle \\
&\quad \le \|\omega_*-\omega(n)\|_2\cdot\|g_n^e-\bar{g}_n^e\|_2 + \langle \omega_*-\omega(n), \bar{g}_n^e-\bar{g}_*^e\rangle \\
&\quad \le 2R_{\rm c}\cdot\|g_n^e-\bar{g}_n^e\|_2 + \langle \omega_*-\omega(n), \bar{g}_n^e-\bar{g}_*^e\rangle, \qquad \text{(A.5.32)}
\end{aligned}
$$

where we use the fact that $\omega(n), \omega_* \in \mathcal{B}(\omega_0, R_{\rm c})$ in the last line. Further, by the definitions in (A.5.29), it holds that

$$
\begin{aligned}
&\langle \omega_*-\omega(n), \bar{g}_n^e-\bar{g}_*^e\rangle \\
&\quad = \mathbb{E}_{\pi_\theta}\big[(\bar{Q}_{\omega(n)}(s_0,a_0)-\bar{Q}_{\omega_*}(s_0,a_0))\cdot\langle \omega_*-\omega(n), \nabla_\omega Q_{\omega_0}(s_0,a_0)\rangle\big] \\
&\quad = \mathbb{E}_{\pi_\theta}\big[(\bar{Q}_{\omega(n)}(s_0,a_0)-\bar{Q}_{\omega_*}(s_0,a_0))\cdot(\bar{Q}_{\omega_*}(s_0,a_0)-\bar{Q}_{\omega(n)}(s_0,a_0))\big] \\
&\quad = -\mathbb{E}_{\pi_\theta}\big[(\bar{Q}_{\omega(n)}(s_0,a_0)-\bar{Q}_{\omega_*}(s_0,a_0))^2\big] = -\mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega(n)}-\bar{Q}_{\omega_*})^2\big], \qquad \text{(A.5.33)}
\end{aligned}
$$

where the second equality comes from (A.5.28), and the last equality comes from the fact that the expectation is only taken to the state-action pair (s_0, a_0). Combining (A.5.32) and (A.5.33), we obtain the following upper bound of term (i),

$$\langle \omega_* - \omega(n), g_n^e - \bar{g}_*^e \rangle \leq 2R_{\mathrm{c}} \cdot \|g_n^e - \bar{g}_n^e\|_2 - \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega(n)} - \bar{Q}_{\omega_*})^2\big]. \tag{A.5.34}$$

Upper Bound of Term (iv). We now upper bound term (iv) on the RHS of (A.5.31). It holds by Cauchy-Schwarz inequality that

$$\begin{aligned}
&\mathbb{E}_{\pi_\theta}\big[\|g_n - \bar{g}_*^e\|_2^2 \,|\, \omega(n)\big] \\
&\quad \leq 2\mathbb{E}_{\pi_\theta}\big[\|g_n - g_n^e\|_2^2 \,|\, \omega(n)\big] + 2\|g_n^e - \bar{g}_*^e\|_2^2 \\
&\quad \leq 2\underbrace{\mathbb{E}_{\pi_\theta}\big[\|g_n - g_n^e\|_2^2 \,|\, \omega(n)\big]}_{\text{(iv).a}} + 4\underbrace{\|g_n^e - \bar{g}_n^e\|_2^2}_{\text{(iv).b}} + 4\underbrace{\|\bar{g}_n^e - \bar{g}_*^e\|_2^2}_{\text{(iv).c}}.
\end{aligned} \tag{A.5.35}$$

We upper bound term (iv).a, term (iv).b, and term (iv).c in the sequel.

Upper Bound of Term (iv).a. We now upper bound term (iv).a on the RHS of (A.5.35). By expanding the square, we have

$$\mathbb{E}_{\pi_\theta}\big[\|g_n - g_n^e\|_2^2 \,|\, \omega(n)\big] = \mathbb{E}_{\pi_\theta}\big[\|g_n\|_2^2 - \|g_n^e\|_2^2 \,|\, \omega(n)\big] \leq \mathbb{E}_{\pi_\theta}\big[\|g_n\|_2^2 \,|\, \omega(n)\big]. \tag{A.5.36}$$

Meanwhile, by the definition of g_n in (A.5.29), it holds that

$$\|g_n\|_2^2 = \big(Q_{\omega(n)}(s_0, a_0) - \gamma \cdot Q_\omega(s_1, a_1) - (1 - \gamma) \cdot r_0\big)^2 \cdot \|\nabla_\omega Q_{\omega(n)}(s_0, a_0)\|_2^2. \tag{A.5.37}$$

We first upper bound Q_ω as follows,

$$Q_\omega^2 = x^{(H_\mathrm{c})\top} b b^\top x^{(H_\mathrm{c})} = x^{(H_\mathrm{c})\top} x^{(H_\mathrm{c})} = \|x^{(H_\mathrm{c})}\|_2^2,$$

where $x^{(H_\mathrm{c})}$ is the output of the H_c-th layer of the DNN Q_ω. Further combining Lemma A.4.4, it holds that

$$|Q_\omega| \leq 2. \tag{A.5.38}$$

Similarly, we have

$$|Q_{\omega(n)}| \leq 2. \tag{A.5.39}$$

Combining Lemma A.4.2, (A.5.36), (A.5.37), (A.5.38), and (A.5.39), we have

$$\mathbb{E}_{\pi_\theta}\big[\|g_n - g_n^e\|_2^2 \,|\, \omega(n)\big] = O(H_\mathrm{c}^2). \tag{A.5.40}$$

Upper Bound of Term (iv).b. We now upper bound term (iv).b on the RHS of (A.5.35). It holds that

$$
\begin{aligned}
&\|g_n^e - \bar{g}_n^e\|_2 \\
&\quad = \Big\|\mathbb{E}_{\pi_\theta}\big[\big(Q_{\omega(n)}(s_0,a_0) - \gamma \cdot Q_\omega(s_1,a_1) - (1-\gamma)\cdot r_0\big)\cdot \nabla_\omega Q_{\omega(n)}(s_0,a_0) \\
&\qquad\qquad - \big(\bar{Q}_{\omega(n)}(s_0,a_0) - \gamma \cdot Q_\omega(s_1,a_1) - (1-\gamma)\cdot r_0\big)\cdot \nabla_\omega Q_{\omega_0}(s_0,a_0)\big]\Big\|_2 \\
&\quad \le \mathbb{E}_{\pi_\theta}\big[\big\|\big(\gamma\cdot Q_\omega(s_1,a_1) + (1-\gamma)\cdot r_t\big)\cdot(\nabla_\omega Q_{\omega_0}(s_0,a_0) - \nabla_\omega Q_{\omega(n)}(s_0,a_0))\big\|_2\big] \\
&\qquad\quad + \mathbb{E}_{\rho_{\pi_\theta}}\big[\|Q_{\omega(n)}\nabla_\omega Q_{\omega(n)} - \bar{Q}_{\omega(n)}\nabla_\omega Q_{\omega_0}\|_2\big]
\end{aligned}
$$

(A.5.41)

$$
\begin{aligned}
&\quad \le \mathbb{E}_{\pi_\theta}\big[\big\|\big(\gamma\cdot Q_\omega(s_1,a_1) + (1-\gamma)\cdot r_0\big)\cdot(\nabla_\omega Q_{\omega_0}(s_0,a_0) - \nabla_\omega Q_{\omega(n)}(s_0,a_0))\big\|_2\big] \\
&\qquad\quad + \mathbb{E}_{\rho_{\pi_\theta}}\big[\|(Q_{\omega(n)} - \bar{Q}_{\omega(n)})\cdot\nabla_\omega Q_{\omega_0}\|_2\big] + \mathbb{E}_{\rho_{\pi_\theta}}\big[\|Q_{\omega(n)}\cdot(\nabla_\omega Q_{\omega(n)} - \nabla_\omega Q_{\omega_0})\|_2\big].
\end{aligned}
$$

We now upper bound the three terms on the RHS of (A.5.41) in the sequel, respectively.

For the term $\mathbb{E}_{\rho_{\pi_\theta}}[\|(Q_{\omega(n)} - \bar{Q}_{\omega(n)})\cdot\nabla_\omega Q_{\omega_0}\|_2]$ on the RHS of (A.5.41), following from Lemmas A.4.2 and A.4.3, it holds with probability at least $1-\exp(-\Omega(R_{\rm c}^{2/3}m_{\rm c}^{2/3}H_{\rm c}))$ that

$$\text{(A.5.42)}\qquad \mathbb{E}_{\rho_{\pi_\theta}}\big[\|(Q_{\omega(n)} - \bar{Q}_{\omega(n)})\cdot\nabla_\omega Q_{\omega_0}\|_2\big] = O\big(R_{\rm c}^{4/3}m_{\rm c}^{-1/6}H_{\rm c}^{7/2}(\log m_{\rm c})^{1/2}\big).$$

For the term $\mathbb{E}_{\rho_{\pi_\theta}}[\|Q_{\omega(n)}\cdot(\nabla_\omega Q_{\omega(n)} - \nabla_\omega Q_{\omega_0})\|_2]$ on the RHS of (A.5.41), following from (A.5.39) and Lemma A.4.2, with probability at least $1-\exp(-\Omega(R_{\rm c}^{2/3}m_{\rm c}^{2/3}H_{\rm c}))$, we have

$$\text{(A.5.43)}\qquad \mathbb{E}_{\rho_{\pi_\theta}}\big[\|Q_{\omega(n)}\cdot(\nabla_\omega Q_{\omega(n)} - \nabla_\omega Q_{\omega_0})\|_2\big] = O\big(R_{\rm c}^{1/3}m_{\rm c}^{-1/6}H_{\rm c}^{5/2}(\log m_{\rm c})^{1/2}\big).$$

For the term $\mathbb{E}_{\pi_\theta}[\|(\gamma\cdot Q_\omega(s_1,a_1)+(1-\gamma)\cdot r_0)\cdot(\nabla_\omega Q_{\omega_0}(s_0,a_0)-\nabla_\omega Q_{\omega(n)}(s_0,a_0))\|_2]$ on the RHS of (A.5.41), we first upper bound $|\gamma\cdot Q_\omega(s_1,a_1)+(1-\gamma)\cdot r_0|$ as follows,

$$|\gamma\cdot Q_\omega(s_1,a_1)+(1-\gamma)\cdot r_0|\le 2+\mathcal{R}_{\max},$$

where we use (A.5.38) and the fact that $|\mathcal{R}(s,a)|\le\mathcal{R}_{\max}$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. Further combining Lemma A.4.2, with probability at least $1-\exp(-\Omega(R_\mathrm{c}^{2/3}m_\mathrm{c}^{2/3}H_\mathrm{c}))$, we have

$$\begin{aligned}&\mathbb{E}_{\pi_\theta}\big[\big\|\big(\gamma\cdot Q_\omega(s_1,a_1)+(1-\gamma)\cdot r_0\big)\cdot(\nabla_\omega Q_{\omega_0}(s_0,a_0)-\nabla_\omega Q_{\omega(n)}(s_0,a_0))\big\|_2\big]\\ &\quad=O\big(R_\mathrm{c}^{1/3}m_\mathrm{c}^{-1/6}H_\mathrm{c}^{5/2}(\log m_\mathrm{c})^{1/2}\big).\end{aligned}\tag{A.5.44}$$

Now, combining (A.5.41), (A.5.42), (A.5.43), and (A.5.44), it holds with probability at least $1-\exp(-\Omega(R_\mathrm{c}^{2/3}m_\mathrm{c}^{2/3}H_\mathrm{c}))$ that

$$\|g_n^e-\bar g_n^e\|_2^2=O(R_\mathrm{c}^{8/3}m_\mathrm{c}^{-1/3}H_\mathrm{c}^7\log m_\mathrm{c}).\tag{A.5.45}$$

Upper Bound of Term (iv).c. We now upper bound term (iv).c on the RHS of (A.5.35). It holds that

$$\|\bar g_n^e-\bar g_*^e\|_2^2=\big\|\mathbb{E}_{\rho_{\pi_\theta}}[(\bar Q_{\omega(n)}-\bar Q_{\omega_*})\nabla_\omega Q_{\omega_0}]\big\|_2^2\le\mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar Q_{\omega(n)}-\bar Q_{\omega_*})^2\cdot\|\nabla_\omega Q_{\omega_0}\|_2^2\big].$$

Further combining Lemma A.4.2, it holds that

$$\mathbb{E}_{\pi_\theta}\big[\|\bar g_n^e-\bar g_*^e\|_2^2\,|\,\omega(n)\big]\le O(H_\mathrm{c}^2)\cdot\mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar Q_{\omega(n)}-\bar Q_{\omega_*})^2\big].\tag{A.5.46}$$

Combining (A.5.35), (A.5.40), (A.5.45), and (A.5.46), we obtain the following upper bound for term (iv) on the RHS of (A.5.31),

(A.5.47)
$$\mathbb{E}_{\pi_\theta}\big[\|g_n - \bar{g}^e_*\|_2^2 \,|\, \omega(n)\big] \leq O(R_{\rm c}^{8/3} m_{\rm c}^{-1/3} H_{\rm c}^7 \log m_{\rm c}) + O(H_{\rm c}^2) \cdot \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega(n)} - \bar{Q}_{\omega_*})^2\big].$$

We continue upper bounding (A.5.31). By plugging (A.5.34) and (A.5.47) into (A.5.31), it holds with probability at least $1 - \exp(-\Omega(R_{\rm c}^{2/3} m_{\rm c}^{2/3} H_{\rm c}))$ that

$$\mathbb{E}_{\pi_\theta}\big[\|\omega(n+1) - \omega_*\|_2^2 \,|\, \omega(n)\big]$$
$$\leq \|\omega(n) - \omega_*\|_2^2 + 2\eta \cdot \Big(O\big(R_{\rm c}^{7/3} m_{\rm c}^{-1/6} H_{\rm c}^{7/2} (\log m_{\rm c})^{1/2}\big) - \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega(n)} - \bar{Q}_{\omega_*})^2\big]\Big)$$

(A.5.48)
$$+ \eta^2 \cdot \Big(O\big(R_{\rm c}^{8/3} m_{\rm c}^{-1/3} H_{\rm c}^7 \log m_{\rm c}\big) + O(H_{\rm c}^2) \cdot \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega(n)} - \bar{Q}_{\omega_*})^2\big]\Big).$$

Rearranging terms in (A.5.48), it holds with probability at least $1-\exp(-\Omega(R_{\rm c}^{2/3} m_{\rm c}^{2/3} H_{\rm c}))$ that

$$(2\eta - \eta^2 \cdot O(H_{\rm c}^2)) \cdot \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega(n)} - \bar{Q}_{\omega_*})^2\big]$$

(A.5.49)
$$\leq \|\omega(n) - \omega_*\|_2^2 - \mathbb{E}_{\rho_{\pi_\theta}}[\|\omega(n+1) - \omega_*\|_2^2 \,|\, \omega(n)] + \eta \cdot O(R_{\rm c}^{8/3} m_{\rm c}^{-1/3} H_{\rm c}^7 \log m_{\rm c}).$$

By telescoping the sum and using Jensen's inequality in (A.5.49), we have

$$\begin{aligned}\mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\bar{\omega}} - \bar{Q}_{\omega_*})^2\big] &\leq \frac{1}{N_\mathrm{c}} \cdot \sum_{n=0}^{N_\mathrm{c}-1} \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega(n)} - \bar{Q}_{\omega_*})^2\big] \\ &\leq 1/N_\mathrm{c} \cdot \big(2\eta - \eta^2 \cdot O(H_\mathrm{c}^2)\big)^{-1} \cdot \big(\|\omega_0 - \omega_*\|_2^2 + \eta N_\mathrm{c} \cdot O(R_\mathrm{c}^{8/3} m_\mathrm{c}^{-1/6} H_\mathrm{c}^7 \log m_\mathrm{c})\big) \\ &\leq N_\mathrm{c}^{-1/2} \cdot \|\theta_0 - \theta_*\|_2^2 + O(R_\mathrm{c}^{8/3} m_\mathrm{c}^{-1/6} H_\mathrm{c}^7 \log m_\mathrm{c}),\end{aligned}$$

where the last line comes from the choices that $\eta = N_\mathrm{c}^{-1/2}$ and $H_\mathrm{c} = O(N_\mathrm{c}^{1/4})$. Further combining Lemma A.4.3 and using triangle inequality, we have

$$\mathbb{E}_{\rho_{\pi_\theta}}\big[(Q_{\bar{\omega}} - \bar{Q}_{\omega_*})^2\big] = O(R_\mathrm{c}^2 N_\mathrm{c}^{-1/2} + R_\mathrm{c}^{8/3} m_\mathrm{c}^{-1/6} H_\mathrm{c}^7 \log m_\mathrm{c}). \tag{A.5.50}$$

To establish the upper bound of $\mathbb{E}_{\rho_{\pi_\theta}}[(\bar{Q}_{\omega_*} - \widetilde{Q})^2]$, we upper bound $\mathbb{E}_{\rho_{\pi_\theta}}[(\bar{Q}_{\omega_*} - \widetilde{Q})^2]$ in the sequel. By the definition of ω_* in (A.5.30), following a similar argument to derive (A.5.26), we have

$$\omega_* = \underset{\omega^\dagger \in \mathcal{B}(\omega_0, R_\mathrm{c})}{\operatorname{argmin}} \ \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega^\dagger}(s_0, a_0) - \widetilde{Q}(s_0, a_0))^2\big]. \tag{A.5.51}$$

From the fact that $\widetilde{Q} \in \mathcal{U}(m_\mathrm{c}, H_\mathrm{c}, R_\mathrm{c})$ by Assumption A.2.2, we know that $\widetilde{Q} = Q_{\widetilde{\omega}}$ for some $\widetilde{\omega} \in \mathcal{B}(\omega_0, R_\mathrm{c})$. Therefore, by (A.5.51), with probability at least $1 - \exp(-\Omega(R_\mathrm{c}^{2/3} m_\mathrm{c}^{2/3} H_\mathrm{c}))$, we have

$$\mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega_*} - \widetilde{Q})^2\big] \leq \mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\widetilde{\omega}} - \widetilde{Q})^2\big] = O(R_\mathrm{c}^{8/3} m_\mathrm{c}^{-1/3} H_\mathrm{c}^5 \log m_\mathrm{c}), \tag{A.5.52}$$

where we use Lemma A.4.3 in the last inequality. Now, combining (A.5.50) and (A.5.52), by triangle inequality, with probability at least $1-\exp(-\Omega(R_{\mathrm{c}}^{2/3} m_{\mathrm{c}}^{2/3} H_{\mathrm{c}}))$, we have

$$\begin{aligned}\mathbb{E}_{\rho_{\pi_\theta}}\big[(Q_{\bar\omega}-\widetilde{Q})^2\big] &\le 2\mathbb{E}_{\rho_{\pi_\theta}}\big[(Q_{\bar\omega}-\bar{Q}_{\omega_*})^2\big] + 2\mathbb{E}_{\rho_{\pi_\theta}}\big[(\bar{Q}_{\omega_*}-\widetilde{Q})^2\big] \\ &= O(R_{\mathrm{c}}^2 N_{\mathrm{c}}^{-1/2} + R_{\mathrm{c}}^{8/3} m_{\mathrm{c}}^{-1/6} H_{\mathrm{c}}^7 \log m_{\mathrm{c}}),\end{aligned}$$

which concludes the proof of Proposition A.2.4.

A.6. Proofs of Lemmas

A.6.1. Proof of Lemma 2.4.1

W denote by $\widetilde{Q} = \mathcal{T}^{\pi_{\theta_k}} Q_{\omega_k}$. In the sequel, we upper bound $\mathbb{E}_{\rho_{k+1}}[(Q_{\omega_{k+1}} - Q_{\bar\omega_{k+1}})^2]$, where $\bar\omega_{k+1} = \Gamma_R(\widetilde\omega_{k+1})$ and $\widetilde\omega_{k+1}$ is defined in (2.2.4). Note that by the fact that $\|\varphi(s,a)\|_2 \le 1$ uniformly, it suffices to upper bound $\|\omega_{k+1} - \widetilde\omega_{k+1}\|_2$. By the definitions of ω_{k+1} and $\widetilde\omega_{k+1}$ in (2.2.5) and (2.2.4), respectively, we have

$$\|\omega_{k+1} - \bar\omega_{k+1}\|_2 \le \|\widehat\Phi\widehat{v} - \Phi v\|_2 \le \|\Phi\|_2 \cdot \|\widehat{v} - v\|_2 + \|\widehat\Phi - \Phi\|_2 \cdot \|\widehat{v}\|_2. \tag{A.6.1}$$

Here, we use the fact that the projection $\Gamma_R(\cdot)$ is a contraction in the first inequality, and triangle inequality in the second inequality. Also, for notational convenience, we denote by $\widehat{\Phi}$, Φ, $\widehat{v}$, and v in (A.6.1) as follows,

$$\widehat{\Phi} = \Bigl(\frac{1}{N}\sum_{\ell=1}^{N} \varphi(s_{\ell,1}, a_{\ell,1})\varphi(s_{\ell,1}, a_{\ell,1})^\top\Bigr)^{-1}, \quad \Phi = \bigl(\mathbb{E}_{\rho_{k+1}}[\varphi(s,a)\varphi(s,a)^\top]\bigr)^{-1},$$

$$\widehat{v} = \frac{1}{N}\sum_{\ell=1}^{N}\bigl((1-\gamma)r_{\ell,2} + \gamma Q_{\omega_k}(s'_{\ell,2}, a'_{\ell,2})\bigr)\cdot \varphi(s_{\ell,2}, a_{\ell,2}),$$

$$v = \mathbb{E}_{\rho_{k+1}}\bigl[\bigl((1-\gamma)\mathcal{R} + \gamma \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}\bigr)(s,a)\cdot\varphi(s,a)\bigr].$$

By the fact that $\|\varphi(s,a)\|_2 \le 1$, $|\mathcal{R}(s,a)| \le \mathcal{R}_{\max}$, and $\|\omega_k\|_2 \le R$ we have

$$\|\Phi\|_2 \le 1/\sigma^*, \qquad \|\widehat{v}\|_2 \le \mathcal{R}_{\max} + R. \tag{A.6.2}$$

Now, following from matrix Bernstein inequality (Tropp, 2015) and Assumption 2.3.3, with probability at least $1 - p/2$, we have

$$\|\widehat{\Phi} - \Phi\|_2 \le \frac{4}{\sqrt{N}(\sigma^*)^2}\cdot\log(N/p + d/p), \tag{A.6.3}$$

where σ^* is defined in Assumption 2.3.3. Similarly, with probability at least $1 - p/2$, we have

$$\|\widehat{v} - v\|_2 \le 4(\mathcal{R}_{\max} + R)/\sqrt{N}\cdot\log(N/p + d/p). \tag{A.6.4}$$

Now, combining (A.6.1), (A.6.2), (A.6.3), and (A.6.4), we have

$$\|\omega_{k+1} - \bar{\omega}_{k+1}\|_2 \le \frac{16(\mathcal{R}_{\max} + R)}{\sqrt{N}(\sigma^*)^2}\cdot\log(N/p + d/p).$$

Therefore, it holds with probability at least $1-p$ that

$$(Q_{\omega_{k+1}} - Q_{\bar{\omega}_{k+1}})^2 \leq \frac{32(\mathcal{R}_{\max}+R)^2}{N(\sigma^*)^2} \cdot \log^2(N/p + d/p). \tag{A.6.5}$$

Meanwhile, by Assumption 2.3.2 and the definition of $\bar{\omega}_{k+1}$, we have

$$\widetilde{Q}(s,a) = Q_{\bar{\omega}_{k+1}}(s,a) \tag{A.6.6}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Combining (A.6.5) and (A.6.6) and a union bound argument, with probability at least $1-\delta$, it holds for any $k \in \{0,1,\ldots,K\}$ that

$$\mathbb{E}_{\rho_{k+1}}\big[(Q_{\omega_{k+1}}(s,a) - \widetilde{Q}(s,a))^2\big] \leq \frac{32(\mathcal{R}_{\max}+R)^2}{N(\sigma^*)^4} \cdot \log^2(NK/p + dK/p),$$

which concludes the proof of Lemma 2.4.1.

A.6.2. Proof of Lemma A.3.1

Following from the definitions of $\mathbb{P}^\pi$ and $\mathbb{P}$ in (2.1.3), we have

$$A_{1,k}(s,a) = \big[\gamma(\mathbb{P}^{\pi^*} - \mathbb{P}^{\pi_{\theta_{k+1}}})Q_{\omega_k}\big](s,a) = \big[\gamma\mathbb{P}\langle Q_{\omega_k}, \pi^* - \pi_{\theta_{k+1}}\rangle\big](s,a). \tag{A.6.7}$$

By invoking Lemma A.4.1 and combining (A.6.7), it holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$A_{1,k}(s,a) = \big[\gamma(\mathbb{P}^{\pi^*} - \mathbb{P}^{\pi_{\theta_{k+1}}})Q_{\omega_k}\big](s,a) \leq \big[\gamma\beta \cdot \mathbb{P}(\vartheta_k + \epsilon^{\mathrm{a}}_{k+1})\big](s,a),$$

where ϑ_k and $\epsilon^{\mathrm{a}}_{k+1}$ are defined in (A.3.4) and (A.3.5) of Lemma A.3.1, respectively. We conclude the proof of Lemma A.3.1.

A.6.3. Proof of Lemma A.3.2

By the definition that Q^* is the action-value function of an optimal policy π^*, we know that $Q^*(s,a) \geq Q^\pi(s,a)$ for any policy π and state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Therefore, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$A_{2,k}(s,a) = \big[\gamma \mathbb{P}^{\pi^*}(Q^{\pi_{\theta_{k+1}}} - Q_{\omega_k})\big](s,a) \leq \big[\gamma \mathbb{P}^{\pi^*}(Q^* - Q_{\omega_k})\big](s,a). \tag{A.6.8}$$

In the sequel, we upper bound $Q^*(s,a) - Q_{\omega_k}(s,a)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. We define

$$\widetilde{Q}_{k+1} = (1-\gamma) \cdot \mathcal{R} + \gamma \cdot \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}.$$

By its definition, we know that $\widetilde{Q}_{k+1} = \mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k}$. It holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$\begin{aligned}
&Q^*(s,a) - Q_{\omega_{k+1}}(s,a) \\
&\quad = Q^*(s,a) - \widetilde{Q}_{k+1}(s,a) + \widetilde{Q}_{k+1}(s,a) - Q_{\omega_{k+1}}(s,a) \\
&\quad = \Big[\big((1-\gamma) \cdot \mathcal{R} + \gamma \cdot \mathbb{P}^{\pi^*} Q^*\big) - \big((1-\gamma) \cdot \mathcal{R} + \gamma \cdot \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}\big)\Big](s,a) + \epsilon^{\mathrm{c}}_{k+1}(s,a) \\
&\quad = \gamma \cdot [\mathbb{P}^{\pi^*} Q^* - \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}](s,a) + \epsilon^{\mathrm{c}}_{k+1}(s,a) \\
&\quad = \gamma \cdot [\mathbb{P}^{\pi^*} Q^* - \mathbb{P}^{\pi^*} Q_{\omega_k}](s,a) + \gamma \cdot [\mathbb{P}^{\pi^*} Q_{\omega_k} - \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}](s,a) + \epsilon^{\mathrm{c}}_{k+1}(s,a) \\
&\quad = \gamma \cdot \big[\mathbb{P}^{\pi^*}(Q^* - Q_{\omega_k})\big](s,a) + A_{1,k}(s,a) + \epsilon^{\mathrm{c}}_{k+1}(s,a) \\
&\quad \leq \gamma \cdot \big[\mathbb{P}^{\pi^*}(Q^* - Q_{\omega_k})\big](s,a) + \gamma\beta \cdot \big[\mathbb{P}(\vartheta_k + \epsilon^{\mathrm{a}}_{k+1})\big](s,a) + \epsilon^{\mathrm{c}}_{k+1}(s,a),
\end{aligned} \tag{A.6.9}$$

where $\epsilon^{\mathrm{c}}_{k+1}$ and $A_{1,k}$ are defined in (A.3.6) and (A.3.3), respectively. Here, we use Lemma A.3.1 to upper bound $A_{1,k}$ in the last line. We remark that (A.6.9) upper bounds $Q^* - Q_{\omega_{k+1}}$ using $Q^* - Q_{\omega_k}$. By recursively applying a similar argument as in (A.6.9), we have

$$\begin{aligned} &Q^*(s,a) - Q_{\omega_k}(s,a) \\ &\quad \le \big[(\gamma\mathbb{P}^{\pi^*})^k(Q^* - Q_{\omega_0})\big](s,a) + \gamma\beta \cdot \sum_{i=0}^{k-1}\big[(\gamma\mathbb{P}^{\pi^*})^{k-i-1}\mathbb{P}(\vartheta_i + \epsilon^{\mathrm{a}}_{i+1})\big](s,a) \\ &\qquad + \sum_{i=0}^{k-1}\big[(\gamma\mathbb{P}^{\pi^*})^{k-i-1}\epsilon^{\mathrm{c}}_{i+1}\big](s,a). \end{aligned} \tag{A.6.10}$$

Combining (A.6.8) and (A.6.10), it holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$\begin{aligned} A_{2,k}(s,a) &\le \big[\gamma\mathbb{P}^{\pi^*}(Q^* - Q_{\omega_k})\big](s,a) \\ &\le \big[(\gamma\mathbb{P}^{\pi^*})^{k+1}(Q^* - Q_{\omega_0})\big](s,a) + \gamma\beta \cdot \sum_{i=0}^{k-1}\big[(\gamma\mathbb{P}^{\pi^*})^{k-i}\mathbb{P}(\vartheta_i + \epsilon^{\mathrm{a}}_{i+1})\big](s,a) \\ &\quad + \sum_{i=0}^{k-1}\big[(\gamma\mathbb{P}^{\pi^*})^{k-i}\epsilon^{\mathrm{c}}_{i+1}\big](s,a), \end{aligned}$$

where ϑ_i, $\epsilon^{\mathrm{a}}_{i+1}$, and $\epsilon^{\mathrm{c}}_{i+1}$ are defined in (A.3.4) of Lemma A.3.1, (A.3.5) of Lemma A.3.1, and (A.3.6) of Lemma A.3.2, respectively. We conclude the proof of Lemma A.3.2.

A.6.4. Proof of Lemma A.3.3

Note that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$
\begin{aligned}
A_{3,k}(s,a) &= [\mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k} - Q^{\pi_{\theta_{k+1}}}](s,a) \\
&= \Big[\big((1-\gamma)\cdot\mathcal{R} + \gamma\mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}\big) - Q^{\pi_{\theta_{k+1}}}\Big](s,a) \\
&= \Big[\big((1-\gamma)\cdot\mathcal{R} + \gamma\mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}\big) - \sum_{t=0}^{\infty}(1-\gamma)(\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^t\mathcal{R}\Big](s,a) \\
&= \Big[\sum_{t=1}^{\infty}\big((\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^t Q_{\omega_k} - (\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{t+1} Q_{\omega_k}\big) - \sum_{t=1}^{\infty}(1-\gamma)(\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^t\mathcal{R}\Big](s,a) \\
&= \sum_{t=1}^{\infty}\Big[(\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^t\big(Q_{\omega_k} - \gamma\mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k} - (1-\gamma)\cdot\mathcal{R}\big)\Big](s,a) \\
&= \sum_{t=1}^{\infty}\Big[(\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^t\big(Q_{\omega_k} - \mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k}\big)\Big](s,a) \\
&= \sum_{t=1}^{\infty}\big[(\gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^t e_{k+1}\big](s,a) = \big[\gamma\mathbb{P}^{\pi_{\theta_{k+1}}}(I - \gamma\mathbb{P}^{\pi_{\theta_{k+1}}})^{-1} e_{k+1}\big](s,a),
\end{aligned}
$$

where the term e_{k+1} in the last line is defined in (A.3.7). We conclude the proof of Lemma A.3.3.

A.6.5. Proof of Lemma A.3.4

We invoke Lemma A.4.1 in §A.4, which gives

$$\begin{aligned}
&\beta^{-1} \cdot \langle Q_{\omega_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\rangle \\
&\qquad \le \big\langle \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)) - \beta^{-1}\cdot Q_{\omega_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\big\rangle \\
&\qquad\qquad - \mathrm{KL}(\pi_{\theta_k}(\cdot\,|\,s)\,\|\,\pi_{\theta_{k+1}}(\cdot\,|\,s))
\end{aligned}$$

$$\le \big\langle \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)) - \beta^{-1}\cdot Q_{\omega_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\big\rangle = \epsilon_{k+1}^{\mathrm{b}}(s). \tag{A.6.11}$$

Combining (A.6.11) and the definition of $\mathbb{P}^\pi$ in (2.1.3), we have

$$[\mathbb{P}^{\pi_{\theta_k}} Q_{\omega_k} - \mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k}](s,a) \le \beta[\mathbb{P}\epsilon_{k+1}^{\mathrm{b}}](s). \tag{A.6.12}$$

By the definition of e_{k+1} in (A.3.7), we have

$$\begin{aligned}
e_{k+1}(s,a) &= \big[Q_{\omega_k} - \gamma\cdot\mathbb{P}^{\pi_{\theta_{k+1}}} Q_{\omega_k} - (1-\gamma)\cdot\mathcal{R}\big](s,a) \\
&\le \big[Q_{\omega_k} - \gamma\cdot\mathbb{P}^{\pi_{\theta_k}} Q_{\omega_k} - (1-\gamma)\cdot\mathcal{R}\big](s,a) + \beta\gamma\cdot[\mathbb{P}\epsilon_{k+1}^{\mathrm{b}}](s,a) \qquad \text{(A.6.13)} \\
&= \big[\widetilde{Q}_k - \gamma\cdot\mathbb{P}^{\pi_{\theta_k}}\widetilde{Q}_k - (1-\gamma)\cdot\mathcal{R}\big](s,a) + \big[\beta\gamma\mathbb{P}\epsilon_{k+1}^{\mathrm{b}} - (I - \gamma\mathbb{P}^{\pi_{\theta_k}})\epsilon_k^{\mathrm{c}}\big](s,a),
\end{aligned}$$

where we use (A.6.12) in the first inequality, and

$$\widetilde{Q}_k = (1-\gamma)\cdot\mathcal{R} + \gamma\cdot\mathbb{P}^{\pi_{\theta_k}} Q_{\omega_{k-1}}. \tag{A.6.14}$$

For the first term on the RHS of (A.6.13), by (A.6.14), it holds that

$$\widetilde{Q}_k - \gamma \cdot \mathbb{P}^{\pi_{\theta_k}} \widetilde{Q}_k - (1-\gamma) \cdot \mathcal{R}$$

$$= (1-\gamma) \cdot \mathcal{R} + \gamma \cdot \mathbb{P}^{\pi_{\theta_k}} Q_{\omega_{k-1}} - \gamma(1-\gamma) \cdot \mathbb{P}^{\pi_{\theta_k}} \mathcal{R} - (\gamma \mathbb{P}^{\pi_{\theta_k}})^2 Q_{\omega_{k-1}} - (1-\gamma) \cdot \mathcal{R}$$

(A.6.15)

$$= \gamma \cdot \mathbb{P}^{\pi_{\theta_k}} \big(Q_{\omega_{k-1}} - \gamma \mathbb{P}^{\pi_{\theta_k}} Q_{\omega_{k-1}} - (1-\gamma)\mathcal{R} \big) = \gamma \cdot \mathbb{P}^{\pi_{\theta_k}} e_k.$$

Combining (A.6.13) and (A.6.15), we have for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ that

$$e_{k+1}(s,a) \le [\gamma \mathbb{P}^{\pi_{\theta_k}} e_k](s,a) + \big[\beta\gamma\mathbb{P}\epsilon_{k+1}^{\mathrm{b}} - (I - \gamma \mathbb{P}^{\pi_{\theta_k}})\epsilon_k^{\mathrm{c}}\big](s,a). \tag{A.6.16}$$

By telescoping (A.6.16), it holds that

$$e_{k+1}(s,a) \le \bigg[\Big(\prod_{s=1}^{k} \gamma \mathbb{P}^{\pi_{\theta_s}}\Big) e_1 + \sum_{i=1}^{k} \gamma^{k-i} \Big(\prod_{s=i+1}^{k} \mathbb{P}^{\pi_{\theta_s}}\Big)\big(\beta\gamma\mathbb{P}\epsilon_{i+1}^{\mathrm{b}} - (I - \gamma\mathbb{P}^{\pi_{\theta_i}})\epsilon_i^{\mathrm{c}}\big)\bigg](s,a).$$

This finishes the proof of the lemma.

A.6.6. Proof of Lemma A.3.5

Note that $\|\omega_0\|_2 \le R$ and $|r(s,a)| \le r_{\max}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, which implies that $|Q_{\omega_0}(s,a)| \le R$ and $|Q^*(s,a)| \le r_{\max}$ by their definitions. Thus, for M_1, we have

$$|M_1| \le \mathbb{E}_\rho\Big[(I - \gamma\mathbb{P}^{\pi^*})^{-1} \sum_{k=0}^{K} (\gamma \mathbb{P}^{\pi^*})^{k+1} |Q^* - Q_{\omega_0}|\Big]$$

$$\le 4(1-\gamma)^{-1} \sum_{k=0}^{K} \gamma^{k+1} \cdot (\mathcal{R}_{\max} + R) \le 4(1-\gamma)^{-2} \cdot (\mathcal{R}_{\max} + R). \tag{A.6.17}$$

For M_2, by the definition of e_1 in (A.3.7), $|\omega_k| \leq R$, $|\phi(s,a)| \leq 1$, and $|r(s,a)| \leq r_{\max}$, we have

$$\begin{aligned} |e_1(s,a)| &= \big|[Q_{\omega_k} - \mathcal{T}^{\pi_{\theta_{k+1}}} Q_{\omega_k}](s,a)\big| \\ &= \big|\omega_k^\top \phi(s,a) - \gamma \cdot \omega_k^\top [\mathbb{P}^{\pi_{\theta_{k+1}}} \phi](s,a) - (1-\gamma) \cdot r(s,a)\big| \\ &\leq 2R + r_{\max} \end{aligned} \tag{A.6.18}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Therefore, we have

$$|M_2| \leq (1-\gamma)^{-3} \cdot (2R + \mathcal{R}_{\max}). \tag{A.6.19}$$

Meanwhile, by the initialization $\tau_0 = \infty$ in Algorithm 1, the initial policy $\pi_{\theta_0}(\cdot \,|\, s)$ is a uniform distribution over $\mathcal{A}$. Therefore, it holds for any $s \in \mathcal{S}$ that

$$\begin{aligned} \mathrm{KL}\big(\pi^*(\cdot \,|\, s) \,\|\, \pi_{\theta_0}(\cdot \,|\, s)\big) &= \int_{\mathcal{A}} \pi^*(a \,|\, s) \log \frac{\pi^*(a \,|\, s)}{\pi_{\theta_0}(a \,|\, s)} \mathrm{d}a \\ &= \int_{\mathcal{A}} \pi^*(a \,|\, s) \log \pi^*(a \,|\, s) \mathrm{d}a - \int_{\mathcal{A}} \pi^*(a \,|\, s) \log \pi_{\theta_0}(a \,|\, s) \mathrm{d}a \\ &\leq - \int_{\mathcal{A}} \pi^*(a \,|\, s) \log \pi_{\theta_0}(a \,|\, s) \mathrm{d}a \\ &= \int_{\mathcal{A}} \pi^*(a \,|\, s) \log |\mathcal{A}| \mathrm{d}a = \log |\mathcal{A}|. \end{aligned} \tag{A.6.20}$$

Therefore, by (A.6.20), we have

$$M_3 \leq (1-\gamma)^{-2} \cdot \log |\mathcal{A}| \cdot K^{1/2}, \tag{A.6.21}$$

where we use $\beta = K^{1/2}$. We see that (A.6.17), (A.6.19), and (A.6.21) upper bound M_1, M_2, and M_3, respectively. We conclude the proof of Lemma A.3.5.

A.6.7. Proof of Lemma A.3.6

For M_4, by changing the index of summation, we have

$$\begin{aligned} |M_4| &= \Big|\mathbb{E}_\rho\Big[\sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{j=0}^{\infty}(\gamma\mathbb{P}^{\pi^*})^{k-i+j}\epsilon_{i+1}^{\mathrm{c}}\Big]\Big| \\ &= \Big|\mathbb{E}_\rho\Big[\sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{t=k-i}^{\infty}(\gamma\mathbb{P}^{\pi^*})^{t}\epsilon_{i+1}^{\mathrm{c}}\Big]\Big| \\ &\le \sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{t=k-i}^{\infty}\big|\mathbb{E}_\rho\big[(\gamma\mathbb{P}^{\pi^*})^{t}\epsilon_{i+1}^{\mathrm{c}}\big]\big|, \end{aligned} \tag{A.6.22}$$

where we expand $(I - \gamma\mathbb{P}^{\pi^*})^{-1}$ into an infinite sum in the first equality. Further, by changing the measure of the expectation from ρ to ρ^* on the RHS of (A.6.22), we have

$$\sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{t=k-i}^{\infty}\big|\mathbb{E}_\rho\big[(\gamma\mathbb{P}^{\pi^*})^{t}\epsilon_{i+1}^{\mathrm{c}}\big]\big| \le \sum_{k=0}^{K}\sum_{i=0}^{k}\sum_{t=k-i}^{\infty}\gamma^t c(t)\cdot\mathbb{E}_{\rho^*}[|\epsilon_{i+1}^{\mathrm{c}}|], \tag{A.6.23}$$

where $c(t)$ is defined in Assumption 2.3.1. Further, by changing the index of summation on the RHS of (A.6.23), combining (A.6.22), we have

$$\begin{aligned} |M_4| &\le \sum_{k=0}^{K}\sum_{t=0}^{\infty}\sum_{i=\max\{0,k-t\}}^{k}\gamma^t c(t)\cdot\varepsilon_Q \\ &\le \sum_{k=0}^{K}\sum_{t=0}^{\infty}2t\gamma^t c(t)\cdot\varepsilon_Q \\ &\le \gamma\sum_{k=0}^{K}2C_{\rho,\rho^*}\cdot\varepsilon_Q \le 3KC_{\rho,\rho^*}\cdot\varepsilon_Q, \end{aligned} \tag{A.6.24}$$

where $\varepsilon_Q = \max_i \mathbb{E}_{\rho^*}[|\epsilon^{\rm c}_{i+1}|]$, and C_{ρ,ρ^*} is defined in Assumption 2.3.1.

Now, for M_5, by a similar argument as in the derivation of (A.6.24), we have

$$\begin{aligned} M_5 &\le \sum_{i=0}^{\infty}\sum_{k=0}^{K}\sum_{j=0}^{\infty}\sum_{\ell=1}^{k} \gamma^{i+j+k-\ell+1} c(i+j+k-\ell+1)\cdot \varepsilon_Q \\ &= \sum_{i=0}^{\infty}\sum_{k=0}^{K}\sum_{j=0}^{\infty}\sum_{t=i+j+1}^{i+j+k} \gamma^t c(t)\cdot\varepsilon_Q \le \sum_{k=0}^{K}\sum_{t=1}^{\infty} t^2\gamma^t c(t)\cdot\varepsilon_Q \le KC_{\rho,\rho^*}\cdot\varepsilon_Q. \end{aligned} \tag{A.6.25}$$

We see that (A.6.24) and (A.6.25) upper bound M_4 and M_5, respectively. We conclude the proof of Lemma A.3.6.

A.6.8. Proof of Lemma A.3.7

Part 1. We first show that the first inequality holds. Note that

$$\pi_{\theta_k}(a\,|\,s) = \exp(\tau_k^{-1} f_{\theta_k}(s,a))/Z_{\theta_k}(s), \qquad \pi_{\theta_{k+1}}(a\,|\,s) = \exp(\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a))/Z_{\theta_{k+1}}(s),$$

Here $Z_{\theta_k}(s), Z_{\theta_{k+1}}(s) \in \mathbb{R}$ are normalization factors, which are defined as

$$Z_{\theta_k}(s) = \sum_{a'\in\mathcal{A}} \exp(\tau_k^{-1} f_{\theta_k}(s,a')), \qquad Z_{\theta_{k+1}}(s) = \sum_{a'\in\mathcal{A}} \exp(\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a')).$$

Thus, we have

$$\begin{aligned} &\langle \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)) - \beta^{-1} Q_{\omega_k}(s,\cdot), \pi^*(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\rangle \\ &\quad = \langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi^*(\cdot\,|\,s) - \pi_{\theta_k}(\cdot\,|\,s)\rangle, \end{aligned} \tag{A.6.26}$$

where we use the fact that

$$\begin{aligned}&\langle \log Z_{\theta_{k+1}}(s) - \log Z_{\theta_k}(s), \pi^*(\cdot \,|\, s) - \pi_{\theta_{k+1}}(\cdot \,|\, s)\rangle \\ &\qquad = (\log Z_{\theta_{k+1}}(s) - \log Z_{\theta_k}(s)) \cdot \sum_{a'\in\mathcal{A}} (\pi^*(a' \,|\, s) - \pi_{\theta_{k+1}}(a' \,|\, s)) = 0.\end{aligned}$$

Thus, it remains to upper bound the right-hand side of (A.6.26). We have

(A.6.27)

$$\begin{aligned}&\langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi^*(\cdot \,|\, s) - \pi_{\theta_{k+1}}(\cdot \,|\, s)\rangle \\ &\quad = \bigg\langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi_{\theta_k}(\cdot \,|\, s) \cdot \bigg(\frac{\pi^*(\cdot \,|\, s)}{\pi_{\theta_k}(\cdot \,|\, s)} - \frac{\pi_{\theta_{k+1}}(\cdot \,|\, s)}{\pi_{\theta_k}(\cdot \,|\, s)} \bigg) \bigg\rangle.\end{aligned}$$

Taking expectation with respect to $s \sim \nu^*$ on the both sides of (A.6.27) and using the Cauchy-Schwarz inequality, we obatin

$$\begin{aligned}&\mathbb{E}_{\nu^*}\big[\big|\langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi^*(\cdot \,|\, s) - \pi_{\theta_{k+1}}(\cdot \,|\, s)\rangle\big|\big]\big| \\ &\quad = \int_{\mathcal{S}} \bigg| \bigg\langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \\ &\qquad\qquad \pi_{\theta_k}(\cdot \,|\, s) \cdot \nu_k(s) \cdot \bigg(\frac{\pi^*(\cdot \,|\, s)}{\pi_{\theta_k}(\cdot \,|\, s)} - \frac{\pi_{\theta_{k+1}}(\cdot \,|\, s)}{\pi_{\theta_k}(\cdot \,|\, s)} \bigg) \bigg\rangle \bigg| \cdot \bigg| \frac{\nu^*(s)}{\nu_k(s)} \bigg| \mathrm{d}s \\ &\quad = \int_{\mathcal{S}\times\mathcal{A}} \big| \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a) - (\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a)) \big| \\ &\qquad\qquad \cdot \bigg| \frac{\rho^*(a \,|\, s)}{\rho_k(a \,|\, s)} - \frac{\pi_{\theta_{k+1}}(a \,|\, s) \cdot \nu^*(s)}{\rho_k(a \,|\, s)} \bigg| \mathrm{d}\rho_k(s,a) \\ &\quad \le \sqrt{\mathbb{E}_{\rho_k}\big[\big(\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a) - (\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a))\big)^2\big] \mathbb{E}_{\rho_k}\bigg[\bigg| \frac{\mathrm{d}\rho^*}{\mathrm{d}\rho_k} - \frac{\mathrm{d}(\pi_{\theta_{k+1}}\nu^*)}{\mathrm{d}\rho_k} \bigg|^2\bigg]} \\ &\quad \le \sqrt{2}\tau_{k+1}^{-1} \cdot \varepsilon_{k+1,f} \cdot (\phi_k^* + \psi_k^*),\end{aligned}$$

where in the last inequality we use the error bound in (A.3.20) and the definition of ϕ_k^* and ψ_k^* in Assumption A.2.1. This finishes the proof of the first inequality.

Part 2. The proof of the second inequality follows from a similar argument as above. We have

$$\begin{aligned}&\langle \log(\pi_{\theta_{k+1}}(\cdot\,|\,s)/\pi_{\theta_k}(\cdot\,|\,s)) - \beta^{-1}Q_{\omega_k}(s,\cdot), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\rangle \\ &\quad = \langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta^{-1}Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\rangle,\end{aligned} \tag{A.6.28}$$

where we use the fact that

$$\begin{aligned}&\langle \log Z_{\theta_{k+1}}(s) - \log Z_{\theta_k}(s), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\rangle \\ &\quad = (\log Z_{\theta_{k+1}}(s) - \log Z_{\theta_k}(s)) \cdot \sum_{a'\in\mathcal{A}} (\pi_{\theta_k}(a'\,|\,s) - \pi_{\theta_{k+1}}(a'\,|\,s)) = 0.\end{aligned}$$

Thus, it remains to upper bound the right-hand side of (A.6.28). We have

(A.6.29)

$$\begin{aligned}&\langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1}Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi_{\theta_k}(\cdot\,|\,s) - \pi_{\theta_{k+1}}(\cdot\,|\,s)\rangle \\ &\quad = \biggl\langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1}Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi_{\theta_k}(\cdot\,|\,s)\cdot\biggl(1 - \frac{\pi_{\theta_{k+1}}(\cdot\,|\,s)}{\pi_{\theta_k}(\cdot\,|\,s)}\biggr)\biggr\rangle.\end{aligned}$$

Taking expectation with respect to $s \sim \nu^*$ on the both sides of (A.6.29) and using the Cauchy-Schwarz inequality, we obatin

$$
\begin{aligned}
&\mathbb{E}_{\nu^*}\big[\big|\big\langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi_{\theta_k}(\cdot \,|\, s) - \pi_{\theta_{k+1}}(\cdot \,|\, s)\big\rangle\big|\big] \\
&\quad = \int_{\mathcal{S}} \bigg| \bigg\langle \tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - (\beta_k^{-1} Q_{\omega_k}(s,\cdot) + \tau_k^{-1} f_{\theta_k}(s,\cdot)), \pi_{\theta_k}(\cdot \,|\, s)\nu_k(s)\bigg(1 - \frac{\pi_{\theta_{k+1}}(\cdot \,|\, s)}{\pi_{\theta_k}(\cdot \,|\, s)}\bigg)\bigg\rangle\bigg| \\
&\qquad\qquad \cdot \Big|\frac{\nu^*(s)}{\nu_k(s)}\Big| \mathrm{d}s \\
&\quad = \int_{\mathcal{S}\times\mathcal{A}} \big|\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a) - (\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a))\big| \bigg|1 - \frac{\pi_{\theta_{k+1}}(a \,|\, s)\cdot \nu^*(s)}{\rho_k(s,a)}\bigg| \mathrm{d}\rho_k \\
&\quad \le \mathbb{E}_{\rho_k}\big[\big(\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a) - (\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a))\big)^2\big]^{1/2} \mathbb{E}_{\rho_k}\bigg[\bigg|1 - \frac{\mathrm{d}(\pi_{\theta_{k+1}}\nu^*)}{\mathrm{d}\rho_k}\bigg|^2\bigg]^{1/2} \\
&\quad \le \sqrt{2}\tau_{k+1}^{-1} \cdot \varepsilon_{k+1,f} \cdot (1 + \psi_k^*),
\end{aligned}
$$

where in the last inequality we use the error bound in (A.3.20) and the definition of ψ_k^* in Assumption A.2.1. This finishes the proof of the second inequality.

APPENDIX B

Supplemental Materials in Chapter 3

B.1. Notations in the Appendix

In the proof, for convenience, for any invertible matrix M, we denote by $M^{-\top} = (M^{-1})^\top = (M^\top)^{-1}$ and $\|M\|_{\mathrm{F}}$ the Frobenius norm. We also denote by $\mathrm{svec}(M)$ the symmetric vectorization of the symmetric matrix M, which is the vectorization of the upper triangular matrix of the symmetric matrix M, with off-diagonal entries scaled by $\sqrt{2}$. We denote by $\mathrm{smat}(\cdot)$ the inverse operation. For any matrices G and H, we denote by $G \otimes H$ the Kronecker product, and $G \otimes_s H$ the symmetric Kronecker product, which is defined as a mapping on a vector $\mathrm{svec}(M)$ such that $(G \otimes_s H)\mathrm{svec}(M) = 1/2 \cdot \mathrm{svec}(HMG^\top + GMH^\top)$.

For notational simplicity, we write $\mathbb{E}_\pi(\cdot)$ to emphasize that the expectation is taken following the policy π.

B.2. Auxiliary Algorithms and Analysis

B.2.1. Results in D-LQR

In this section, we provide auxiliary results in analyzing Problem 3.1.2. First, we introduce the value functions of the Markov decision process (MDP) induced by Problem 3.1.2. We define the state- and action-value functions $V_{K,b}(x)$ and $Q_{K,b}(x,u)$ as follows

$$V_{K,b}(x) = \sum_{t=0}^{\infty}\Bigl\{\mathbb{E}\bigl[c(x_t,u_t)\,|\,x_0 = x\bigr] - J(K,b)\Bigr\}, \tag{B.2.1}$$

$$Q_{K,b}(x,u) = c(x,u) - J(K,b) + \mathbb{E}\bigl[V_{K,b}(x_1)\,|\,x_0 = x, u_0 = u\bigr], \tag{B.2.2}$$

where x_t follows the state transition, and u_t follows the policy $\pi_{K,b}$ given x_t. In other words, we have $u_t = -Kx_t + b + \sigma\eta_t$, where $\eta_t \sim \mathcal{N}(0,I)$. The following proposition establishes the close forms of these value functions.

Proposition B.2.1. The state-value function $V_{K,b}(x)$ takes the form of

$$V_{K,b}(x) = x^\top P_K x - \operatorname{tr}(P_K\Phi_K) + 2f_{K,b}^\top(x-\mu_{K,b}) - \mu_{K,b}^\top P_K \mu_{K,b}, \tag{B.2.3}$$

and the action-value function $Q_{K,b}(x,u)$ takes the form of

$$\begin{aligned} Q_{K,b}(x,u) = {}& \begin{pmatrix} x \\ u \end{pmatrix}^\top \Upsilon_K \begin{pmatrix} x \\ u \end{pmatrix} + 2\begin{pmatrix} p_{K,b} \\ q_{K,b} \end{pmatrix}^\top \begin{pmatrix} x \\ u \end{pmatrix} - \operatorname{tr}(P_K\Phi_K) - \sigma^2\cdot\operatorname{tr}(R + P_K BB^\top) \\ & - b^\top R b + 2b^\top RK\mu_{K,b} - \mu_{K,b}^\top(Q + K^\top RK + P_K)\mu_{K,b} \\ & + 2f_{K,b}^\top\bigl[(\overline{A}\mu + d) - \mu_{K,b}\bigr] + (\overline{A}\mu + d)^\top P_K(\overline{A}\mu + d), \end{aligned} \tag{B.2.4}$$

where $f_{K,b} = (I - A + BK)^{-\top}[(A-BK)^\top P_K(Bb + \overline{A}\mu + d) - K^\top Rb]$, and Υ_K, $p_{K,b}$, and $q_{K,b}$ are defined in (3.2.7).

Proof. See §B.5.6 for a detailed proof. □

By Proposition B.2.1, we know that $V_{K,b}(x)$ is quadratic in x, while $Q_{K,b}(x,u)$ is quadratic in $(x^\top, u^\top)^\top$. Now, we show that (3.2.5) holds.

Proposition B.2.2. The expected total cost $J(K, b)$ defined in Problem 3.1.2 takes the form of

$$J(K, b) = J_1(K) + J_2(K, b) + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top \overline{Q} \mu,$$

where

$$\begin{aligned} J_1(K) &= \mathrm{tr}\big[(Q + K^\top R K)\Phi_K\big] = \mathrm{tr}(P_K \Psi_\epsilon), \\ J_2(K, b) &= \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}^\top \begin{pmatrix} Q + K^\top R K & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}. \end{aligned}$$

Here $\mu_{K,b}$ is defined in (3.2.2), Φ_K is defined in (3.2.3), and P_K is defined in (3.2.4).

Proof. See §B.5.3 for a detailed proof. □

The following proposition establishes the gradients of $J_1(K)$ and $J_2(K, b)$, respectively.

Proposition B.2.3. The gradient of $J_1(K)$ and the gradient of $J_2(K, b)$ with respect to b take the forms of

$$\nabla_K J_1(K) = 2(\Upsilon_K^{22} K - \Upsilon_K^{21}) \cdot \Phi_K, \qquad \nabla_b J_2(K, b) = 2\big[\Upsilon_K^{22}(-K\mu_{K,b} + b) + \Upsilon_K^{21}\mu_{K,b} + q_{K,b}\big],$$

where Υ_K and $q_{K,b}$ are defined in (3.2.7).

Proof. See §B.5.5 for a detailed proof. □

The following theorem establishes the convergence of Algorithm 3.

Theorem B.2.4 (Convergence of Algorithm 3)**.** Assume that $\rho(A-BK_0)<1$. Let $\varepsilon>0$ be a sufficiently small tolerance. We set

$$
\begin{aligned}
&\gamma \leq \bigl[\|R\|_2 + \|B\|_2^2 \cdot J(K_0,b_0)\cdot \sigma_{\min}^{-1}(\Psi_\epsilon)\bigr]^{-1},\\
&N \geq C\cdot \|\Phi_{K^*}\|_2 \cdot \gamma^{-1}\cdot \log\Bigl\{4\bigl[J(K_0,b_0)-J(K^*,b^*)\bigr]\cdot \varepsilon^{-1}\Bigr\},\\
&T_n \geq \operatorname{poly}\bigl(\|K_n\|_{\mathrm{F}}, \|b_0\|_2, \|\mu\|_2, J(K_0,b_0)\bigr)\cdot \lambda_{K_n}^{-4}\cdot \bigl[1-\rho(A-BK_n)\bigr]^{-9}\cdot \varepsilon^{-5},\\
&\widetilde{T}_n \geq \operatorname{poly}\bigl(\|K_n\|_{\mathrm{F}}, \|b_0\|_2, \|\mu\|_2, J(K_0,b_0)\bigr)\cdot \lambda_{K_n}^{-2}\cdot \bigl[1-\rho(A-BK_n)\bigr]^{-12}\cdot \varepsilon^{-12},\\
&\gamma_{n,t} = \gamma_0\cdot t^{-1/2},\\
&\gamma^b \leq \min\Bigl\{1-\rho(A-BK_N),\\
&\qquad\qquad \bigl[1-\rho(A-BK_N)\bigr]^{-2}\cdot \bigl(\|B\|_2^2\cdot\|K_N\|_2^2\cdot\|R\|_2 + \|B\|_2^2\cdot\|Q\|_2\bigr)\Bigr\},\\
&H \geq C_0\cdot \nu_{K_N}^{-1}\cdot(\gamma^b)^{-1}\cdot \log\Bigl\{4\bigl[J(K_N,b_0)-J(K_N,b^{K_N})\bigr]\cdot\varepsilon^{-1}\Bigr\},\\
&T_h^b \geq \operatorname{poly}\bigl(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N,b_0)\bigr)\cdot \lambda_{K_N}^{-4}\cdot \nu_{K_N}^{-4}\cdot \bigl[1-\rho(A-BK_N)\bigr]^{-11}\cdot \varepsilon^{-5},\\
&\widetilde{T}_h^b \geq \operatorname{poly}\bigl(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N,b_0)\bigr)\cdot \lambda_{K_N}^{-4}\cdot \nu_{K_N}^{-2}\cdot \bigl[1-\rho(A-BK_N)\bigr]^{-17}\cdot \varepsilon^{-8},\\
&\gamma_{h,t}^b = \gamma_0\cdot t^{-1/2},
\end{aligned}
$$

where C, C_0, and γ_0 are positive absolute constants, $\{K_n\}_{n\in[N]}$ and $\{b_h\}_{h\in[H]}$ are the sequences generated by Algorithm 3, λ_{K_n} is specified in Proposition B.2.6, and ν_{K_N} is specified in Proposition 3.2.3. Then it holds with probability at least $1-\varepsilon^{10}$ that

$$
\begin{aligned}
&J(K_N,b_H)-J(K^*,b^*) < \varepsilon, \qquad \|b_H-b^*\|_2 \leq M_b(\mu)\cdot\varepsilon^{1/2},\\
&\|K_N-K^*\|_{\mathrm{F}} \leq \bigl[\sigma_{\min}^{-1}(\Psi_\epsilon)\cdot\sigma_{\min}^{-1}(R)\cdot\varepsilon\bigr]^{1/2}, \qquad \|\widehat{\mu}_{K_N,b_H}-\mu_{K^*,b^*}\|_2 \leq \varepsilon,
\end{aligned}
$$

where $M_b(\mu)$ is defined in (3.3.3).

Proof. See §B.4.2 for a detailed proof. □

By Theorem B.2.4, given any mean-field state μ, Algorithm 3 converges linearly to the optimal policy π_μ^* of Problem 3.1.2.

B.2.2. Primal-Dual Policy Evaluation Algorithm

Note that the critic update steps in Algorithm 3 are built upon the estimators of the matrix Υ_K and the vector $q_{K,b}$. We now derive a policy evaluation algorithm to establish the estimators of Υ_K and $q_{K,b}$, which is based on gradient temporal difference algorithm (Sutton et al., 2009a).

We define the feature vector as

$$\psi(x,u) = \begin{pmatrix} \varphi(x,u) \\ x - \mu_{K,b} \\ u - (-K\mu_{K,b} + b) \end{pmatrix}, \tag{B.2.5}$$

where

$$\varphi(x,u) = \operatorname{svec}\left[\begin{pmatrix} x - \mu_{K,b} \\ u - (-K\mu_{K,b} + b) \end{pmatrix}\begin{pmatrix} x - \mu_{K,b} \\ u - (-K\mu_{K,b} + b) \end{pmatrix}^\top\right].$$

Recall svec(M) gives the symmetric vectorization of the symmetric matrix M. We also define

$$\alpha_{K,b} = \begin{pmatrix} \operatorname{svec}(\Upsilon_K) \\ \Upsilon_K \begin{pmatrix} \mu_{K,b} \\ -K\mu_{K,b} + b \end{pmatrix} + \begin{pmatrix} p_{K,b} \\ q_{K,b} \end{pmatrix} \end{pmatrix}, \tag{B.2.6}$$

where Υ_K, $p_{K,b}$, and $q_{K,b}$ are defined in (3.2.7). To estimate Υ_K and $q_{K,b}$, it suffices to estimate $\alpha_{K,b}$. Meanwhile, we define

$$\Theta_{K,b} = \mathbb{E}_{\pi_{K,b}}\Big\{\psi(x,u)\big[\psi(x,u) - \psi(x',u')\big]^\top\Big\}, \tag{B.2.7}$$

where (x', u') is the state-action pair after (x, u) following the policy $\pi_{K,b}$ and the state transition. The following proposition characterizes the connection between $\Theta_{K,b}$ and $\alpha_{K,b}$.

Proposition B.2.5. It holds that

$$\begin{pmatrix} 1 & 0 \\ \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] & \Theta_{K,b} \end{pmatrix} \begin{pmatrix} J(K,b) \\ \alpha_{K,b} \end{pmatrix} = \begin{pmatrix} J(K,b) \\ \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big] \end{pmatrix},$$

where $\psi(x, u)$ is defined in (B.2.5), $\alpha_{K,b}$ is defined in (B.2.6), and $\Theta_{K,b}$ is defined in (B.2.7).

Proof. See §B.5.7 for a detailed proof. □

By Proposition B.2.5, to obtain $\alpha_{K,b}$, it suffices to solve the following linear system in $\zeta = (\zeta_1, \zeta_2^\top)^\top$,

$$\widetilde{\Theta}_{K,b} \cdot \zeta = \begin{pmatrix} J(K,b) \\ \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big] \end{pmatrix}, \tag{B.2.8}$$

where for notational convenience, we define

$$\widetilde{\Theta}_{K,b} = \begin{pmatrix} 1 & 0 \\ \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] & \Theta_{K,b} \end{pmatrix}. \tag{B.2.9}$$

The following proposition shows that $\Theta_{K,b}$ is invertible.

Proposition B.2.6. If $\rho(A-BK) < 1$, then the matrix $\Theta_{K,b}$ is invertible, and $\|\Theta_{K,b}\|_2 \leq 4(1+\|K\|_{\mathrm{F}}^2)^2 \cdot \|\Phi_K\|_2^2$. Also, $\sigma_{\min}(\widetilde{\Theta}_{K,b}) \geq \lambda_K$, where λ_K only depends on $\|K\|_2$ and $\rho(A-BK)$.

Proof. See §B.5.8 for a detailed proof. □

By Proposition B.2.6, $\Theta_{K,b}$ is invertible. Therefore, (B.2.8) admits the unique solution $\zeta_{K,b} = (J(K,b), \alpha_{K,b}^\top)^\top$.

Now, we present the primal-dual gradient temporal difference algorithm.

Primal-Dual Gradient Method. Instead of solving (B.2.8) directly, we minimize the following loss function with respect to $\zeta = ((\zeta^1)^\top, (\zeta^2)^\top)$,

$$\big[\zeta^1 - J(K,b)\big]^2 + \Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2^2. \tag{B.2.10}$$

By Fenchel's duality, the minimization of (B.2.10) is equivalent to the following primal-dual min-max problem,

$$\min_{\zeta\in\mathcal{V}_\zeta}\max_{\xi\in\mathcal{V}_\xi} F(\zeta,\xi) = \Big\{\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\}^\top \xi^2 \\ + \big[\zeta^1 - J(K,b)\big]\cdot\xi^1 - \|\xi\|_2^2/2, \tag{B.2.11}$$

where we restrict the primal variable ζ in a compact set $\mathcal{V}_\zeta$ and the dual variable ξ in a compact set $\mathcal{V}_\xi$, which are specified in Definition B.2.7. It holds that

$$\nabla_{\zeta^1} F = \xi^1 + \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]^\top \xi^2, \qquad \nabla_{\zeta^2} F = \Theta_{K,b}^\top \xi^2, \qquad \nabla_{\xi^1} F = \zeta^1 - J(K,b) - \xi^1,$$

(B.2.12)

$$\nabla_{\xi^2} F = \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big] - \xi^2.$$

The primal-dual gradient method updates ζ and ξ via

$$\begin{aligned} &\zeta^1 \leftarrow \zeta^1 - \gamma \cdot \nabla_{\zeta^1} F(\zeta,\xi), \qquad &&\zeta^2 \leftarrow \zeta^2 - \gamma \cdot \nabla_{\zeta^2} F(\zeta,\xi) \\ &\xi^1 \leftarrow \xi^1 - \gamma \cdot \nabla_{\xi^1} F(\zeta,\xi), \qquad &&\xi^2 \leftarrow \xi^2 - \gamma \cdot \nabla_{\xi^2} F(\zeta,\xi). \end{aligned} \tag{B.2.13}$$

Estimation of Mean-Field State $\mu_{K,b}$. To utilize the primal-dual gradient method in (B.2.13), it remains to evaluate the feature vector $\psi(x,u)$. Note that by (B.2.5), the evaluation of the feature vector $\psi(x,u)$ requires the mean-field state $\mu_{K,b}$. In what follows, we establish the estimator $\widehat{\mu}_{K,b}$ of the mean-field state $\mu_{K,b}$ by simulating the MDP following the policy $\pi_{K,b}$ for $\widetilde{T}$ steps, and calculate the estimated feature vector $\widehat{\psi}(x,u)$ by

$$\widehat{\psi}(x,u) = \begin{pmatrix} \widehat{\varphi}(x,u) \\ x - \widehat{\mu}_{K,b} \\ u - (-K\widehat{\mu}_{K,b} + b) \end{pmatrix}, \tag{B.2.14}$$

where $\widehat{\varphi}(x,u)$ takes the form of

$$\widehat{\varphi}(x,u) = \operatorname{svec}\biggl[\begin{pmatrix} x - \widehat{\mu}_{K,b} \\ u - (-K\widehat{\mu}_{K,b} + b) \end{pmatrix} \begin{pmatrix} x - \widehat{\mu}_{K,b} \\ u - (-K\widehat{\mu}_{K,b} + b) \end{pmatrix}^\top \biggr].$$

We now define the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$ in (B.2.11).

Definition B.2.7. Given K_0 and b_0 such that $\rho(A - BK_0) < 1$ and $J(K_0, b_0) < \infty$, we define the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$ as

$$\mathcal{V}_\zeta = \Bigl\{\zeta \colon 0 \le \zeta^1 \le J(K_0, b_0), \|\zeta^2\|_2 \le M_{\zeta,1} + M_{\zeta,2} \cdot (1 + \|K\|_{\mathrm{F}}) \cdot \bigl[1 - \rho(A - BK)\bigr]^{-1}\Bigr\},$$

$$\mathcal{V}_\xi = \Bigl\{\xi \colon |\xi^1| \le J(K_0, b_0), \|\xi^2\|_2 \le M_\xi \cdot \bigl(1 + \|K\|_{\mathrm{F}}^2\bigr)^3 \cdot \bigl[1 - \rho(A - BK)\bigr]^{-1}\Bigr\}.$$

Here $M_{\zeta,1}$, $M_{\zeta,2}$, and M_ξ are constants independent of K and b, which take the forms of

$$\begin{aligned} M_{\zeta,1} = {} & \Bigl[\bigl(\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}}\bigr) + \bigl(\|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2\bigr) \cdot \sqrt{d} \cdot J(K_0, b_0) \cdot \sigma_{\min}^{-1}(\Psi_\omega)\Bigr] \\ & + \bigl(\|A\|_2 + \|B\|_2\bigr) \cdot J(K_0, b_0)^2 \cdot \sigma_{\min}^{-1}(\Psi_\omega) \cdot \sigma_{\min}^{-1}(Q), \\ & + \Bigl[\bigl(\|Q\|_2 + \|R\|_2\bigr) + \bigl(\|A\|_2 + \|B\|_2\bigr)^2 \cdot J(K_0, b_0) \cdot \sigma_{\min}^{-1}(\Psi_\omega)\Bigr] \\ & \cdot J(K_0, b_0) \cdot \bigl[\sigma_{\min}^{-1}(Q) + \sigma_{\min}^{-1}(R)\bigr] \end{aligned}$$

$$M_{\zeta,2} = \bigl(\|A\|_2 + \|B\|_2\bigr) \cdot (\kappa_Q + \kappa_R), \qquad M_\xi = C \cdot (M_{\zeta,1} + M_{\zeta,2}) \cdot J(K_0, b_0)^2 \cdot \sigma_{\min}^{-2}(Q),$$

where C is a positive absolute constant, and κ_Q and κ_R are condition numbers of Q and R, respectively.

We summarize the primal-dual gradient temporal difference algorithm in Algorithm 7. Hereafter, for notational convenience, we denote by $\widehat{\psi}_t$ the estimated feature vector $\widehat{\psi}(x_t, u_t)$.

Algorithm 7 Primal-Dual Gradient Temporal Difference Algorithm.

```
Input: Policy $\pi_{K,b}$, mean-field state $\mu$, numbers of iteration $\widetilde{T}$ and $T$, stepsizes $\{\gamma_t\}_{t\in[T]}$, parameters $K_0$ and $b_0$.
Define the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$ via Definition B.2.7 with $K_0$ and $b_0$.
Initialize the parameters by $\zeta_0 \in \mathcal{V}_\zeta$ and $\xi_0 \in \mathcal{V}_\xi$.
Sample $\widetilde{x}_0$ from the the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
for $t = 0, \ldots, \widetilde{T} - 1$ do
    Given the mean-field state $\mu$, take action $\widetilde{u}_t$ following $\pi_{K,b}$ and generate the next state $\widetilde{x}_{t+1}$.
end for
Set $\widehat{\mu}_{K,b} \leftarrow 1/\widetilde{T} \cdot \sum_{t=1}^{\widetilde{T}} \widetilde{x}_t$ and compute the estimated feature vector $\widehat{\psi}$ via (B.2.14).
Sample $x_0$ from the the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
for $t = 0, \ldots, T - 1$ do
    Given the mean-field state $\mu$, take action $u_t$ following $\pi_{K,b}$, observe the cost $c_t$, and generate the next state $x_{t+1}$.
    Set $\delta_{t+1} \leftarrow \zeta_t^1 + (\widehat{\psi}_t - \widehat{\psi}_{t+1})^\top \zeta_t^2 - c_t$.
    Update parameters via
        $\zeta_{t+1}^1 \leftarrow \zeta_t^1 - \gamma_{t+1} \cdot (\xi_t^1 + \widehat{\psi}_t^\top \xi_t^2)$,    $\zeta_{t+1}^2 \leftarrow \zeta_t^2 - \gamma_{t+1} \cdot \widehat{\psi}_t (\widehat{\psi}_t - \widehat{\psi}_{t+1})^\top \xi_t^2$,
        $\xi_{t+1}^1 \leftarrow (1 - \gamma_{t+1}) \cdot \xi_t^1 + \gamma_{t+1} \cdot (\zeta_t^1 - c_t)$,    $\xi_{t+1}^2 \leftarrow (1 - \gamma_{t+1}) \cdot \xi_t^2 + \gamma_{t+1} \cdot \delta_{t+1} \cdot \widehat{\psi}_t$.
    Project $\zeta_{t+1}$ and $\xi_{t+1}$ to $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$, respectively.
end for
Set $\widehat{\alpha}_{K,b} \leftarrow (\sum_{t=1}^T \gamma_t)^{-1} \cdot (\sum_{t=1}^T \gamma_t \cdot \zeta_t^2)$, and
        $\widehat{\Upsilon}_K \leftarrow \mathrm{smat}(\widehat{\alpha}_{K,b,1})$,    $\begin{pmatrix} \widehat{p}_{K,b} \\ \widehat{q}_{K,b} \end{pmatrix} \leftarrow \widehat{\alpha}_{K,b,2} - \widehat{\Upsilon}_K \begin{pmatrix} \widehat{\mu}_{K,b} \\ -K\widehat{\mu}_{K,b} + b \end{pmatrix}$,
    where $\widehat{\alpha}_{K,b,1} = (\widehat{\alpha}_{K,b})_1^{(k+d+1)(k+d)/2}$ and $\widehat{\alpha}_{K,b,2} = (\widehat{\alpha}_{K,b})_{(k+d+1)(k+d)/2+1}^{(k+d+3)(k+d)/2}$.
Output: Estimators $\widehat{\mu}_{K,b}$, $\widehat{\Upsilon}_K$, and $\widehat{q}_{K,b}$.
```

We now characterize the rate of convergence of Algorithm 7.

Theorem B.2.8 (Convergence of Algorithm 7)**.** Given K_0, b_0, K, and b such that $\rho(A - BK_0) < 1$ and $J(K, b) \leq J(K_0, b_0)$, we define the sets $\mathcal{V}_\zeta$ and $\mathcal{V}_\xi$ through Definition B.2.7.

Let $\gamma_t = \gamma_0 t^{-1/2}$, where γ_0 is a positive absolute constant. Let $\rho \in (\rho(A - BK), 1)$. For $\widetilde{T} \geq \mathrm{poly}_0(\|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0, b_0)) \cdot (1-\rho)^{-6}$ and a sufficiently large T, it holds with probability at least $1 - T^{-4} - \widetilde{T}^{-6}$ that

$$\begin{aligned}&\|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2^2 \\ &\qquad \leq \lambda_K^{-2} \cdot \mathrm{poly}_1\big(\|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0, b_0)\big) \cdot \biggl[\frac{\log^6 T}{T^{1/2} \cdot (1-\rho)^4} + \frac{\log \widetilde{T}}{\widetilde{T}^{1/4} \cdot (1-\rho)^2}\biggr],\end{aligned}$$

where λ_K is defined in Proposition B.2.6. Same bounds for $\|\widehat{\Upsilon}_K - \Upsilon_K\|_{\mathrm{F}}^2$, $\|\widehat{p}_{K,b} - p_{K,b}\|_2^2$, and $\|\widehat{q}_{K,b} - q_{K,b}\|_2^2$ hold. Meanwhile, it holds with probability at least $1 - \widetilde{T}^{-6}$ that

$$\|\widehat{\mu}_{K,b} - \mu_{K,b}\|_2 \leq \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}} \cdot (1-\rho)^{-2} \cdot \mathrm{poly}_2\big(\|\Phi_K\|_2, \|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0, b_0)\big).$$

Proof. See §B.4.3 for a detailed proof. □

B.2.3. Temporal Difference Policy Evaluation Algorithm

Besides the primal-dual gradient temporal difference algorithm, we can also evaluate $\alpha_{K,b}$ by TD(0) method (Sutton and Barto, 2018) in practice, which is presented in Algorithm 8.

Note that in related literature (Bhandari et al., 2018; Korda and La, 2015), non-asymptotic convergence analysis of TD(0) method with linear function approximation is only applied to discounted MDP. As for our ergodic setting, the convergence of TD(0) method is only shown asymptotically (Borkar and Meyn, 2000; Kushner and Yin, 2003)

Algorithm 8 Temporal Difference Policy Evaluation Algorithm.

```
Input: Policy $\pi_{K,b}$, number of iteration $\widetilde{T}$ and $T$, stepsizes $\{\gamma_t\}_{t\in[T]}$.
Sample $\widetilde{x}_0$ from the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
for $t = 0, \ldots, \widetilde{T}-1$ do
    Take action $\widetilde{u}_t$ under the policy $\pi_{K,b}$ and generate the next state $\widetilde{x}_{t+1}$.
end for
Set $\widehat{\mu}_{K,b} \leftarrow 1/\widetilde{T} \cdot \sum_{t=1}^{\widetilde{T}} \widetilde{x}_t$.
Sample $x_0$ from the the stationary distribution $\mathcal{N}(\mu_{K,b}, \Phi_K)$.
for $t = 0, \ldots, T$ do
    Given the mean-field state $\mu$, take action $u_t$ following $\pi_{K,b}$, observe the cost $c_t$,
    and generate the next state $x_{t+1}$.
    Set $\delta_{t+1} \leftarrow \zeta_t^1 + (\widehat{\psi}_t - \widehat{\psi}_{t+1})^\top \zeta_t^2 - c_t$.
    Update parameters via $\zeta_{t+1}^1 \leftarrow (1-\gamma_{t+1}) \cdot \zeta_t^1 + \gamma_{t+1} \cdot c_t$ and $\zeta_{t+1}^2 \leftarrow \zeta_t^2 - \gamma_{t+1} \cdot \delta_{t+1} \cdot \widehat{\psi}_t$.
    Project $\zeta_t$ to $\mathcal{V}'_\zeta$, where $\mathcal{V}'_\zeta$ is a compact set.
end for
Set $\widehat{\alpha}_{K,b} \leftarrow (\sum_{t=1}^T \gamma_t)^{-1} \cdot (\sum_{t=1}^T \gamma_t \cdot \zeta_t^2)$, and
    $$\widehat{\Upsilon}_K \leftarrow \mathrm{smat}(\widehat{\alpha}_{K,b,1}), \qquad \begin{pmatrix} \widehat{p}_{K,b} \\ \widehat{q}_{K,b} \end{pmatrix} \leftarrow \widehat{\alpha}_{K,b,2} - \widehat{\Upsilon}_K \begin{pmatrix} \widehat{\mu}_{K,b} \\ -K\widehat{\mu}_{K,b} + b \end{pmatrix},$$
    where $\widehat{\alpha}_{K,b,1} = (\widehat{\alpha}_{K,b})_1^{(k+d+1)(k+d)/2}$ and $\widehat{\alpha}_{K,b,2} = (\widehat{\alpha}_{K,b})_{(k+d+1)(k+d)/2+1}^{(k+d+3)(k+d)/2}$.
Output: Estimators $\widehat{\mu}_{K,b}$, $\widehat{\Upsilon}_K$, and $\widehat{q}_{K,b}$.
```

using ordinary differential equation method. Therefore, in the convergence theorem proposed in §3.2, we only focus on the primal-dual gradient temporal difference method (Algorithm 7) to establish non-asymptotic convergence result.

B.3. General Formulation

Compared with Problem 3.1.3, a more general formulation includes an additional term $x_t^\top P\mu$ in the cost function. For the completeness of this paper, we define this general formulation here. Following from a same argument as in §3.1, it suffices to study the setting where t is sufficiently large. First, we propose the following general drifted LQR (general D-LQR) problem, which is parallel to Problem 3.1.2.

Problem B.3.1 (General D-LQR)**.** For any given mean-field state $\mu \in \mathbb{R}^m$, consider the following formulation

$$
\begin{aligned}
&x_{t+1} = Ax_t + Bu_t + \overline{A}\mu + d + \omega_t,\\
&\widetilde{c}_\mu(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + \mu^\top \overline{Q}\mu + 2x_t^\top P\mu,\\
&\widetilde{J}_\mu(\pi) = \lim_{T\to\infty} \mathbb{E}\Bigg[\frac{1}{T}\sum_{t=0}^{T} \widetilde{c}_\mu(x_t, u_t)\Bigg],
\end{aligned}
$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. We aim to find an optimal policy π_μ^* such that $\widetilde{J}_\mu(\pi_\mu^*) = \inf_{\pi\in\Pi} \widetilde{J}_\mu(\pi)$.

In Problem B.3.1, the unique optimal policy $\pi_\mu^*(\cdot)$ still admits a linear form $\pi_\mu^*(x_t) = -K_{\pi_\mu^*}x_t + b_{\pi_\mu^*}$ (Anderson and Moore, 2007), where the matrix $K_{\pi_\mu^*} \in \mathbb{R}^{k\times m}$ and the vector $b_{\pi_\mu^*} \in \mathbb{R}^k$ are the parameters of the policy π. It then suffices to find the optimal policy in the class Π defined in (3.1.1).

Parallel to Problem 3.1.3, we define the general LQ-MFG problem as follows.

Problem B.3.2 (General LQ-MFG)**.** We consider the following formulation

$$
\begin{aligned}
&x_{t+1} = Ax_t + Bu_t + \overline{A}\mu + d + \omega_t,\\
&\widetilde{c}(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t + \mu^\top \overline{Q}\mu + 2x_t^\top P\mu,\\
&\widetilde{J}(\pi, \mu) = \lim_{T\to\infty} \mathbb{E}\Bigg[\frac{1}{T}\sum_{t=0}^{T} \widetilde{c}(x_t, u_t)\Bigg],
\end{aligned}
$$

where $x_t \in \mathbb{R}^m$ is the state vector, $u_t \in \mathbb{R}^k$ is the action vector generated by the policy π, $\mu \in \mathbb{R}^m$ is the mean-field state, $\omega_t \in \mathbb{R}^m$ is an independent random noise term following the Gaussian distribution $\mathcal{N}(0, \Psi_\omega)$, and $d \in \mathbb{R}^m$ is a drift term. We aim to find a pair (μ^*, π^*) such that (i) $\widetilde{J}(\pi^*, \mu^*) = \inf_{\pi\in\Pi} \widetilde{J}(\pi, \mu^*)$; (ii) $\mathbb{E}x_t^*$ converges to μ^* as $t \to \infty$, where $\{x_t^*\}_{t\geq 0}$ is the Markov chain of states generated by the policy π^*.

One can see that Problem B.3.2 aims to find a Nash equilibrium pair (μ^*, π^*).

Similar to the discussions in §3.2.2, to solve Problem B.3.2, one can design an algorithm similar to Algorithm 2, which solves Problem B.3.1 and obtain the new mean-field state at each iteration. We omit the detailed algorithm here. Now, we focus on solving Problem B.3.1 in the sequel.

Similar to §3.2.3, we drop the subscript μ when we focus on Problem B.3.1 for a fixed μ. We write $\pi_{K,b}(x) = -Kx + b + \sigma \cdot \eta$ to emphasize the dependence on K and b, and $\widetilde{J}(K, b) = \widetilde{J}(\pi_{K,b})$ consequently. We derive an explicit form of the expected total cost $\widetilde{J}(K, b)$ in the following proposition.

Proposition B.3.3. The expected total cost $\widetilde{J}(K, b)$ in Problem B.3.1 is decomposed as

$$\widetilde{J}(K, b) = \widetilde{J}_1(K) + \widetilde{J}_2(K, b) + \sigma^2 \cdot \operatorname{tr}(R) + \mu^\top \overline{Q} \mu,$$

where $\widetilde{J}_1(K)$ and $\widetilde{J}_2(K, b)$ take the forms of

$$\begin{aligned}
\widetilde{J}_1(K) &= \operatorname{tr}\big[(Q + K^\top RK)\Phi_K\big] = \operatorname{tr}(P_K \Psi_\epsilon), \\
\widetilde{J}_2(K, b) &= \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix}^\top \begin{pmatrix} Q + K^\top RK & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu_{K,b} \\ b \end{pmatrix} + 2\mu^\top P \mu_{K,b}.
\end{aligned}$$

Here $\mu_{K,b}$ is given in (3.2.2), Φ_K is given in (3.2.3), and P_K is given in (3.2.4).

Proof. The proof is similar to the one of Proposition B.2.2. Thus we omit it here. □

Compared with the form of $J(K, b)$ given in (3.2.5), we see that the only difference is that $\widetilde{J}(K, b)$ contains an extra term $2\mu^\top P\mu_{K,b}$ in $\widetilde{J}_2(K, b)$, which is only a linear term in b (recall that $\mu_{K,b}$ is linear in b by (3.2.2)). Thus, $\widetilde{J}_2(K, b)$ is still strongly convex in b, as shown in the proposition below.

Proposition B.3.4. Given any K, the function $\widetilde{J}_2(K, b)$ is ν_K-strongly convex in b, here $\nu_K = \sigma_{\min}(Y_{1,K}^\top Y_{1,K} + Y_{2,K}^\top Y_{2,K})$, where $Y_{1,K} = R^{1/2}K(I - A + BK)^{-1}B - R^{1/2}$ and $Y_{2,K} = Q^{1/2}(I - A + BK)^{-1}B$. Also, $\widetilde{J}_2(K, b)$ has ι_K-Lipschitz continuous gradient in b, where ι_K is upper bounded such that $\iota_K \le [1 - \rho(A - BK)]^{-2} \cdot (\|B\|_*^2 \cdot \|K\|_*^2 \cdot \|R\|_* + \|B\|_*^2 \cdot \|Q\|_*)$.

Proof. The proof is similar to the one of Proposition 3.2.3. Thus we omit it here. □

Parallel to Proposition 3.2.4, we derive a similar proposition in the sequel.

Proposition B.3.5. Denote by $\widetilde{b}^K = \operatorname{argmin}_b \widetilde{J}_2(K, b)$, then $\widetilde{J}_2(K, \widetilde{b}^K)$ takes the form

$$\widetilde{J}_2(K, \widetilde{b}^K) = \begin{pmatrix} \overline{A}\mu + d \\ P^\top \mu \end{pmatrix}^\top \begin{pmatrix} S & S(I - A)Q^{-1} \\ Q^{-1}(I - A)^\top S & 3Q^{-1}(I - A)^\top S(I - A)Q^{-1} - Q^{-1} \end{pmatrix} \begin{pmatrix} \overline{A}\mu + d \\ P^\top \mu \end{pmatrix},$$

which is independent of K. Here $S = [(I - A)Q^{-1}(I - A)^\top + BR^{-1}B^\top]^{-1}$. And $\widetilde{b}^K$ takes the form

$$\widetilde{b}^K = \big[KQ^{-1}(I - A)^\top - R^{-1}B^\top\big] \cdot S \cdot \big[(\overline{A}\mu + d) + (I - A)Q^{-1}P^\top \mu\big] - KQ^{-1}P^\top \mu.$$

Proof. The proof is similar to the one of Proposition 3.2.4. Thus we omit it here. □

Similar to Problem 3.1.2, we define the state- and action-value functions as

$$\widetilde{V}_{K,b}(x) = \sum_{t=0}^{\infty}\Big\{\mathbb{E}\big[\widetilde{c}(x_t,u_t)\,|\,x_0=x, u_t=-Kx_t+b+\sigma\eta_t\big]-\widetilde{J}(K,b)\Big\},$$

$$\widetilde{Q}_{K,b}(x,u) = \widetilde{c}(x,u)-\widetilde{J}(K,b)+\mathbb{E}\big[\widetilde{V}_{K,b}(x')\,|\,x,u\big],$$

where the x' is the state generated by the state transition after the state-action pair (x,u). A slight modification of Proposition B.2.1 gives the proposition below.

Proposition B.3.6. For Problem B.3.1, the state-value function $\widetilde{V}_{K,b}(x)$ takes the form

$$\widetilde{V}_{K,b}(x) = x^\top P_K x - \operatorname{tr}(P_K\Phi_K) + 2\widetilde{f}_{K,b}^\top(x-\mu_{K,b}) - (\mu_{K,b})^\top P_K\mu_{K,b},$$

and the action-value function $\widetilde{Q}_{K,b}(x,u)$ takes the form

$$\begin{aligned}\widetilde{Q}_{K,b}(x,u) = {} & \begin{pmatrix} x \\ u \end{pmatrix}^\top \Upsilon_K \begin{pmatrix} x \\ u \end{pmatrix} + 2\begin{pmatrix} \widetilde{p}_{K,b} \\ \widetilde{q}_{K,b} \end{pmatrix}^\top \begin{pmatrix} x \\ u \end{pmatrix} \\ & - \operatorname{tr}(P_K\Phi_K) - \sigma^2\cdot\operatorname{tr}(R+P_KBB^\top) - b^\top Rb \\ & + 2b^\top RK\mu_{K,b} - (\mu_{K,b})^\top(Q+K^\top RK+P_K)\mu_{K,b} + 2\widetilde{f}_{K,b}^\top\big[(\overline{A}\mu+d)-\mu_{K,b}\big] \\ & + (\overline{A}\mu+d)^\top P_K(\overline{A}\mu+d) - 2\mu^\top P\mu_{K,b}.\end{aligned}$$

Here the matrix Υ_K is given in (3.2.7), and the vectors $\widetilde{p}_{K,b}, \widetilde{q}_{K,b}$ are given as

$$\begin{pmatrix} \widetilde{p}_{K,b} \\ \widetilde{q}_{K,b} \end{pmatrix} = \begin{pmatrix} A^\top\big[P_K\cdot(\overline{A}\mu+d)+\widetilde{f}_{K,b}\big]+P\mu \\ B^\top\big[P_K\cdot(\overline{A}\mu+d)+\widetilde{f}_{K,b}\big] \end{pmatrix}, \tag{B.3.1}$$

where the vector $\widetilde{f}_{K,b} = (I - A + BK)^{-\top}[(A - BK)^\top P_K(Bb + \overline{A}\mu + d) - K^\top Rb + P\mu]$.

Proof. The proof is similar to the one of Proposition B.2.1. Thus we omit it here. □

Now we establish the gradients of $\widetilde{J}(K,b)$ for Problem B.3.1.

Proposition B.3.7. The gradient of $\widetilde{J}_1(K)$ and the gradient of $\widetilde{J}_2(K,b)$ w.r.t. b takes the form

$$\nabla_K \widetilde{J}_1(K) = 2(\Upsilon_K^{22}K - \Upsilon_K^{21}) \cdot \Phi_K,$$

$$\nabla_b \widetilde{J}_2(K,b) = 2\big[\Upsilon_K^{22}(-K\mu_{K,b} + b) + \Upsilon_K^{21}\mu_{K,b} + \widetilde{q}_{K,b}\big],$$

where the matrix Υ_K is given in (3.2.7), and the vector $\widetilde{q}_{K,b}$ is given in (B.3.1).

Proof. The proof is similar to the one of Proposition B.2.3. Thus we omit it here. □

Equipped with above results, parallel to the analysis in §3.2, it is clear that by slight modification of Algorithms 2, 3, and 7, we derive similar actor-critic algorithms to solve both Problem B.3.2 and Problem B.3.1, where all the non-asymptotic convergence results hold. We omit the algorithms and the convergence results here.

B.4. Proofs of Theorems

B.4.1. Proof of Theorem 3.3.1

We define $\mu_{s+1}^* = \Lambda(\mu_s)$, which is the mean-field state generated by the optimal policy $\pi_{K^*(\mu_s),b^*(\mu_s)} = \Lambda_1(\mu_s)$ under the current mean-field state μ_s. By Proposition 3.2.4, the optimal $K^*(\mu)$ is independent of the mean-field state μ. Therefore, we write $K^* = K^*(\mu)$

hereafter for notational convenience. By (3.2.2), we know that

$$\mu^*_{s+1} = (I - A + BK^*)^{-1} \cdot \bigl[Bb^*(\mu_s) + \overline{A}\mu_s + d\bigr].$$

We define

$$\widetilde{\mu}_{s+1} = (I - A + BK_s)^{-1}(Bb_s + \overline{A}\mu_s + d),$$

which is the mean-field state generated by the policy π_s under the current mean-field state μ_s, where K_s and b_s are the parameters of the policy π_s. By triangle inequality, we have

$$\|\mu_{s+1} - \mu^*\|_2 \le \underbrace{\|\mu_{s+1} - \widetilde{\mu}_{s+1}\|_2}_{E_1} + \underbrace{\|\widetilde{\mu}_{s+1} - \mu^*_{s+1}\|_2}_{E_2} + \underbrace{\|\mu^*_{s+1} - \mu^*\|_2}_{E_3}, \tag{B.4.1}$$

where μ_{s+1} is generated by Algorithm 2. We upper bound E_1, E_2, and E_3 in the sequel.

Upper Bound of E_1. By Theorem B.2.4, it holds with probability at least $1 - \varepsilon^{10}$ that

$$E_1 = \|\mu_{s+1} - \widetilde{\mu}_{s+1}\|_2 < \varepsilon_s \le \varepsilon/8 \cdot 2^{-s}, \tag{B.4.2}$$

where ε_s is given in (3.3.2).

Upper Bound of E_2. By the triangle inequality, we have

$$\begin{aligned} E_2 &= \Bigl\|(I - A + BK_s)^{-1}(Bb_s + \overline{A}\mu_s + d) - (I - A + BK^*)^{-1} \cdot \bigl[Bb^*(\mu_s) + \overline{A}\mu_s + d\bigr]\Bigr\|_2 \\ &\le \bigl\|Bb^*(\mu_s) + \overline{A}\mu_s + d\bigr\|_2 \cdot \Bigl\|\bigl[I - A + BK^* + B(K_s - K^*)\bigr]^{-1} - (I - A + BK^*)^{-1}\Bigr\|_2 \\ &\quad + \bigl\|(I - A + BK_s)^{-1}\bigr\|_2 \cdot \|B\|_2 \cdot \bigl\|b_s - b^*(\mu_s)\bigr\|_2. \end{aligned} \tag{B.4.3}$$

By Taylor's expansion, we have

$$
\begin{aligned}
&\Big\|\big[I - A + BK^* + B(K_s - K^*)\big]^{-1} - (I - A + BK^*)^{-1}\Big\|_2 \\
&\qquad = \Big\|(I - A + BK^*)^{-1}\big[I + (I - A + BK^*)^{-1}B(K_s - K^*)\big]^{-1} - (I - A + BK^*)^{-1}\Big\|_2 \\
&\qquad \le 2\big\|(I - A + BK^*)^{-1}B(K_s - K^*)(I - A + BK^*)^{-1}\big\|_2. \qquad \text{(B.4.4)}
\end{aligned}
$$

Meanwhile, by Taylor's expansion, it holds with probability at least $1 - \varepsilon^{10}$ that

$$
\begin{aligned}
&\big\|(I - A + BK_s)^{-1}\big\|_2 \\
&\qquad = \Big\|\big(I - A + BK^* + B(K_s - K^*)\big)^{-1}\Big\|_2 \\
&\qquad = \Big\|(I - A + BK^*)^{-1}\big(I + (I - A + BK^*)^{-1}B(K_s - K^*)\big)^{-1}\Big\|_2 \\
&\qquad \le \big[1 - \rho(A - BK^*)\big]^{-1} \cdot \Big(1 + \big\|(I - A + BK^*)^{-1}B\big\|_2 \cdot \|K^* - K_s\|_2\Big) \\
&\qquad \le 2\big[1 - \rho(A - BK^*)\big]^{-2}, \qquad \text{(B.4.5)}
\end{aligned}
$$

where the last inequality comes from Theorem B.2.4. By plugging (B.4.4) and (B.4.5) in (B.4.3), it holds with probability at least $1 - \varepsilon^{10}$ that

$$
\begin{aligned}
E_2 &\le 2\big\|Bb^*(\mu_s) + \overline{A}\mu_s + d\big\|_2 \cdot \big\|(I - A + BK^*)^{-1}B(K_s - K^*)(I - A + BK^*)^{-1}\big\|_2 \\
&\qquad + \big\|(I - A + BK_s)^{-1}\big\|_2 \cdot \|B\|_2 \cdot \big\|b_s - b^*(\mu_s)\big\|_2 \\
&\le 2\big\|Bb^*(\mu_s) + \overline{A}\mu_s + d\big\|_2 \cdot \big[1 - \rho(A - BK^*)\big]^{-2} \cdot \|B\|_2 \cdot \|K_s - K^*\|_2 \qquad \text{(B.4.6)} \\
&\qquad + 2\big[1 - \rho(A - BK^*)\big]^{-2} \cdot \|B\|_2 \cdot \big\|b_s - b^*(\mu_s)\big\|_2.
\end{aligned}
$$

By Proposition 3.2.4, it holds that

$$\begin{aligned}\big\|Bb^*(\mu_s) + \overline{A}\mu_s + d\big\|_2 &\le L_1 \cdot \|B\|_2 \cdot \|\mu_s\|_2 + \|\overline{A}\|_2 \cdot \|\mu_s\|_2 + \|d\|_2 \\ &\le \big(L_1 \cdot \|B\|_2 + \|\overline{A}\|_2\big) \cdot \|\mu_s\|_2 + \|d\|_2, \qquad \text{(B.4.7)}\end{aligned}$$

where the scalar L_1 is defined in Assumption 3.2.1. Meanwhile, by Theorem B.2.4, it holds with probability at least $1 - \varepsilon^{10}$ that

$$\text{(B.4.8)} \quad \|K_s - K^*\|_{\mathrm{F}} \le \big[\sigma_{\min}^{-1}(\Psi_\epsilon) \cdot \sigma_{\min}^{-1}(R) \cdot \varepsilon_s\big]^{1/2}, \qquad \big\|b_s - b^*(\mu_s)\big\|_2 \le M_b(\mu_s) \cdot \varepsilon_s^{1/2},$$

where $M_b(\mu_s)$ is defined in (3.3.3). Combining (B.4.6), (B.4.7), (B.4.8), and the choice of ε_s in (3.3.2), it holds with probability at least $1 - \varepsilon^{10}$ that

$$\text{(B.4.9)} \qquad E_2 \le \varepsilon/8 \cdot 2^{-s}.$$

Upper Bound of E_3. By Proposition 3.2.2, we have

$$\text{(B.4.10)} \qquad E_3 = \|\mu_{s+1}^* - \mu^*\|_2 = \big\|\Lambda(\mu_s) - \Lambda(\mu^*)\big\|_2 \le L_0 \cdot \|\mu_s - \mu^*\|_2,$$

where $L_0 = L_1 L_3 + L_2$ by Assumption 3.2.1.

By plugging (B.4.2), (B.4.9), and (B.4.10) in (B.4.1), we know that

$$\text{(B.4.11)} \qquad \|\mu_{s+1} - \mu^*\|_2 \le L_0 \cdot \|\mu_s - \mu^*\|_2 + \varepsilon \cdot 2^{-s-2},$$

which holds with probability at least $1-\varepsilon^{10}$. Following from (B.4.11) and a union bound argument with $S=\mathcal{O}(\log(1/\varepsilon))$, it holds with probability at least $1-\varepsilon^5$ that

$$\|\mu_S-\mu^*\|_2 \le L_0^S \cdot \|\mu_0-\mu^*\|_2+\varepsilon/2,$$

where we use the fact that $L_0<1$ by Assumption 3.2.1. By the choice of S in (3.3.1), it further holds with probability at least $1-\varepsilon^6$ that

$$\|\mu_S-\mu^*\| \le \varepsilon. \tag{B.4.12}$$

By Theorem B.2.4 and the choice of ε_s in (3.3.2), it holds with probability at least $1-\varepsilon^5$ that

$$\|K_S-K^*\|_{\mathrm{F}}=\big\|K_S-K^*(\mu_S)\big\|_{\mathrm{F}} \le \big[\sigma_{\min}^{-1}(\Psi_\epsilon)\cdot\sigma_{\min}^{-1}(R)\cdot\varepsilon_S\big]^{1/2} \le \varepsilon. \tag{B.4.13}$$

Meanwhile, by the triangle inequality and the choice of ε_s in (3.3.2), it holds with probability at least $1-\varepsilon^5$ that

$$\begin{aligned}\|b_S-b^*\|_2 &\le \big\|b_S-b^*(\mu_S)\big\|_2+\big\|b^*(\mu_S)-b^*\big\|_2\\ &\le M_b(\mu_S)\cdot\varepsilon_S^{1/2}+L_1\cdot\|\mu_S-\mu^*\|_2\\ &\le (1+L_1)\cdot\varepsilon, \qquad\qquad \text{(B.4.14)}\end{aligned}$$

where the second inequality comes from Theorem B.2.4 and Proposition 3.2.4, and the last inequality comes from (B.4.12). By (B.4.12), (B.4.13), and (B.4.14), we conclude the proof of the theorem.

B.4.2. Proof of Theorem B.2.4

Proof. We first show that $J_1(K_N) - J_1(K^*) < \varepsilon/2$ with a high probability, then show that $J_2(K_N, b_H) - J_2(K^*, b^*) < \varepsilon/2$ with a high probability. Then we have

$$J(K_N, b_N) - J(K^*, b^*) = J_1(K_N) + J_2(K_N, b_H) - J_1(K^*) - J_2(K^*, b^*) < \varepsilon$$

with a high probability, which proves Theorem B.2.4.

Part 1. We show that $J_1(K_N) - J_1(K^*) < \varepsilon/2$ with a high probability.

We first bound $J_1(K_1) - J_1(K_2)$ for any K_1 and K_2. By Proposition B.2.2, $J_1(K)$ takes the form of

$$J_1(K) = \mathrm{tr}(P_K \Psi_\epsilon) = \mathbb{E}_{y \sim \mathcal{N}(0, \Psi_\epsilon)}(y^\top P_K y). \tag{B.4.15}$$

The following lemma calculates $y^\top P_{K_1} y - y^\top P_{K_2} y$ for any K_1 and K_2.

Lemma B.4.1. Assume that $\rho(A - BK_1) < 1$ and $\rho(A - BK_2) < 1$. For any state vector y, we denote by $\{y_t\}_{t \geq 0}$ the sequence generated by the state transition $y_{t+1} = (A - BK_2) y_t$ with initial state $y_0 = y$. It holds that

$$y^\top P_{K_2} y - y^\top P_{K_1} y = \sum_{t \geq 0} D_{K_1, K_2}(y_t),$$

where

$$D_{K_1, K_2}(y) = 2y^\top (K_2 - K_1)(\Upsilon_{K_1}^{22} K_1 - \Upsilon_{K_1}^{21}) y + y^\top (K_2 - K_1)^\top \Upsilon_{K_1}^{22} (K_2 - K_1) y.$$

Here Υ_K is defined in (3.2.7).

Proof. See §B.6.1 for a detailed proof. □

The following lemma shows that $J_1(K)$ is gradient dominant.

Lemma B.4.2. Let K^* be the optimal parameter and K be a parameter such that $J_1(K) < \infty$, then it holds that

$$J_1(K) - J_1(K^*) \leq \sigma_{\min}^{-1}(R) \cdot \|\Phi_{K^*}\|_2 \cdot \operatorname{tr}\bigl[(\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22} K - \Upsilon_K^{21})\bigr], \tag{B.4.16}$$

$$J_1(K) - J_1(K^*) \geq \sigma_{\min}(\Psi_\omega) \cdot \|\Upsilon_K^{22}\|_2^{-1} \cdot \operatorname{tr}\bigl[(\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22} K - \Upsilon_K^{21})\bigr]. \tag{B.4.17}$$

Proof. See §B.6.2 for a detailed proof. □

Recall that from Algorithm 3, the parameter K is updated via

$$K_{n+1} = K_n - \gamma \cdot (\widehat{\Upsilon}_{K_n}^{22} K_n - \widehat{\Upsilon}_{K_n}^{21}), \tag{B.4.18}$$

where $\widehat{\Upsilon}_{K_n}$ is the output of Algorithm 7. We upper bound $|J_1(K_{n+1}) - J_1(K^*)|$ in the sequel. First, we show that if $J_1(K_n) - J_1(K^*) \geq \varepsilon/2$ holds for any $n \leq N$, we obtain that

$$J_1(K_N) \leq J_1(K_{N-1}) \leq \cdots \leq J_1(K_0), \tag{B.4.19}$$

which holds with probability at least $1 - \varepsilon^{13}$. We prove (B.4.19) by mathematical induction. Suppose that

$$J_1(K_n) \leq J_1(K_{n-1}) \leq \cdots \leq J_1(K_0), \tag{B.4.20}$$

which holds for $n = 0$. In what follows, we define $\widetilde{K}_{n+1}$ as

$$\widetilde{K}_{n+1} = K_n - \gamma \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n}), \tag{B.4.21}$$

where Υ_{K_n} is given in (3.2.7). By (B.4.21), we have

$$\begin{aligned} J_1(\widetilde{K}_{n+1}) - J_1(K_n) &= \mathbb{E}_{y\sim\mathcal{N}(0,\Psi_\epsilon)}\big[y^\top (P_{\widetilde{K}_{n+1}} - P_{K_n}) y\big] \\ &= -2\gamma \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big] \\ &\qquad + \gamma^2 \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top \Upsilon^{22}_{K_n} (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big] \\ &\le -2\gamma \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big] \qquad \text{(B.4.22)} \\ &\qquad + \gamma^2 \cdot \|\Upsilon^{22}_{K_n}\|_2 \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})^\top (\Upsilon^{22}_{K_n} K_n - \Upsilon^{21}_{K_n})\big], \end{aligned}$$

where the first equality comes from (B.4.15), the second equality comes from Lemma B.4.1, and the last inequality comes from the trace inequality. By the definition of Υ_K in (3.2.7), we obtain that

$$\begin{aligned} \|\Upsilon^{22}_{K_n}\|_2 &\le \|R\|_2 + \|B\|_2^2 \cdot \|P_{K_n}\|_2 \le \|R\|_2 + \|B\|_2^2 \cdot J_1(K_n) \cdot \sigma^{-1}_{\min}(\Psi_\epsilon) \\ &\le \|R\|_2 + \|B\|_2^2 \cdot J_1(K_0) \cdot \sigma^{-1}_{\min}(\Psi_\epsilon), \qquad \text{(B.4.23)} \end{aligned}$$

where the second inequality comes from Proposition B.2.2. By plugging (B.4.23) and the choice of stepsize $\gamma \leq [\|R\|_2 + \|B\|_2^2 \cdot J_1(K_0) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon)]^{-1}$ into (B.4.22), we obtain that

$$\begin{aligned} J_1(\widetilde{K}_{n+1}) - J_1(K_n) &\leq -\gamma \cdot \mathrm{tr}\big[\Phi_{\widetilde{K}_{n+1}} \cdot (\Upsilon_{K_n}^{22} K_n - \Upsilon_{K_n}^{21})^\top (\Upsilon_{K_n}^{22} K_n - \Upsilon_{K_n}^{21})\big] \\ &\leq -\gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \mathrm{tr}\big[(\Upsilon_{K_n}^{22} K_n - \Upsilon_{K_n}^{21})^\top (\Upsilon_{K_n}^{22} K_n - \Upsilon_{K_n}^{21})\big] \\ &\leq -\gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_2^{-1} \cdot \big[J_1(K_n) - J_1(K^*)\big] < 0, \end{aligned} \tag{B.4.24}$$

where the last inequality comes from Lemma B.4.2.

The following lemma upper bounds $|J_1(\widetilde{K}_{n+1}) - J_1(K_{n+1})|$.

Lemma B.4.3. Assume that $J_1(K_n) \leq J_1(K_0)$. It holds with probability at least $1 - \varepsilon^{15}$ that

$$\big|J_1(\widetilde{K}_{n+1}) - J_1(K_{n+1})\big| \leq \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_2^{-1} \cdot \varepsilon/4,$$

where K_{n+1} and $\widetilde{K}_{n+1}$ are defined in (B.4.18) and (B.4.21), respectively.

Proof. See §B.6.3 for a detailed proof. □

Combining (B.4.24) and Lemma B.4.3, if $J_1(K_n) - J_1(K^*) \geq \varepsilon/2$, it holds with probability at least $1 - \varepsilon^{15}$ that

$$\begin{aligned} J_1(K_{n+1}) - J_1(K_n) &\leq J_1(\widetilde{K}_{n+1}) - J_1(K_n) + \big|J_1(\widetilde{K}_{n+1}) - J_1(K_{n+1})\big| \\ &\leq -\gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_2^{-1} \cdot \varepsilon/4 < 0. \end{aligned} \tag{B.4.25}$$

Combining (B.4.20) and (B.4.25), it holds with probability at least $1-\varepsilon^{15}$ that

$$J_1(K_{n+1}) \le J_1(K_n) \le \cdots \le J_1(K_0).$$

Finally, following from a union bound argument and the choice of N in Theorem B.2.4, if $J_1(K_n) - J_1(K^*) \ge \varepsilon/2$ holds for any $n \le N$, we have

$$J_1(K_N) \le J_1(K_{N-1}) \le \cdots \le J_1(K_0),$$

which holds with probability at least $1-\varepsilon^{13}$. Thus, we complete the proof of (B.4.19).

Combining (B.4.24) and (B.4.25), for $J_1(K_n) - J_1(K^*) \ge \varepsilon/2$, we have

$$J_1(K_{n+1}) - J_1(K^*) \le \bigl[1 - \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_2^{-1}\bigr] \cdot \bigl[J_1(K_n) - J_1(K^*)\bigr],$$

which holds with probability at least $1-\varepsilon^{13}$. Meanwhile, following from a union bound argument and the choice of N in Theorem B.2.4, it holds with probability at least $1-\varepsilon^{11}$ that

$$J_1(K_N) - J_1(K^*) \le \varepsilon/2. \tag{B.4.26}$$

The following lemma upper bounds $\|K_N - K^*\|_{\mathrm{F}}$.

Lemma B.4.4. For any K, we have

$$\|K - K^*\|_{\mathrm{F}}^2 \le \sigma_{\min}^{-1}(\Psi_\epsilon) \cdot \sigma_{\min}^{-1}(R) \cdot \bigl[J_1(K) - J_1(K^*)\bigr].$$

Proof. See §B.6.4 for a detailed proof. □

Combining (B.4.26) and Lemma B.4.4, we have

$$\|K_N - K^*\|_{\mathrm{F}} \le \big[\sigma_{\min}^{-1}(\Psi_\epsilon) \cdot \sigma_{\min}^{-1}(R) \cdot \varepsilon/2\big]^{1/2}, \tag{B.4.27}$$

which holds with probability $1 - \varepsilon^{11}$.

Part 2. We show that $J_2(K_N, b_H) - J_2(K^*, b^*) < \varepsilon/2$ with high probability. Following from Proposition 3.2.4, it holds that $J_2(K^*, b^*) = J_2(K_N, b^{K_N})$. Therefore, it suffices to show that $J_2(K_N, b_H) - J_2(K_N, b^{K_N}) < \varepsilon/2$.

First, we show that if $J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \ge \varepsilon/2$ for any $h \le H$, we obtain that

$$J_2(K_N, b_H) \le J_2(K_N, b_{H-1}) \le \cdots \le J_2(K_N, b_1) \le J_2(K_N, b_0), \tag{B.4.28}$$

which holds with probability at least $1 - \varepsilon^{13}$. We prove (B.4.28) by mathematical induction. Suppose that

$$J_2(K_N, b_h) \le J_2(K_N, b_{h-1}) \le \cdots \le J_2(K_N, b_0), \tag{B.4.29}$$

Recall that by Algorithm 3, the parameter b is updated via

$$b_{h+1} = b_h - \gamma^b \cdot \widehat{\nabla}_b J_2(K_N, b_h). \tag{B.4.30}$$

Here

$$\widehat{\nabla}_b J_2(K_N, b_h) = \widehat{\Upsilon}_{K_N}^{22}(-K_N \widehat{\mu}_{K_N,b_h} + b_h) + \widehat{\Upsilon}_{K_N}^{21} \widehat{\mu}_{K_N,b_h} + \widehat{q}_{K_N,b_h}, \tag{B.4.31}$$

where $\widehat{\Upsilon}_{K_N}$ and $\widehat{q}_{K_N,b_h}$ are the outputs of Algorithm 7. We define $\widetilde{b}_{h+1}$ as

$$\widetilde{b}_{h+1} = b_h - \gamma^b \cdot \nabla_b J_2(K_N, b_h). \tag{B.4.32}$$

Here

$$\nabla_b J_2(K_N, b_h) = \Upsilon^{22}_{K_N}(-K_N \mu_{K_N,b_h} + b_h) + \Upsilon^{21}_{K_N}\mu_{K_N,b_h} + q_{K_N,b_h}, \tag{B.4.33}$$

where Υ_{K_N} and q_{K_N,b_h} are defined in (3.2.7). We upper bound $J_2(K_N, b_{h+1}) - J_2(K_N, b^{K_N})$ in the sequel. Following from (B.4.32) and Proposition 3.2.3, we have

$$\begin{aligned} J_2(K_N, \widetilde{b}_{h+1}) - J_2(K_N, b_h) &\leq -\gamma^b/2 \cdot \big\|\nabla_b J_2(K_N, b_h)\big\|_2^2 \\ &\leq -\nu_{K_N} \cdot \gamma^b \cdot \big[J_2(K_N, b_h) - J_2(K_N, b^{K_N})\big] \\ &\leq -\nu_{K_N} \cdot \gamma^b \cdot \varepsilon < 0, \end{aligned} \tag{B.4.34}$$

where ν_{K_N} is specified in Proposition 3.2.3. The following lemma upper bounds $|J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1})|$.

Lemma B.4.5. Assume that $J_2(K_N, b_h) \leq J_2(K_N, b_0)$. It holds with probability at least $1 - \varepsilon^{15}$ that

$$\big|J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1})\big| \leq \nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2,$$

where b_{h+1} and $\widetilde{b}_{h+1}$ are defined in (B.4.30) and (B.4.32), respectively.

Proof. See §B.6.5 for a detailed proof. □

Combining (B.4.34) and Lemma B.4.5, we know that if $J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \geq \varepsilon$, it holds with probability at least $1 - \varepsilon^{15}$ that

$$J_2(K_N, b_{h+1}) - J_2(K_N, b_h) \leq J_2(K_N, \widetilde{b}_{h+1}) - J_2(K_N, b_h) + \big|J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1})\big|$$

$$\leq -\nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2 < 0. \tag{B.4.35}$$

Combining (B.4.29) and (B.4.35), it holds with probability at least $1 - \varepsilon^{15}$ that

$$J_2(K_N, b_{h+1}) \leq J_2(K_N, b_h) \leq \cdots \leq J_2(K_N, b_0).$$

Following from a union bound argument and the choice of H in Theorem B.2.4, if $J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \geq \varepsilon$ holds for any $h \leq H$, we have

$$J_2(K_N, b_H) \leq J_2(K_N, b_{H-1}) \leq \cdots \leq J_2(K_N, b_0),$$

which holds with probability at least $1 - \varepsilon^{13}$. Thus, we finish the proof of (B.4.28).

Combining (B.4.34) and Lemma B.4.5, for $J_2(K_N, b_h) - J_2(K_N, b^{K_N}) \geq \varepsilon/2$, we have

$$J_2(K_N, b_{h+1}) - J_2(K_N, b^{K_N}) \leq (1 - \nu_{K_N} \cdot \gamma^b) \cdot \big[J_2(K_N, b_h) - J_2(K_N, b^{K_N})\big],$$

which holds with probability at least $1 - \varepsilon^{13}$. Meanwhile, following from a union bound argument and the choice of H in Theorem B.2.4, it holds with probability at least $1 - \varepsilon^{11}$ that

$$J_2(K_N, b_H) - J_2(K_N, b^{K_N}) \leq \varepsilon/2. \tag{B.4.36}$$

By Proposition 3.2.3 and (B.4.36), it holds with probability at least $1-\varepsilon^{11}$ that

$$\|b_H - b^{K_N}\|_2 \leq (2\varepsilon/\nu_{K^*})^{1/2}. \tag{B.4.37}$$

Following from Proposition 3.2.4, we know that

$$\begin{aligned} b^{K_N} - b^* = {} & (K_N - K^*)Q^{-1}(I-A)^\top \\ & \cdot \big[(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top\big]^{-1} \cdot (\overline{A}\mu + d). \end{aligned} \tag{B.4.38}$$

Combining (B.4.27), (B.4.37), and (B.4.38), it holds with probability $1-\varepsilon^{10}$ that

$$\|b_H - b^{K_N}\|_2 \leq M_b \cdot \varepsilon^{1/2},$$

where

$$\begin{aligned} M_b(\mu) = 4 & \Big\| Q^{-1}(I-A)^\top \cdot \big[(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top\big]^{-1} \cdot (\overline{A}\mu + d)\Big\|_2 \\ & \cdot \big[\nu_{K^*}^{-1} + \sigma_{\min}^{-1}(\Psi_\epsilon) \cdot \sigma_{\min}^{-1}(R)\big]^{1/2}. \end{aligned}$$

We finish the proof of the theorem. □

B.4.3. Proof of Theorem B.2.8

Proof. We follow the proof of Theorem 4.2 in Yang et al. (2019b), where they only consider LQR without drift terms. Since our proof requires much more delicate analysis, we present it here.

Part 1. We denote by $\widehat{\zeta}$ and $\widehat{\xi}$ the primal and dual variables generated by Algorithm 7. We define the primal-dual gap of (B.2.11) as

$$\text{gap}(\widehat{\zeta}, \widehat{\xi}) = \max_{\xi \in \mathcal{V}_\xi} F(\widehat{\zeta}, \xi) - \min_{\zeta \in \mathcal{V}_\zeta} F(\zeta, \widehat{\xi}). \tag{B.4.39}$$

In the sequel, we upper bound $\|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2$ using (B.4.39).

We define $\zeta_{K,b}$ and $\xi(\zeta)$ as

$$\zeta_{K,b} = \bigl(J(K,b), \alpha_{K,b}^\top\bigr)^\top, \qquad \xi(\zeta) = \operatorname*{argmax}_{\xi} F(\zeta, \xi). \tag{B.4.40}$$

Following from (B.2.12), we know that

(B.4.41)

$$\xi^1(\zeta) = \zeta^1 - J(K,b), \quad \xi^2(\zeta) = \mathbb{E}_{\pi_{K,b}}\bigl[\psi(x,u)\bigr]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\bigl[c(x,u)\psi(x,u)\bigr].$$

The following lemma shows that $\zeta_{K,b} \in \mathcal{V}_\zeta$ and $\xi(\zeta) \in \mathcal{V}_\xi$ for any $\zeta \in \mathcal{V}_\zeta$.

Lemma B.4.6. Under the assumptions in the statement of Theorem B.2.8, we have $\zeta_{K,b} = (J(K,b), \alpha_{K,b}^\top)^\top \in \mathcal{V}_\zeta$. Also, for any $\zeta \in \mathcal{V}_\zeta$, the vector $\xi(\zeta)$ defined in (B.4.40) satisfies that $\xi(\zeta) \in \mathcal{V}_\xi$.

Proof. See §B.6.6 for a detailed proof. □

By (B.2.12), we know that $\nabla_\zeta F(\zeta_{K,b}, 0) = 0$ and $\nabla_\xi F(\zeta_{K,b}, 0) = 0$. Combining Lemma B.4.6, it holds that $(\zeta_{K,b}, 0)$ is a saddle point of the function $F(\zeta, \xi)$ defined in (B.2.11).

Following from (B.4.39), it holds that

$$\Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\widehat{\zeta}^1 + \Theta_{K,b}\widehat{\zeta}^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2^2 + \big|\widehat{\zeta}^1 - J(K,b)\big|^2$$

$$= F\big(\widehat{\zeta}, \xi(\widehat{\zeta})\big) = \max_{\xi\in\mathcal{V}_\xi} F(\widehat{\zeta}, \xi) = \operatorname{gap}(\widehat{\zeta}, \widehat{\xi}) + \min_{\zeta\in\mathcal{V}_\zeta} F(\zeta, \widehat{\xi}), \tag{B.4.42}$$

where the first equality comes from (B.4.41), and the second equality comes from the fact that $\xi(\widehat{\zeta}) = \operatorname{argmax}_{\xi\in\mathcal{V}_\xi} F(\widehat{\zeta}, \xi)$ by (B.4.40) and Lemma B.4.6. We upper bound the RHS of (B.4.42) and lower bound the LHS of (B.4.42) in the sequel.

As for the RHS of (B.4.42), it holds for any $\xi \in \mathcal{V}_\xi$ that

$$\min_{\zeta\in\mathcal{V}_\zeta} F(\zeta, \xi) \le \min_{\zeta\in\mathcal{V}_\zeta}\max_{\xi\in\mathcal{V}_\xi} F(\zeta, \xi) = \min_{\zeta\in\mathcal{V}_\zeta} F\big(\zeta, \xi(\zeta)\big)$$

$$= \frac{1}{2}\min_{\zeta\in\mathcal{V}_\zeta}\bigg\{\Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2^2 + \big|\zeta^1 - J(K,b)\big|^2\bigg\}$$

$$= 0, \tag{B.4.43}$$

where the first equality comes from the fact that $\xi(\zeta) = \operatorname{argmax}_{\xi\in\mathcal{V}_\xi} F(\zeta, \xi)$ by (B.4.40) and Lemma B.4.6, the second equality comes from (B.4.41), and the last equality holds by taking $\zeta = \zeta_{K,b} \in \mathcal{V}_\zeta$. Meanwhile, we lower bound the LHS of (B.4.42) as

$$\Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\widehat{\zeta}^1 + \Theta_{K,b}\widehat{\zeta}^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2^2 + \big|\widehat{\zeta}^1 - J(K,b)\big|^2$$

$$= \big\|\widetilde{\Theta}_{K,b}(\widehat{\zeta} - \zeta_{K,b})\big\|_2^2 \ge \lambda_K^2 \cdot \|\widehat{\zeta} - \zeta_{K,b}\|_2^2 \ge \lambda_K^2 \cdot \|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2^2, \tag{B.4.44}$$

where the first equality comes from the definition of $\widetilde{\Theta}_{K,b}$ in (B.2.9), and the first inequality comes from Proposition B.2.6. Here λ_K is defined in Proposition B.2.6. Combining

(B.4.42), (B.4.43), and (B.4.44), it holds that

$$\|\widehat{\alpha}_{K,b} - \alpha_{K,b}\|_2^2 \le \lambda_K^{-2} \cdot \mathrm{gap}(\widehat{\zeta}, \widehat{\xi}), \tag{B.4.45}$$

which finishes the proof of this part.

Part 2. We now upper bound $\mathrm{gap}(\widehat{\zeta}, \widehat{\xi})$. We denote by $\widetilde{z}_t = (\widetilde{x}_t^\top, \widetilde{u}_t^\top)^\top$ for $t \in [\widetilde{T}]$, where $\widetilde{x}_t$ and $\widetilde{u}_t$ are generated in Line 6 of Algorithm 7. Following from the state transition in Problem 3.1.3 and the form of the linear policy, $\{\widetilde{z}_t\}_{t\in[\widetilde{T}]}$ follows the following transition,

$$\widetilde{z}_{t+1} = L\widetilde{z}_t + \nu + \delta_t, \tag{B.4.46}$$

where

$$\nu = \begin{pmatrix} \overline{A}\mu + d \\ -K(\overline{A}\mu + d) + b \end{pmatrix}, \qquad \delta_t = \begin{pmatrix} \omega_t \\ -K\omega_t + \sigma\eta \end{pmatrix}, \qquad L = \begin{pmatrix} A & B \\ -KA & -KB \end{pmatrix}.$$

Note that we have

$$L = \begin{pmatrix} A & B \\ -KA & -KB \end{pmatrix} = \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix}.$$

Then by the property of spectral radius, it holds that

$$\rho(L) = \rho\left(\begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} I \\ -K \end{pmatrix}\right) = \rho(A - BK) < 1.$$

Thus, the Markov chain generated by (B.4.46) admits a unique stationary distribution $\mathcal{N}(\mu_z, \Sigma_z)$, where

$$\mu_z = (I - L)^{-1}\nu, \qquad \Sigma_z = L\Sigma_z L^\top + \begin{pmatrix} \Psi_\omega & -\Psi_\omega K^\top \\ -K\Psi_\omega & K\Psi_\omega K^\top + \sigma^2 I \end{pmatrix}. \tag{B.4.47}$$

The following lemma characterizes the average

$$\widehat{\mu}_z = 1/\widetilde{T} \cdot \sum_{t=1}^{\widetilde{T}} \widetilde{z}_t. \tag{B.4.48}$$

Lemma B.4.7. It holds that

$$\widehat{\mu}_z \sim \mathcal{N}\Big(\mu_z + \frac{1}{\widetilde{T}}\mu_{\widetilde{T}},\ \frac{1}{\widetilde{T}}\widetilde{\Sigma}_{\widetilde{T}}\Big),$$

where $\|\mu_{\widetilde{T}}\|_2 \le M_\mu \cdot (1-\rho)^{-2} \cdot \|\mu_z\|_2$ and $\|\widetilde{\Sigma}_{\widetilde{T}}\|_{\mathrm{F}} \le M_\Sigma \cdot (1-\rho)^{-1} \cdot \|\Sigma_z\|_{\mathrm{F}}$. Here M_μ and M_Σ are positive absolute constants. Moreover, it holds with probability at least $1 - \widetilde{T}^{-6}$ that

$$\|\widehat{\mu}_z - \mu_z\|_2 \le \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}} \cdot (1-\rho)^{-2} \cdot \mathrm{poly}\big(\|\Phi_K\|_2, \|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2\big).$$

Proof. See §B.6.7 for a detailed proof. □

Lemma B.4.7 gives that

$$\|\widehat{\mu}_{K,b} - \mu_{K,b}\|_2 \le \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}} \cdot (1-\rho)^{-2} \cdot \mathrm{poly}\big(\|\Phi_K\|_2, \|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2\big),$$

which holds with probability at least $1 - \widetilde{T}^{-6}$.

We now apply a truncation argument to show that $\operatorname{gap}(\widehat{\zeta}, \widehat{\xi})$ is upper bounded. We define the event $\mathcal{E}$ in the sequel. Following from Lemma B.4.7, it holds for any $z \sim \mathcal{N}(\mu_z, \Sigma_z)$ that

$$z - \widehat{\mu}_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}} \sim \mathcal{N}(0, \Sigma_z + 1/\widetilde{T} \cdot \widetilde{\Sigma}_{\widetilde{T}}).$$

By Lemma B.7.3, there exists a positive absolute constant C_0 such that

(B.4.49)

$$\mathbb{P}\Big[\big|\|z - \widehat{\mu}_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2^2 - \operatorname{tr}(\widetilde{\Sigma}_z)\big| > \tau\Big] \le 2\exp\Big[-C_0 \cdot \min\big(\tau^2 \|\widetilde{\Sigma}_z\|_{\mathrm{F}}^{-2},\ \tau\|\widetilde{\Sigma}_z\|_2^{-1}\big)\Big],$$

where we write $\widetilde{\Sigma}_z = \Sigma_z + 1/\widetilde{T} \cdot \widetilde{\Sigma}_{\widetilde{T}}$ for notational convenience. By taking $\tau = C_1 \cdot \log T \cdot \|\widetilde{\Sigma}_z\|_{\mathrm{F}}$ in (B.4.49) for a sufficiently large positive absolute constant C_1, it holds that

$$\mathbb{P}\Big[\big|\|z - \widehat{\mu}_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2^2 - \operatorname{tr}(\widetilde{\Sigma}_z)\big| > C_1 \cdot \log T \cdot \|\widetilde{\Sigma}_z\|_{\mathrm{F}}\Big] \le T^{-6}. \tag{B.4.50}$$

We define the event $\mathcal{E}_{t,1}$ for any $t \in [T]$ as

$$\mathcal{E}_{t,1} = \Big\{\big|\|z_t - \widehat{\mu}_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2^2 - \operatorname{tr}(\widetilde{\Sigma}_z)\big| \le C_1 \cdot \log T \cdot \|\widetilde{\Sigma}_z\|_{\mathrm{F}}\Big\}.$$

Then by (B.4.50), it holds for any $t \in [T]$ that

$$\mathbb{P}(\mathcal{E}_{t,1}) \ge 1 - T^{-6}. \tag{B.4.51}$$

Also, we define

$$\mathcal{E}_1 = \bigcap_{t \in [T]} \mathcal{E}_{t,1}. \tag{B.4.52}$$

Following from a union bound argument and (B.4.51), it holds that

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - T^{-5}. \tag{B.4.53}$$

Also, conditioning on $\mathcal{E}_1$, it holds for sufficiently large $\widetilde{T}$ that

$$\begin{aligned}
&\max_{t\in[T]} \|z_t - \widehat{\mu}_z\|_2^2 \\
&\qquad \leq C_1 \cdot \log T \cdot \|\widetilde{\Sigma}_z\|_{\mathrm{F}} + \mathrm{tr}(\widetilde{\Sigma}_z) + \|1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2^2 \\
&\qquad \leq 2\widetilde{C}_1 \cdot \big[1 + M_\Sigma(1-\rho)^{-1}/\widetilde{T}^2\big] \cdot \log T \cdot \|\Sigma_z\|_2 + M_\mu(1-\rho)^{-2}/\widetilde{T}^2 \cdot \|\mu_z\|_2^2 \\
&\qquad \leq C_2 \cdot \log T \cdot \big(1 + \|K\|_{\mathrm{F}}^2\big) \cdot \|\Phi_K\|_2 \cdot (1-\rho)^{-1} + C_3 \cdot \big(\|b\|_2^2 + \|\mu\|_2^2\big) \cdot (1-\rho)^{-4} \cdot \widetilde{T}^{-2} \\
&\qquad \leq 2C_2 \cdot \log T \cdot \big(1 + \|K\|_{\mathrm{F}}^2\big) \cdot \|\Phi_K\|_2 \cdot (1-\rho)^{-1},
\end{aligned} \tag{B.4.54}$$

where $\widetilde{C}_1$, C_2, and C_3 are positive absolute constants. Here, the first inequality comes from the definition of $\mathcal{E}_1$ in (B.4.52), the second inequality comes from Lemma B.4.7, and the third inequality comes from (B.4.47). Also, we define the following event

$$\mathcal{E}_2 = \big\{\|\widehat{\mu}_z - \mu_z + 1/\widetilde{T} \cdot \mu_{\widetilde{T}}\|_2 \leq C_1\big\}. \tag{B.4.55}$$

Then by Lemma B.4.7, we know that

$$\mathbb{P}(\mathcal{E}_2) \geq 1 - \widetilde{T}^{-6} \tag{B.4.56}$$

for $\widetilde{T}$ sufficiently large. We define the event $\mathcal{E}$ as

$$\mathcal{E} = \mathcal{E}_1 \bigcap \mathcal{E}_2.$$

Then following from (B.4.53), (B.4.56), and a union bound argument, we know that

$$\mathbb{P}(\mathcal{E}) \geq 1 - T^{-5} - \widetilde{T}^{-6}.$$

Now, we define the truncated feature vector $\widetilde{\psi}(x,u)$ as $\widetilde{\psi}(x,u) = \widehat{\psi}(x,u)\,\mathbb{1}_{\mathcal{E}}$, the truncated cost function $\widetilde{c}(x,u)$ as $\widetilde{c}(x,u) = c(x,u)\,\mathbb{1}_{\mathcal{E}}$, and also the truncated objective function $\widetilde{F}(\zeta,\xi)$ as

(B.4.57)

$$\widetilde{F}(\zeta,\xi) = \Bigl\{\mathbb{E}(\widetilde{\psi})\zeta^1 + \mathbb{E}\bigl[(\widetilde{\psi} - \widetilde{\psi}')\widetilde{\psi}^\top\bigr]\zeta^2 - \mathbb{E}(\widetilde{c}\widetilde{\psi})\Bigr\}^\top \xi^2 + \bigl[\zeta^1 - \mathbb{E}(\widetilde{c})\bigr]\cdot\xi^1 - \|\xi\|_2^2/2,$$

where we write $\widetilde{\psi} = \widetilde{\psi}(x,u)$ and $\widetilde{c} = \widetilde{c}(x,u)$ for notational convenience. Here the expectation is taken following the policy $\pi_{K,b}$ and the state transition. The following lemma establishes the upper bound of $|F(\zeta,\xi) - \widetilde{F}(\zeta,\xi)|$, where $F(\zeta,\xi)$ and $\widetilde{F}(\zeta,\xi)$ are defined in (B.2.11) and (B.4.57), respectively.

Lemma B.4.8. It holds with probability at least $1 - \widetilde{T}^{-6}$ that

$$\bigl|F(\zeta,\xi) - \widetilde{F}(\zeta,\xi)\bigr| \leq \biggl(\frac{1}{2T} + \frac{\log\widetilde{T}}{\widetilde{T}^{1/4}}\biggr)\cdot(1-\rho)^{-2}\cdot\operatorname{poly}\bigl(\|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0,b_0)\bigr).$$

Proof. See §B.6.8 for a detailed proof. □

Following from (B.4.39) and Lemma B.4.8, it holds with probability at least $1 - \widetilde{T}^{-6}$ that

$$\begin{aligned}
&\bigl|\mathrm{gap}(\widehat{\zeta}, \widehat{\xi}) - \widetilde{\mathrm{gap}}(\widehat{\zeta}, \widehat{\xi})\bigr| \\
&\quad \le \biggl(\frac{1}{2T} + \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}}\biggr) \cdot (1-\rho)^{-2} \cdot \mathrm{poly}\bigl(\|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0, b_0)\bigr). \qquad \text{(B.4.58)}
\end{aligned}$$

where we define $\widetilde{\mathrm{gap}}(\widehat{\zeta}, \widehat{\xi})$ as

$$\widetilde{\mathrm{gap}}(\widehat{\zeta}, \widehat{\xi}) = \max_{\xi \in \mathcal{V}_\xi} \widetilde{F}(\widehat{\zeta}, \xi) - \min_{\zeta \in \mathcal{V}_\zeta} \widetilde{F}(\zeta, \widehat{\xi}).$$

Therefore, to upper bound of $\mathrm{gap}(\zeta, \xi)$, we only need to upper bound $\widetilde{\mathrm{gap}}(\zeta, \xi)$.

Part 3. We upper bound $\widetilde{\mathrm{gap}}(\zeta, \xi)$ in the sequel. We first show that the trajectory generated by the policy $\pi_{K,b}$ and the state transition in Problem 3.1.2 is β-mixing.

Lemma B.4.9. Consider a linear system $y_{t+1} = Dy_t + \vartheta + \upsilon_t$, where $\{y_t\}_{t\ge 0} \subset \mathbb{R}^m$, the matrix $D \in \mathbb{R}^{m\times m}$ satisfying $\rho(D) < 1$, the vector $\vartheta \in \mathbb{R}^m$, and $\upsilon_t \sim \mathcal{N}(0, \Sigma)$ is the Gaussians. We denote by ϖ_t the marginal distribution of y_t for any $t \ge 0$. Meanwhile, assume that the stationary distribution of $\{y_t\}_{t\ge 0}$ is a Gaussian distribution $\mathcal{N}((I - D)^{-1}\vartheta, \Sigma_\infty)$, where Σ_∞ is the covariance matrix. We define the β-mixing coefficients for any $n \ge 1$ as follows

$$\beta(n) = \sup_{t\ge 0} \mathbb{E}_{y\sim\varpi_t}\Bigl[\bigl\|\mathbb{P}_{y_n}(\cdot \,|\, y_0 = y) - \mathbb{P}_{\mathcal{N}((I-D)^{-1}\vartheta, \Sigma_\infty)}(\cdot)\bigr\|_{\mathrm{TV}}\Bigr].$$

Then, for any $\rho \in (\rho(D), 1)$, the β-mixing coefficients satisfy that

$$\beta(n) \le C_{\rho, D, \vartheta} \cdot \bigl[\mathrm{tr}(\Sigma_\infty) + m \cdot (1-\rho)^{-2}\bigr]^{1/2} \cdot \rho^n,$$

where $C_{\rho,D,\vartheta}$ is a constant, which only depends on ρ, D, and ϑ. We say that the sequence $\{y_t\}_{t\geq 0}$ is β-mixing with parameter ρ.

Proof. See Proposition 3.1 in Tu and Recht (2017) for details. □

Recall that by (3.2.1), the sequence $\{x_t\}_{t\geq 0}$ follows

$$x_{t+1} = (A - BK)x_t + (Bb + \overline{A}\mu + d) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \Psi_\epsilon),$$

where the matrix $A - BK$ satisfies that $\rho(A - BK) < 1$. Therefore, by Lemma B.4.9, the sequence $\{z_t\}_{t\geq 0}$ is β-mixing with parameter $\rho \in (\rho(A - BK), 1)$, where $z_t = (x_t^\top, u_t^\top)^\top$. The following lemma upper bounds the primal-dual gap for a convex-concave problem.

Lemma B.4.10. Let $\mathcal{X}$ and $\mathcal{Y}$ be two compact and convex sets such that $\|x - x'\|_2 \leq M$ and $\|y - y'\|_2 \leq M$ for any $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. We consider solving the following minimax problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} H(x, y) = \mathbb{E}_{\epsilon\sim\varpi_\epsilon}\big[G(x, y; \epsilon)\big],$$

where the objective function $H(x, y)$ is convex in x and concave in y. In addition, we assume that the distribution ϖ_ϵ is β-mixing with $\beta(n) \leq C_\epsilon \cdot \rho^n$, where C_ϵ is a constant. Meanwhile, we assume that it holds almost surely that $G(x, y; \epsilon)$ is $\widetilde{L}_0$-Lipschitz in both x and y, the gradient $\nabla_x G(x, y; \epsilon)$ is $\widetilde{L}_1$-Lipschitz in x for any $y \in \mathcal{Y}$, the gradient $\nabla_y G(x, y; \epsilon)$ is $\widetilde{L}_1$-Lipschitz in y for any $x \in \mathcal{X}$, where $C_\epsilon, \widetilde{L}_0, \widetilde{L}_1 > 1$. Each step of our gradient-based method takes the following forms,

$$x_{t+1} = \Gamma_{\mathcal{X}}\big[x_t - \gamma_{t+1} \cdot \nabla_x G(x_t, y_t; \epsilon_t)\big], \qquad y_{t+1} = \Gamma_{\mathcal{Y}}\big[y_t - \gamma_{t+1} \cdot \nabla_y G(x_t, y_t; \epsilon_t)\big],$$

where the operators $\Gamma_{\mathcal{X}}$ and $\Gamma_{\mathcal{Y}}$ projects the variables back to $\mathcal{X}$ and $\mathcal{Y}$, respectively, and the stepsizes take the form $\gamma_t = \gamma_0 \cdot t^{-1/2}$ for a constant $\gamma_0 > 0$. Moreover, let $\widehat{x} = (\sum_{t=1}^{T} \gamma_t)^{-1}(\sum_{t=1}^{T} \gamma_t x_t)$ and $\widehat{y} = (\sum_{t=1}^{T} \gamma_t)^{-1}(\sum_{t=1}^{T} \gamma_t y_t)$ be the final output of the gradient method after T iterations, then there exists a positive absolute constant C, such that for any $\delta \in (0,1)$, the primal-dual gap to the minimax problem is upper bounded as

$$\max_{x\in\mathcal{X}} H(\widehat{x}, y) - \min_{y\in\mathcal{Y}} H(x, \widehat{y}) \leq \frac{C \cdot (M^2 + \widetilde{L}_0^2 + \widetilde{L}_0 \widetilde{L}_1 M)}{\log(1/\rho)} \cdot \frac{\log^2 T + \log(1/\delta)}{\sqrt{T}} + \frac{C \cdot C_\epsilon \widetilde{L}_0 M}{T},$$

which holds with probability at least $1 - \delta$.

Proof. See Theorem 5.4 in Yang et al. (2019b) for details. □

To use Lemma B.4.10, we define the function $G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}')$ as

$$G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}') = \big[\widetilde{\psi}\zeta^1 + (\widetilde{\psi} - \widetilde{\psi}')\widetilde{\psi}^\top \zeta^2 - \widetilde{c}\widetilde{\psi}\big]^\top \xi^2 + (\zeta^1 - \widetilde{c}) \cdot \xi^1 - 1/2 \cdot \|\xi\|_2^2,$$

where $\widetilde{\psi} = \widetilde{\psi}(x, u)$ and $\widetilde{\psi}' = \widetilde{\psi}(x', u')$. Note that the gradients of $G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}')$ take the form

$$\nabla_\zeta G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}') = \begin{pmatrix} \widetilde{\psi}^\top \xi^2 + \xi^1 \\ \widetilde{\psi}(\widetilde{\psi} - \widetilde{\psi}')^\top \xi^2 \end{pmatrix},$$

$$\nabla_\xi G(\zeta, \xi; \widetilde{\psi}, \widetilde{\psi}') = \begin{pmatrix} \zeta^1 - \widetilde{c} - \xi^1 \\ \widetilde{\psi}\zeta^1 + (\widetilde{\psi} - \widetilde{\psi}')\widetilde{\psi}^\top \zeta^2 - \widetilde{c}\widetilde{\psi} - \xi^2 \end{pmatrix}.$$

By Definition B.2.7 and Lemma B.4.6, we know that

$$\big\|\nabla_\zeta G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')\big\|_2 \le \mathrm{poly}\big(\|K\|_{\mathrm{F}}, J(K_0,b_0)\big)\cdot\log^2 T\cdot(1-\rho)^{-2},$$

$$\big\|\nabla_\xi G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')\big\|_2 \le \mathrm{poly}\big(\|K\|_{\mathrm{F}}, \|\mu\|_2, J(K_0,b_0)\big)\cdot\log^2 T\cdot(1-\rho)^{-2}. \tag{B.4.59}$$

This gives the Lipschitz constant $\widetilde{L}_0$ in Lemma B.4.10 for $G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')$. Also, the Hessians of $G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')$ take the forms of

$$\nabla^2_{\zeta\zeta}G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}') = 0,\qquad \nabla^2_{\xi\xi}G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}') = -I,$$

which follows that

$$\big\|\nabla^2_{\zeta\zeta}G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')\big\|_2 = 0,\qquad \big\|\nabla^2_{\xi\xi}G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')\big\|_2 = 1. \tag{B.4.60}$$

This gives the Lipschitz constant $\widetilde{L}_1$ in Lemma B.4.10 for $\nabla_\zeta G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')$ and $\nabla_\xi G(\zeta,\xi;\widetilde{\psi},\widetilde{\psi}')$. Moreover, note that (B.4.54) provides an upper bound of M, combining (B.4.59), (B.4.60) and Lemma B.4.10, it holds with probability at least $1-T^{-5}$ that

$$\widetilde{\mathrm{gap}}(\widehat{\zeta},\widehat{\xi}) \le \frac{\mathrm{poly}\big(\|K\|_{\mathrm{F}},\|\mu\|_2,J(K_0,b_0)\big)\cdot\log^6 T}{(1-\rho)^4\cdot\sqrt{T}}. \tag{B.4.61}$$

Combining (B.4.45), (B.4.58), and (B.4.61), we know that

$$\begin{aligned}&\|\widehat{\alpha}_{K,b}-\alpha_{K,b}\|_2^2\\&\qquad\le \lambda_K^{-2}\cdot\mathrm{poly}_1\big(\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big)\cdot\biggl[\frac{\log^6 T}{T^{1/2}\cdot(1-\rho)^4}+\frac{\log\widetilde{T}}{\widetilde{T}^{1/4}\cdot(1-\rho)^2}\biggr].\end{aligned}$$

Same bounds for $\|\widehat{\Upsilon}_K - \Upsilon_K\|_{\mathrm{F}}^2$, $\|\widehat{p}_{K,b} - p_{K,b}\|_2^2$, and $\|\widehat{q}_{K,b} - q_{K,b}\|_2^2$ hold. We finish the proof of the theorem. □

B.5. Proofs of Propositions

B.5.1. Proof of Proposition 3.2.2

Proof. We follow a similar proof as in the one of Theorem 1.1 in Sznitman (1991) and Theorem 3.2 in Bensoussan et al. (2016). Note that for any policy $\pi_{K,b} \in \Pi$, the parameters K and b uniquely determine the policy. We define the following metric on Π.

Definition B.5.1. For any $\pi_{K_1,b_1}, \pi_{K_2,b_2} \in \Pi$, we define the following metric,

$$\|\pi_{K_1,b_1} - \pi_{K_2,b_2}\|_2 = c_1 \cdot \|K_1 - K_2\|_2 + c_2 \cdot \|b_1 - b_2\|_2,$$

where c_1 and c_2 are positive constants.

One can verify that Definition B.5.1 satisfies the requirement of being a metric. We first evaluate the forms of the operators $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot, \cdot)$.

Forms of the operators $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot, \cdot)$. By the definition of $\Lambda_1(\mu)$, which gives the optimal policy under the mean-field state μ, it holds that

$$\Lambda_1(\mu) = \pi_\mu^*,$$

where π_μ^* solves Problem 3.1.2. This gives the form of $\Lambda_1(\cdot)$. We now turn to $\Lambda_2(\mu, \pi)$, which gives the mean-field state μ_{new} generated by the policy π under the current mean-field state μ. In Problem 3.1.2, the sequence of states $\{x_t\}_{t\geq 0}$ constitutes a Markov chain, which admits a unique stationary distribution. Thus, by the state transition in Problem

3.1.2 and the form of the linear-Gaussian policy, we have

$$\mu_{\text{new}} = (A - BK_\pi)\mu_{\text{new}} + (Bb_\pi + \overline{A}\mu + d), \tag{B.5.1}$$

where K_π and b_π are parameters of the policy π. By solving (B.5.1) for μ_{new}, it holds that

$$\Lambda_2(\mu, \pi) = \mu_{\text{new}} = (I - A + BK_\pi)^{-1}(Bb_\pi + \overline{A}\mu + d).$$

This gives the form of $\Lambda_2(\cdot, \cdot)$.

Next, we compute the Lipschitz constants for $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot, \cdot)$.

Lipschitz constant for $\Lambda_1(\cdot)$. By Proposition 3.2.4, for any $\mu_1, \mu_2 \in \mathbb{R}^m$, the optimal K^* is fixed for Problem 3.1.2. Therefore, by the form of the optimal b^K given in Proposition 3.2.4, it holds that

$$\begin{aligned}\big\|\Lambda_1(\mu_1) - \Lambda_1(\mu_2)\big\|_2 &\leq c_2 \cdot \Big\|\big[(I-A)Q^{-1}(I-A)^\top + BR^{-1}B^\top\big]^{-1}\overline{A}\Big\|_2 \\ &\qquad \cdot \Big\|\big[K^*Q^{-1}(I-A)^\top - R^{-1}B^\top\big]\Big\|_2 \cdot \|\mu_1 - \mu_2\|_2 \\ &= c_2 L_1 \cdot \|\mu_1 - \mu_2\|_2, \end{aligned} \tag{B.5.2}$$

where L_1 is defined in Assumption 3.2.1.

Lipschitz constants for $\Lambda_2(\cdot, \cdot)$. By Proposition 3.2.4, for any $\mu_1, \mu_2 \in \mathbb{R}^m$, the optimal K^* is fixed for Problem 3.1.2. Thus, for any $\pi \in \Pi$ such that π is an optimal policy under some $\mu \in \mathbb{R}^m$, it holds that

$$\begin{aligned}\big\|\Lambda_2(\mu_1, \pi) - \Lambda_2(\mu_2, \pi)\big\|_2 &= \big\|(I - A + BK_\pi)^{-1} \cdot \overline{A} \cdot (\mu_1 - \mu_2)\big\|_2 \\ &\le \big[1 - \rho(A - BK^*)\big]^{-1} \cdot \|\overline{A}\|_2 \cdot \|\mu_1 - \mu_2\|_2 \\ &= L_2 \cdot \|\mu_1 - \mu_2\|_2, \end{aligned} \tag{B.5.3}$$

where L_2 is defined in Assumption 3.2.1, and $K_\pi = K^*$ is the parameter of the policy π. Meanwhile, for any mean-field state $\mu \in \mathbb{R}^m$, and any poicies $\pi_1, \pi_2 \in \Pi$ that are optimal under some mean-field states μ_1, μ_2, respectively, we have

$$\begin{aligned}\big\|\Lambda_2(\mu, \pi_1) - \Lambda_2(\mu, \pi_2)\big\|_2 &= \big\|(I - A + BK^*)^{-1} B \cdot (b_{\pi_1} - b_{\pi_2})\big\|_2 \\ &\le \big[1 - \rho(A - BK^*)\big]^{-1} \cdot \|B\|_2 \cdot \|b_{\pi_1} - b_{\pi_2}\|_2 \\ &= c_2^{-1} L_3 \cdot \|\pi_1 - \pi_2\|_2, \end{aligned} \tag{B.5.4}$$

where in the last equality, we use the fact that $K_{\pi_1} = K_{\pi_2} = K^*$ by Proposition 3.2.4. Here L_3 is defined in Assumption 3.2.1, and b_{π_1} and b_{π_2} are the parameters of the policies π_1 and π_2.

Now we show that the operator $\Lambda(\cdot)$ is a contraction. For any $\mu_1, \mu_2 \in \mathbb{R}^m$, it holds that

$$\begin{aligned}
\big\|\Lambda(\mu_1) - \Lambda(\mu_2)\big\|_2 =& \Big\|\Lambda_2\big(\mu_1, \Lambda_1(\mu_1)\big) - \Lambda_2\big(\mu_2, \Lambda_1(\mu_2)\big)\Big\|_2 \\
\leq& \Big\|\Lambda_2\big(\mu_1, \Lambda_1(\mu_1)\big) - \Lambda_2\big(\mu_1, \Lambda_1(\mu_2)\big)\Big\|_2 + \Big\|\Lambda_2\big(\mu_1, \Lambda_1(\mu_2)\big) - \Lambda_2\big(\mu_2, \Lambda_1(\mu_2)\big)\Big\|_2 \\
\leq& \ c_2^{-1} L_3 \cdot \big\|\Lambda_1(\mu_1) - \Lambda_1(\mu_2)\big\|_2 + L_2 \cdot \|\mu_1 - \mu_2\|_2 \\
\leq& \ c_2^{-1} L_3 \cdot c_2 L_1 \cdot \|\mu_1 - \mu_2\|_2 + L_2 \cdot \|\mu_1 - \mu_2\|_2 = (L_1 L_3 + L_2) \cdot \|\mu_1 - \mu_2\|_2,
\end{aligned}$$

where the first inequality comes from triangle inequality, the second inequality comes from (B.5.3) and (B.5.4), and the last inequality comes from (B.5.2). By Assumption 3.2.1, we know that $L_0 = L_1 L_3 + L_2 < 1$, which shows that the operator $\Lambda(\cdot)$ is a contraction. Moreover, by Banach fixed-point theorem, we obtain that $\Lambda(\cdot)$ has a unique fixed point, which gives the unique equilibrium pair of Problem 3.1.3. We finish the proof of the proposition. □

B.5.2. Proof of Proposition 3.2.4

Proof. By the definition of $J_2(K, b)$ in (3.2.6) and the definition of $\mu_{K,b}$ in (3.2.2), the problem

$$\min_b J_2(K, b)$$

is equivalent to the following constrained problem,

$$\min_{\mu,b} \begin{pmatrix} \mu \\ b \end{pmatrix}^\top \begin{pmatrix} Q + K^\top RK & -K^\top R \\ -RK & R \end{pmatrix} \begin{pmatrix} \mu \\ b \end{pmatrix}$$

$$\text{s.t. } (I - A + BK)\mu - (Bb + \overline{A}\mu + d) = 0. \tag{B.5.5}$$

Following from the KKT conditions of (B.5.5), it holds that

$$2M_K \begin{pmatrix} \mu \\ b \end{pmatrix} + N_K\lambda = 0, \qquad N_K^\top \begin{pmatrix} \mu \\ b \end{pmatrix} + \overline{A}\mu + d = 0, \tag{B.5.6}$$

where

$$M_K = \begin{pmatrix} Q + K^\top RK & -K^\top R \\ -RK & R \end{pmatrix}, \qquad N_K = \begin{pmatrix} -(I - A + BK)^\top \\ B^\top \end{pmatrix}.$$

By solving (B.5.6), the minimizer of (B.5.5) takes the form of

$$\begin{pmatrix} \mu_{K,b^K} \\ b^K \end{pmatrix} = -M_K^{-1}N_K(N_K^\top M_K^{-1}N_K)^{-1}(\overline{A}\mu + d). \tag{B.5.7}$$

By substituting (B.5.7) into the definition of $J_2(K, b)$ in (3.2.6), we have

$$J_2(K, b^K) = (\overline{A}\mu + d)^\top(N_K^\top M_K^{-1}N_K)^{-1}(\overline{A}\mu + d). \tag{B.5.8}$$

Meanwhile, by calculation, we have

$$M_K^{-1} = \begin{pmatrix} Q^{-1} & Q^{-1}K^\top \\ KQ^{-1} & KQ^{-1}K^\top + R^{-1} \end{pmatrix}.$$

Therefore, the term $N_K^\top M_K^{-1} N_K$ in (B.5.8) takes the form of

$$N_K^\top M_K^{-1} N_K = (I - A)Q^{-1}(I - A^\top) + BR^{-1}B^\top. \tag{B.5.9}$$

By plugging (B.5.9) into (B.5.8), we have

$$J_2(K, b^K) = (\overline{A}\mu + d)^\top \big[(I - A)Q^{-1}(I - A^\top) + BR^{-1}B^\top\big]^{-1}(\overline{A}\mu + d).$$

Also, by plugging (B.5.9) into (B.5.7), we have

$$\begin{pmatrix} \mu_{K,b^K} \\ b^K \end{pmatrix} = \begin{pmatrix} Q^{-1}(I - A)^\top \\ KQ^{-1}(I - A)^\top - R^{-1}B^\top \end{pmatrix} \big[(I - A)Q^{-1}(I - A)^\top + BR^{-1}B^\top\big]^{-1}(\overline{A}\mu + d).$$

We finish the proof of the proposition. □

B.5.3. Proof of Proposition B.2.2

Proof. By the definition of the cost function $c(x, u)$ in Problem 3.1.2 (recall that we drop the subscript μ when we focus on Problem 3.1.2), we have

$$\begin{aligned} \mathbb{E}c_t &= \mathbb{E}(x_t^\top Q x_t + u_t^\top R u_t + \mu^\top \overline{Q}\mu) \\ &= \mathbb{E}(x_t^\top Q x_t + x_t^\top K^\top R K x_t - 2b^\top R K x_t + b^\top R b + \sigma^2 \eta_t^\top R \eta_t + \mu^\top \overline{Q}\mu) \\ &= \mathbb{E}\big[x_t^\top (Q + K^\top R K)x_t - 2b^\top R K x_t\big] + b^\top R b + \sigma^2 \cdot \operatorname{tr}(R) + \mu^\top \overline{Q}\mu, \qquad \text{(B.5.10)} \end{aligned}$$

where we write $c_t = c(x_t, u_t)$ for notational convenience. Here in the second line we use $u_t = \pi_{K,b}(x_t) = -Kx_t + b + \sigma\eta_t$. Therefore, combining (B.5.10) and the definition of $J(K, b)$ in Problem 3.1.2, we have

$$
\begin{aligned}
J(K,b) &= \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T}\Big\{\mathbb{E}\big[x_t^\top(Q+K^\top RK)x_t - 2b^\top RKx_t\big] + b^\top Rb + \sigma^2\cdot\mathrm{tr}(R) + \mu^\top\overline{Q}\mu\Big\} \\
&= \mathbb{E}_{x\sim\mathcal{N}(\mu_{K,b},\Phi_K)}\big[x^\top(Q+K^\top RK)x - 2b^\top RKx\big] + b^\top Rb + \sigma^2\cdot\mathrm{tr}(R) + \mu^\top\overline{Q}\mu
\end{aligned}
$$

(B.5.11)

$$
\begin{aligned}
&= \mathrm{tr}\big[(Q+K^\top RK)\Phi_K\big] + \mu_{K,b}^\top(Q+K^\top RK)\mu_{K,b} - 2b^\top RK\mu_{K,b} \\
&\qquad + b^\top Rb + \sigma^2\cdot\mathrm{tr}(R) + \mu^\top\overline{Q}\mu.
\end{aligned}
$$

Now, by iteratively applying (3.2.3) and (3.2.4), we have

$$
\mathrm{tr}\big[(Q+K^\top RK)\Phi_K\big] = \mathrm{tr}(P_K\Psi_\epsilon), \tag{B.5.12}
$$

where P_K is given in (3.2.4). Combining (B.5.11) and (B.5.12), we know that

$$
J(K,b) = J_1(K) + J_2(K,b) + \sigma^2\cdot\mathrm{tr}(R) + \mu^\top\overline{Q}\mu,
$$

where

$$
\begin{aligned}
J_1(K) &= \mathrm{tr}\big[(Q+K^\top RK)\Phi_K\big] = \mathrm{tr}(P_K\Psi_\epsilon), \\
J_2(K,b) &= \begin{pmatrix}\mu_{K,b}\\ b\end{pmatrix}^\top \begin{pmatrix}Q+K^\top RK & -K^\top R\\ -RK & R\end{pmatrix}\begin{pmatrix}\mu_{K,b}\\ b\end{pmatrix}.
\end{aligned}
$$

We finish the proof of the proposition. □

B.5.4. Proof of Proposition 3.2.3

Proof. By calculating the Hessian matrix of $J_2(K, b)$, we have

$$\begin{aligned}\nabla^2_{bb} J_2(K, b) =& B^\top (I - A + BK)^{-\top}(Q + K^\top RK)(I - A + BK)^{-1}B \\ & - \big[RK(I - A + BK)^{-1}B + B^\top (I - A + BK)^{-\top} K^\top R\big] + R \\ =& \big[R^{1/2}K(I - A + BK)^{-1}B - R^{1/2}\big]^\top \big[R^{1/2}K(I - A + BK)^{-1}B - R^{1/2}\big] \\ & + B^\top (I - A + BK)^{-\top} Q(I - A + BK)^{-1}B,\end{aligned}$$

which is a positive definite matrix independent of b. We denote by its minimum singular value as ν_K. Also, note that $\|\nabla^2_{bb} J_2(K, b)\|_2$ is upper bounded as

$$\big\|\nabla^2_{bb} J_2(K, b)\big\|_2 \leq \big[1 - \rho(A - BK)\big]^{-2} \cdot \big(\|B\|_2^2 \cdot \|K\|_2^2 \cdot \|R\|_2 + \|B\|_2^2 \cdot \|Q\|_2\big).$$

Therefore, it holds that

$$\iota_K \leq \big[1 - \rho(A - BK)\big]^{-2} \cdot \big(\|B\|_2^2 \cdot \|K\|_2^2 \cdot \|R\|_2 + \|B\|_2^2 \cdot \|Q\|_2\big),$$

where ι_K is the maximum singular value of $\nabla^2_{bb} J_2(K, b)$. We finish the proof of the proposition. □

B.5.5. Proof of Proposition B.2.3

Proof. Following from Proposition B.2.2, it holds that

$$J_1(K) = \mathrm{tr}(P_K \Psi_\epsilon) = \mathbb{E}_{y\sim\mathcal{N}(0,\Psi_\epsilon)}(y^\top P_K y) = \mathbb{E}_{y\sim\mathcal{N}(0,\Psi_\epsilon)}\big[f_K(y)\big], \qquad \text{(B.5.13)}$$

where $f_K(y) = y^\top P_K y$. By the definition of P_K in (3.2.4), we obtain that

$$\begin{aligned}\nabla_K f_K(y) &= \nabla_K \Big\{ y^\top (Q + K^\top R K) y + \big[(A - BK)y\big]^\top P_K \big[(A - BK)y\big]^\top \Big\} \\ &= 2RKyy^\top + \nabla_K \Big[f_K\big((A - BK)y\big)\Big]. \qquad \text{(B.5.14)}\end{aligned}$$

Also, we have

$$\nabla_K \Big[f_K\big((A - BK)y\big)\Big] = \nabla_K f_K\big((A - BK)y\big) - 2B^\top P_K (A - BK) yy^\top. \qquad \text{(B.5.15)}$$

By plugging (B.5.15) into (B.5.14), we have

$$\nabla_K f_K(y) = 2\big[(R + B^\top P_K B)K - B^\top P_K A\big] yy^\top + \nabla_K f_K\big((A - BK)y\big). \qquad \text{(B.5.16)}$$

By iteratively applying (B.5.16), it holds that

$$\nabla_K f_K(y) = 2\big[(R + B^\top P_K B)K - B^\top P_K A\big] \cdot \sum_{t=0}^{\infty} y_t y_t^\top, \qquad \text{(B.5.17)}$$

where $y_{t+1} = (A - BK)y_t$ with $y_0 = y$. Now, combining (B.5.13) and (B.5.17), it holds that

$$\nabla_K J_1(K) = 2\big[(R + B^\top P_K B)K - B^\top P_K A\big]\Phi_K = 2(\Upsilon_K^{22} K - \Upsilon_K^{21}) \cdot \Phi_K,$$

where Υ_K is defined in (3.2.7). Meanwhile, combining the form of $\mu_{K,b}$ in (3.2.2), it holds by calculation that

$$\nabla_b J_2(K, b) = 2\big[\Upsilon_K^{22}(-K\mu_{K,b} + b) + \Upsilon_K^{21}\mu_{K,b} + q_{K,b}\big],$$

where $q_{K,b}$ is defined in (3.2.7). We finish the proof of the proposition. □

B.5.6. Proof of Proposition B.2.1

Proof. From the definition of $V_{K,b}(x)$ in (B.2.1) and the definition of the cost function $c(x,u)$ in Problem 3.1.2, it holds that

$$V_{K,b}(x) = \sum_{t=0}^{\infty}\Big\{\mathbb{E}\big[x_t^\top (Q + K^\top RK)x_t - 2b^\top RKx_t \\ + b^\top Rb + \sigma^2\eta_t^\top R\eta_t + \mu^\top\overline{Q}\mu \,|\, x_0 = x\big] - J(K,b)\Big\}.$$

Combining (3.2.1), we know that $V_{K,b}(x)$ is a quadratic function taking the form of $V_{K,b}(x) = x^\top Gx + r^\top x + h$, where G, r, and h are functions of K and b. Note that $V_{K,b}(x)$ satisfies that

$$V_{K,b}(x) = c(x, -Kx+b) - J(K,b) + \mathbb{E}\big[V_{K,b}(x') \,|\, x\big], \tag{B.5.18}$$

by substituting the form of $c(x, -Kx+b)$ in Problem 3.1.2 and $J(K,b)$ in (3.2.5) into (B.5.18), we obtain that

$$x^\top Gx + r^\top x + h$$

(B.5.19)

$$\begin{aligned} &= x^\top (Q + K^\top RK)x - 2b^\top RKx + b^\top Rb + \mu^\top \overline{Q}\mu \\ &\quad - \big[\mathrm{tr}(P_K\Psi_\epsilon) + \mu_{K,b}^\top (Q + K^\top RK)\mu_{K,b} - 2b^\top RK\mu_{K,b} + \mu^\top \overline{Q}\mu + b^\top Rb\big] \\ &\quad + \big[(A - BK)x + (Bb + \overline{A}\mu + d)\big]^\top G\big[(A - BK)x + (Bb + \overline{A}\mu + d)\big] \\ &\quad + \mathrm{tr}(G\Psi_\epsilon) + r^\top\big[(A - BK)x + (Bb + \overline{A}\mu + d)\big] + h - \sigma^2 \cdot \mathrm{tr}(R). \end{aligned}$$

By comparing the quadratic terms and linear terms on both the LHS and RHS in (B.5.19), we obtain that

$$G = P_K, \qquad r = 2f_{K,b},$$

where $f_{K,b} = (I - A + BK)^{-\top}[(A - BK)^\top P_K(Bb + \overline{A}\mu + d) - K^\top Rb]$. Also, by the definition of $V_{K,b}(x)$ in (B.2.1), we know that $\mathbb{E}[V_{K,b}(x)] = 0$, where the expectation is taken following the stationary distribution generated by the policy $\pi_{K,b}$ and the state transition. Therefore, we have

$$h = -2f_{K,b}\mu_{K,b} - \mu_{K,b}^\top P_K\mu_{K,b} - \mathrm{tr}(P_K\Phi_K),$$

which shows that

$$V_{K,b}(x) = x^\top P_K x - \operatorname{tr}(P_K \Phi_K) + 2 f_{K,b}^\top (x - \mu_{K,b}) - \mu_{K,b}^\top P_K \mu_{K,b}. \tag{B.5.20}$$

For the action-value function $Q_{K,b}(x,u)$, by plugging (B.5.20) into (B.2.2), we obtain that

$$\begin{aligned} Q_{K,b}(x,u) = & \begin{pmatrix} x \\ u \end{pmatrix}^\top \Upsilon_K \begin{pmatrix} x \\ u \end{pmatrix} + 2 \begin{pmatrix} p_{K,b} \\ q_{K,b} \end{pmatrix}^\top \begin{pmatrix} x \\ u \end{pmatrix} - \operatorname{tr}(P_K \Phi_K) - \sigma^2 \cdot \operatorname{tr}(R + P_K B B^\top) \\ & - b^\top R b + 2 b^\top R K \mu_{K,b} - \mu_{K,b}^\top (Q + K^\top R K + P_K) \mu_{K,b} \\ & + 2 f_{K,b}^\top \big[(\overline{A}\mu + d) - \mu_{K,b} \big] + (\overline{A}\mu + d)^\top P_K (\overline{A}\mu + d). \end{aligned}$$

We finish the proof of the proposition. □

B.5.7. Proof of Proposition B.2.5

Proof. By Proposition B.2.1, it holds that $Q_{K,b}$ takes the following linear form

$$Q_{K,b}(x,u) = \psi(x,u)^\top \alpha_{K,b} + \beta_{K,b}, \tag{B.5.21}$$

where $\beta_{K,b}$ is a scalar independent of x and u. Note that $Q_{K,b}(x,u)$ satisfies that

$$Q_{K,b}(x,u) = c(x,u) - J(K,b) + \mathbb{E}_{\pi_{K,b}} \big[Q_{K,b}(x', u') \,|\, x, u \big], \tag{B.5.22}$$

where (x', u') is the state-action pair after (x, u) following the policy $\pi_{K,b}$ and the state transition. Combining (B.5.21) and (B.5.22), we obtain that

$$\psi(x,u)^\top \alpha_{K,b} = c(x,u) - J(K,b) + \mathbb{E}_{\pi_{K,b}}\big[\psi(x',u') \,|\, x,u\big]^\top \alpha_{K,b}. \tag{B.5.23}$$

By left multiplying $\psi(x, u)$ to both sides of (B.5.23), and taking the expectation, we have

$$\mathbb{E}_{\pi_{K,b}}\Big\{\psi(x,u)\big[\psi(x,u) - \psi(x',u')\big]^\top\Big\} \cdot \alpha_{K,b} + \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] \cdot J(K,b) = \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big].$$

Combining the definition of the matrix $\Theta_{K,b}$ in (B.2.7), we have

$$\begin{pmatrix} 1 & 0 \\ \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] & \Theta_{K,b} \end{pmatrix} \begin{pmatrix} J(K,b) \\ \alpha_{K,b} \end{pmatrix} = \begin{pmatrix} J(K,b) \\ \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big] \end{pmatrix},$$

which concludes the proof of the proposition. □

B.5.8. Proof of Proposition B.2.6

Proof. Invertibility and Upper Bound. We denote by $z_t = (x_t^\top, u_t^\top)^\top$ for any $t \geq 0$. Then following from the state transition and the policy $\pi_{K,b}$, the transition of $\{z_t\}_{t\geq 0}$ takes the form of

$$z_{t+1} = Lz_t + \nu + \delta_t, \tag{B.5.24}$$

where L, ν and δ are defined as

$$L = \begin{pmatrix} A & B \\ -KA & -KB \end{pmatrix}, \qquad \nu = \begin{pmatrix} \overline{A}\mu + d \\ -K(\overline{A}\mu + d) + b \end{pmatrix}, \qquad \delta_t = \begin{pmatrix} \omega_t \\ -K\omega_t + \sigma\eta_t \end{pmatrix}.$$

Note that L also takes the form of

$$L = \begin{pmatrix} I \\ -K \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix}.$$

Combining the fact that $\rho(UV) = \rho(VU)$ for any matrices U and V, we know that $\rho(L) = \rho(A - BK) < 1$, which verifies the stability of (B.5.24). Following from the stability of (B.5.24), we know that the Markov chain generated by (B.5.24) admits a unique stationary distribution $\mathcal{N}(\mu_z, \Sigma_z)$, where μ_z and Σ_z satisfy that

$$\mu_z = L\mu_z + \nu, \qquad \Sigma_z = L\Sigma_z L^\top + \Psi_\delta.$$

where

$$\Psi_\delta = \begin{pmatrix} \Psi_\omega & -\Psi_\omega K^\top \\ -K\Psi_\omega & K\Psi_\omega K^\top + \sigma^2 I \end{pmatrix}.$$

Also, we know that Σ_z takes the form of

(B.5.25)

$$\Sigma_z = \operatorname{Cov}\left[\begin{pmatrix} x \\ u \end{pmatrix}\right] = \begin{pmatrix} \Phi_K & -\Phi_K K^\top \\ -K\Phi_K & K\Phi_K K^\top + \sigma^2 I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 I \end{pmatrix} + \begin{pmatrix} I \\ -K \end{pmatrix} \Phi_K \begin{pmatrix} I \\ -K \end{pmatrix}^\top,$$

where Φ_K is defined in (3.2.3).

The following lemma establishes the form of $\Theta_{K,b}$.

Lemma B.5.2. The matrix $\Theta_{K,b}$ in (B.2.7) takes the form of

$$\Theta_{K,b} = \begin{pmatrix} 2(\Sigma_z \otimes_s \Sigma_z)(I - L \otimes_s L)^\top & 0 \\ 0 & \Sigma_z(I - L)^\top \end{pmatrix}.$$

Proof. See §B.6.9 for a detailed proof. □

Note that since $\rho(L) < 1$, both $I - L \otimes_s L$ and $I - L$ are positive definite. Therefore, by Lemma B.5.2, the matrix $\Theta_{K,b}$ is invertible. This finishes the proof of the invertibility of $\Theta_{K,b}$. Moreover, by (B.5.25) and Lemma B.5.2, we upper bound the spectral norm of $\Theta_{K,b}$ as

$$\|\Theta_{K,b}\|_2 \le 2\max\Big\{\|\Sigma_z\|_2^2 \cdot \big(1 + \|L\|_2^2\big),\ \|\Sigma_z\|_2 \cdot \big(1 + \|L\|_2\big)\Big\} \le 4\big(1 + \|K\|_{\mathrm{F}}^2\big)^2 \cdot \|\Phi_K\|_2^2,$$

which proves the upper bound of $\|\Theta_{K,b}\|_2$.

Minimum singular value. To lower bound $\sigma_{\min}(\widetilde{\Theta}_{K,b})$, we only need to upper bound $\sigma_{\max}(\widetilde{\Theta}_{K,b}^{-1}) = \|\widetilde{\Theta}_{K,b}^{-1}\|_2$. We first calculate $\widetilde{\Theta}_{K,b}^{-1}$. Recall that the matrix $\widetilde{\Theta}_{K,b}$ in (B.2.8) takes the form of

$$\widetilde{\Theta}_{K,b} = \begin{pmatrix} 1 & 0 \\ \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] & \Theta_{K,b} \end{pmatrix}.$$

By the definition of the feature vector $\psi(x,u)$ in (B.2.5), the vector $\widetilde{\sigma}_z = \mathbb{E}_{\pi_{K,b}}[\psi(x,u)]$ takes the form of

$$\widetilde{\sigma}_z = \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big] = \begin{pmatrix} \mathrm{svec}(\Sigma_z) \\ \mathbf{0}_{k+m} \end{pmatrix},$$

where $\mathbf{0}_{k+m}$ denotes the all-zero column vector with dimension $k+m$. Also, since $\Theta_{K,b}$ is invertible, the matrix $\widetilde{\Theta}_{K,b}$ is also invertible, whose inverse takes the form of

$$\widetilde{\Theta}_{K,b}^{-1} = \begin{pmatrix} 1 & 0 \\ -\Theta_{K,b}^{-1} \cdot \widetilde{\sigma}_z & \Theta_{K,b}^{-1} \end{pmatrix}.$$

The following lemma upper bounds the spectral norm of $\widetilde{\Theta}_{K,b}^{-1}$.

Lemma B.5.3. The spectral norm of the matrix $\widetilde{\Theta}_{K,b}^{-1}$ is upper bounded by a positive constant $\widetilde{\lambda}_K$, where $\widetilde{\lambda}_K$ only depends on $\|K\|_2$ and $\rho(A-BK)$.

Proof. See §B.6.10 for a detailed proof. □

By Lemma B.5.3, we know that $\sigma_{\min}(\widetilde{\Theta}_{K,b})$ is lower bounded by a positive constant $\lambda_K = 1/\widetilde{\lambda}_K$, which only depends on $\|K\|_2$ and $\rho(A-BK)$. This concludes the proof of the proposition. □

B.6. Proofs of Lemmas

B.6.1. Proof of Lemma B.4.1

Proof. Following from (3.2.4), it holds that

$$y^\top P_{K_2} y = \sum_{t\geq 0} y^\top \bigl[(A-BK_2)^t\bigr]^\top (Q + K_2^\top R K_2)(A-BK_2)^t y. \tag{B.6.1}$$

Meanwhile, by the state transition $y_{t+1} = (A-BK_2)y_t$, we know that

$$y_t = (A-BK_2)^t y_0 = (A-BK_2)^t y. \tag{B.6.2}$$

By plugging (B.6.2) into (B.6.1), it holds that

$$y^\top P_{K_2} y = \sum_{t\geq 0} y_t^\top (Q + K_2^\top R K_2) y_t = \sum_{t \geq 0} (y_t^\top Q y_t + y_t^\top K_2^\top R K_2 y_t). \tag{B.6.3}$$

Also, it holds that

$$y^\top P_{K_1} y = \sum_{t\geq 0} (y_{t+1}^\top P_{K_1} y_{t+1} - y_t^\top P_{K_1} y_t) \tag{B.6.4}$$

Combining (B.6.3) and (B.6.4), we have

$$y^\top P_{K_2} y - y^\top P_{K_1} y = \sum_{t\geq 0} (y_t^\top Q y_t + y_t^\top K_2^\top R K_2 y_t + y_{t+1}^\top P_{K_1} y_{t+1} - y_t^\top P_{K_1} y_t). \tag{B.6.5}$$

Also, by the state transition $y_{t+1} = (A - BK_2) y_t$, it holds for any $t \geq 0$ that

$$\begin{aligned}
& y_t^\top Q y_t + y_t^\top K_2^\top R K_2 y_t + y_{t+1}^\top P_{K_1} y_{t+1} - y_t^\top P_{K_1} y_t \\
&\quad = y_t^\top \big[Q + (K_2 - K_1 + K_1)^\top R (K_2 - K_1 + K_1) \big] y_t \\
&\qquad + y_t^\top \big[A - BK_1 - B(K_2 - K_1) \big]^\top P_{K_1} \big[A - BK_1 - B(K_2 - K_1) \big] y_t - y_t^\top P_{K_1} y_t \\
&\quad = 2 y_t^\top (K_2 - K_1)^\top \big[(R + B^\top P_{K_1} B) K_1 - B^\top P_{K_1} A \big] y_t \\
&\qquad + y_t^\top (K_2 - K_1)^\top (R + B^\top P_{K_1} B)(K_2 - K_1) y_t
\end{aligned}$$

(B.6.6)

$$= 2 y_t^\top (K_2 - K_1)^\top (\Upsilon_{K_1}^{22} K_1 - \Upsilon_{K_1}^{21}) y_t + y_t^\top (K_2 - K_1)^\top \Upsilon_{K_1}^{22} (K_2 - K_1) y_t,$$

where the matrix Υ_{K_1} is defined in (3.2.7). By plugging (B.6.6) into (B.6.5), we have

$$\begin{aligned} & y^\top P_{K_2} y - y^\top P_{K_1} y \\ & \quad = \sum_{t\geq 0} 2y_t^\top (K_2 - K_1)^\top (\Upsilon_{K_1}^{22} K_1 - \Upsilon_{K_1}^{21}) y_t + y_t^\top (K_2 - K_1)^\top \Upsilon_{K_1}^{22} (K_2 - K_1) y_t \\ & \quad = \sum_{t\geq 0} D_{K_1,K_2}(y_t), \end{aligned}$$

where $D_{K_1,K_2}(y) = 2y^\top (K_2 - K_1)(\Upsilon_{K_1}^{22} K_1 - \Upsilon_{K_1}^{21}) y + y^\top (K_2 - K_1)^\top \Upsilon_{K_1}^{22} (K_2 - K_1) y$. We finish the proof of the lemma. □

B.6.2. Proof of Lemma B.4.2

Proof. We prove (B.4.16) and (B.4.17) separately in the sequel.

Proof of (B.4.16)**.** From the definition of $J_1(K)$ in (3.2.6), we have

$$\begin{aligned} J_1(K) - J_1(K^*) &= \mathrm{tr}(P_K \Psi_\epsilon - P_{K^*} \Psi_\epsilon) = \mathbb{E}_{y\sim\mathcal{N}(0,\Psi_\epsilon)}(y^\top P_K y - y^\top P_{K^*} y) \\ &= -\mathbb{E}\biggl[\sum_{t\geq 0} D_{K,K^*}(y_t)\biggr], \end{aligned} \tag{B.6.7}$$

where in the last equality, we apply Lemma B.4.1 and the expectation is taken following the transition $y_{t+1} = (A - BK^*) y_t$ with initial state $y_0 \sim \mathcal{N}(0, \Psi_\epsilon)$. Here we denote by $D_{K,K^*}(y)$ as

$$D_{K,K^*}(y) = 2y^\top (K^* - K)(\Upsilon_K^{22} K - \Upsilon_K^{21}) y + y^\top (K^* - K)^\top \Upsilon_K^{22} (K^* - K) y.$$

Also, we write $D_{K,K^*}(y)$ as

(B.6.8)

$$\begin{aligned} D_{K,K^*}(y) &= 2y^\top (K^* - K)(\Upsilon_K^{22} K - \Upsilon_K^{21}) y + y^\top (K^* - K)^\top \Upsilon_K^{22} (K^* - K) y \\ &= y^\top \big[K^* - K + (\Upsilon_K^{22})^{-1}(\Upsilon_K^{22} K - \Upsilon_K^{21})\big]^\top \Upsilon_K^{22} \big[K^* - K + (\Upsilon_K^{22})^{-1}(\Upsilon_K^{22} K - \Upsilon_K^{21})\big] y \\ &\qquad - y^\top (\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22})^{-1} (\Upsilon_K^{22} K - \Upsilon_K^{21}) y. \end{aligned}$$

Note that the first term on the RHS of (B.6.8) is positive, due to the fact that it is a quadratic form of a positive definite matrix, we lower bound $D_{K,K^*}(y)$ as

$$D_{K,K^*}(y) \geq -y^\top (\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22})^{-1} (\Upsilon_K^{22} K - \Upsilon_K^{21}) y. \tag{B.6.9}$$

Combining (B.6.7) and (B.6.9), it holds that

$$\begin{aligned} J_1(K) - J_1(K^*) &\leq \bigg\| \mathbb{E}\bigg(\sum_{t \geq 0} y_t y_t^\top \bigg) \bigg\|_2 \cdot \mathrm{tr}\big[(\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22})^{-1} (\Upsilon_K^{22} K - \Upsilon_K^{21})\big] \\ &= \|\Phi_{K^*}\|_2 \cdot \mathrm{tr}\big[(\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22})^{-1} (\Upsilon_K^{22} K - \Upsilon_K^{21})\big] \\ &\leq \big\|(\Upsilon_K^{22})^{-1}\big\|_2 \cdot \|\Phi_{K^*}\|_2 \cdot \mathrm{tr}\big[(\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22} K - \Upsilon_K^{21})\big] \\ &\leq \sigma_{\min}^{-1}(R) \cdot \|\Phi_{K^*}\|_2 \cdot \mathrm{tr}\big[(\Upsilon_K^{22} K - \Upsilon_K^{21})^\top (\Upsilon_K^{22} K - \Upsilon_K^{21})\big], \end{aligned}$$

where the last line comes from the fact that $\Upsilon_K^{22} = R + B^\top P_K B \succeq R$. This complete the proof of (B.4.16).

Proof of (B.4.17)**.** Note that for any $\widetilde{K}$, it holds by the optimality of K^* that

$$J_1(K) - J_1(K^*) \geq J_1(K) - J_1(\widetilde{K}) = -\mathbb{E}\biggl[\sum_{t\geq 0} D_{K,\widetilde{K}}(y_t)\biggr], \tag{B.6.10}$$

where the expectation is taken following the transition $y_{t+1} = (A - B\widetilde{K})y_t$ with initial state $y_0 \sim \mathcal{N}(0, \Psi_\epsilon)$. By taking $\widetilde{K} = K - (\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})$ and following from a similar calculation as in (B.6.8), the function $D_{K,\widetilde{K}}(y)$ takes the form of

$$D_{K,\widetilde{K}}(y) = -y^\top(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})y. \tag{B.6.11}$$

Combining (B.6.10) and (B.6.11), it holds that

$$\begin{aligned} J(K) - J(K^*) &\geq \operatorname{tr}\bigl[\Phi_{\widetilde{K}}(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22})^{-1}(\Upsilon_K^{22}K - \Upsilon_K^{21})\bigr] \\ &\geq \sigma_{\min}(\Psi_\epsilon)\cdot\|\Upsilon_K^{22}\|_2^{-1}\cdot\operatorname{tr}\bigl[(\Upsilon_K^{22}K - \Upsilon_K^{21})^\top(\Upsilon_K^{22}K - \Upsilon_K^{21})\bigr], \end{aligned}$$

where we use the fact that $\Phi_{\widetilde{K}} = (A - B\widetilde{K})\Phi_{\widetilde{K}}(A - B\widetilde{K})^\top + \Psi_\epsilon \succeq \Psi_\epsilon$ in the last line. This finishes the proof of (B.4.17). □

B.6.3. Proof of Lemma B.4.3

Proof. By Proposition B.2.2, we have

$$\bigl|J_1(\widetilde{K}_{n+1}) - J_1(K_{n+1})\bigr| = \operatorname{tr}\bigl[(P_{\widetilde{K}_{n+1}} - P_{K_{n+1}})\Psi_\epsilon\bigr] \leq \|P_{\widetilde{K}_{n+1}} - P_{K_{n+1}}\|_2\cdot\|\Psi_\epsilon\|_{\mathrm{F}}. \tag{B.6.12}$$

The following lemma upper bounds the term $\|P_{\widetilde{K}_{n+1}} - P_{K_{n+1}}\|_2$.

Lemma B.6.1. Suppose that the parameters K and $\widetilde{K}$ satisfy that

$$\|\widetilde{K} - K\|_2 \cdot \big(\|A - BK\|_2 + 1\big) \cdot \|\Phi_K\|_2 \leq \sigma_{\min}(\Psi_\omega)/4 \cdot \|B\|_2^{-1}, \tag{B.6.13}$$

then it holds that

$$\begin{aligned} \|P_{\widetilde{K}} - P_K\|_2 \leq 6 \cdot \sigma_{\min}^{-1}(\Psi_\omega) \cdot \|\Phi_K\|_2 \cdot \|K\|_2 \cdot \|R\|_2 \cdot \|\widetilde{K} - K\|_2 \\ \cdot \big(\|B\|_2 \cdot \|K\|_2) \cdot \|A - BK\|_2 + \|B\|_2 \cdot \|K\|_2 + 1\big). \end{aligned} \tag{B.6.14}$$

Proof. See Lemma 5.7 in Yang et al. (2019b) for a detailed proof. □

To use Lemma B.6.1, it suffices to verify that $\widetilde{K}_{n+1}$ and K_{n+1} satisfy (B.6.13). Note that from the definitions of K_{n+1} and $\widetilde{K}_{n+1}$ in (B.4.18) and (B.4.21), respectively, we have

$$\begin{aligned} &\|\widetilde{K}_{n+1} - K_{n+1}\|_2 \cdot \big(\|A - B\widetilde{K}_{n+1}\|_2 + 1\big) \cdot \|\Phi_{\widetilde{K}_{n+1}}\|_2 \\ &\quad \leq \gamma \cdot \|\widehat{\Upsilon}_{K_n} - \Upsilon_{K_n}\|_{\mathrm{F}} \cdot \big(1 + \|K_n\|_2\big) \cdot \big(\|A - B\widetilde{K}_{n+1}\|_2 + 1\big) \cdot \|\Phi_{\widetilde{K}_{n+1}}\|_2. \end{aligned} \tag{B.6.15}$$

Now, we upper bound the RHS of (B.6.15). For the term $\|A - B\widetilde{K}_{n+1}\|_2$, it holds by the definition of $\widetilde{K}_{n+1}$ in (B.4.21) that

$$\begin{aligned} \|A - B\widetilde{K}_{n+1}\|_2 &\leq \|A - BK_n\|_2 + \gamma \cdot \|B\|_2 \cdot \|\Upsilon_{K_n}^{22} K_n - \Upsilon_{K_n}^{21}\|_2 \\ &\leq \|A - BK_n\|_2 + \gamma \cdot \|B\|_2 \cdot \|\Upsilon_{K_n}\|_2 \cdot \big(1 + \|K_n\|_2\big). \end{aligned} \tag{B.6.16}$$

By the definition of Υ_{K_n} in (3.2.7), we upper bound $\|\Upsilon_{K_n}\|_2$ as

$$\begin{aligned}\|\Upsilon_{K_n}\|_2 &\leq \|Q\|_2 + \|R\|_2 + \big(\|A\|_{\mathrm{F}} + \|B\|_{\mathrm{F}}\big)^2 \cdot \|P_{K_n}\|_2 \\ &\leq \|Q\|_2 + \|R\|_2 + \big(\|A\|_{\mathrm{F}} + \|B\|_{\mathrm{F}}\big)^2 \cdot J_1(K_0) \cdot \sigma_{\min}^{-1}(\Psi_\epsilon), \qquad \text{(B.6.17)}\end{aligned}$$

where the last line comes from the fact that

$$J_1(K_0) \geq J_1(K_n) = \mathrm{tr}\big[(Q + K_n^\top R K_n)\Phi_{K_n}\big] = \mathrm{tr}(P_{K_n}\Psi_\epsilon) \geq \|P_{K_n}\|_2 \cdot \sigma_{\min}(\Psi_\epsilon).$$

As for the term $\|\Phi_{\widetilde{K}_{n+1}}\|_2$ in (B.6.15), from the fact that

$$J_1(K_0) \geq J_1(\widetilde{K}_{n+1}) = \mathrm{tr}\big[(Q + \widetilde{K}_{n+1}^\top R \widetilde{K}_{n+1})\Phi_{\widetilde{K}_{n+1}}\big] \geq \|\Phi_{\widetilde{K}_{n+1}}\|_2 \cdot \sigma_{\min}(Q),$$

it holds that

$$\|\Phi_{\widetilde{K}_{n+1}}\|_2 \leq J_1(K_0) \cdot \sigma_{\min}^{-1}(Q). \qquad \text{(B.6.18)}$$

Therefore, combining (B.6.15), (B.6.16), (B.6.17), and (B.6.18), we know that

$$\begin{aligned}&\|\widetilde{K}_{n+1} - K_{n+1}\|_2 \cdot \big(\|A - B\widetilde{K}_{n+1}\|_2 + 1\big) \cdot \|\Phi_{\widetilde{K}_{n+1}}\|_2 \\ &\quad \leq \mathrm{poly}_1\big(\|K_n\|_2\big) \cdot \|\widehat{\Upsilon}_{K_n} - \Upsilon_{K_n}\|_{\mathrm{F}}. \qquad \text{(B.6.19)}\end{aligned}$$

From Theorem B.2.8, it holds with probability at least $1 - T_n^{-4} - \widetilde{T}_n^{-6}$ that

$$\begin{aligned}\|\widehat{\Upsilon}_{K_n} - \Upsilon_{K_n}\|_{\mathrm{F}} &\leq \frac{\mathrm{poly}_3\big(\|K_n\|_{\mathrm{F}}, \|\mu\|_2\big)}{\lambda_{K_n} \cdot (1-\rho)^2} \cdot \frac{\log^3 T_n}{T_n^{1/4}} \qquad \text{(B.6.20)} \\ &\quad + \frac{\mathrm{poly}_4\big(\|K_n\|_{\mathrm{F}}, \|b_0\|_2, \|\mu\|_2\big)}{\lambda_{K_n}} \cdot \frac{\log^{1/2} \widetilde{T}_n}{\widetilde{T}_n^{1/8} \cdot (1-\rho)},\end{aligned}$$

which holds for any $\rho \in (\rho(A - BK_n), 1)$. Note that from the choice of T_n and $\widetilde{T}_n$ in the statement of Theorem B.2.4 that

$$T_n \geq \text{poly}_5\big(\|K_n\|_{\text{F}}, \|b_0\|_2, \|\mu\|_2\big) \cdot \lambda_{K_n}^{-4} \cdot \big[1 - \rho(A - BK_n)\big]^{-9} \cdot \varepsilon^{-5},$$

$$\widetilde{T}_n \geq \text{poly}_6\big(\|K_n\|_{\text{F}}, \|b_0\|_2, \|\mu\|_2\big) \cdot \lambda_{K_n}^{-2} \cdot \big[1 - \rho(A - BK_n)\big]^{-12} \cdot \varepsilon^{-12},$$

it holds that

$$\frac{\text{poly}_3\big(\|K_n\|_{\text{F}}, \|\mu\|_2\big)}{\lambda_{K_n} \cdot (1-\rho)^2} \cdot \frac{\log^3 T_n}{T_n^{1/4}} + \frac{\text{poly}_4\big(\|K_n\|_{\text{F}}, \|b_0\|_2, \|\mu\|_2\big)}{\lambda_{K_n}} \cdot \frac{\log^{1/2} \widetilde{T}_n}{\widetilde{T}_n^{1/8} \cdot (1-\rho)}$$

(B.6.21)

$$\leq \min\bigg\{ \Big[\text{poly}_1\big(\|K_n\|_2\big)\Big]^{-1} \cdot \sigma_{\min}(\Psi_\omega)/4 \cdot \|B\|_2^{-1}, \\ \Big[\text{poly}_2\big(\|K_n\|_2\big)\Big]^{-1} \cdot \varepsilon/8 \cdot \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_2^{-1} \cdot \|\Psi_\epsilon\|_{\text{F}}^{-1} \bigg\}.$$

Combining (B.6.19), (B.6.20), and (B.6.21), we know that (B.6.13) holds with probability at least $1-\varepsilon^{15}$ for sufficiently small $\varepsilon > 0$. Meanwhile, by (B.6.16), (B.6.17), and (B.6.18), the RHS of (B.6.14) is upper bounded as

$$6 \cdot \sigma_{\min}^{-1}(\Psi_\omega) \cdot \|\Phi_{\widetilde{K}_{n+1}}\|_2 \cdot \|\widetilde{K}_{n+1}\|_2 \cdot \|R\|_2 \cdot \|\widetilde{K}_{n+1} - K_{n+1}\|_2 \\ \cdot \big(\|B\|_2 \cdot \|\widetilde{K}_{n+1}\|_2\big) \cdot \|A - B\widetilde{K}_{n+1}\|_2 + \|B\|_2 \cdot \|\widetilde{K}_{n+1}\|_2 + 1\big)$$

(B.6.22)

$$\leq \text{poly}_2\big(\|K_n\|_2\big) \cdot \|\widehat{\Upsilon}_{K_n} - \Upsilon_{K_n}\|_{\text{F}}.$$

Now, by Lemma B.6.1, it holds with probability at least $1 - \varepsilon^{15}$ that

$$\begin{aligned}\|P_{\widetilde{K}_{n+1}} - P_{K_{n+1}}\|_2 &\leq 6 \cdot \sigma_{\min}^{-1}(\Psi_\omega) \cdot \|\Phi_{\widetilde{K}_{n+1}}\|_2 \cdot \|\widetilde{K}_{n+1}\|_2 \cdot \|R\|_2 \cdot \|\widetilde{K}_{n+1} - K_{n+1}\|_2 \\ &\qquad \cdot \big(\|B\|_2 \cdot \|\widetilde{K}_{n+1}\|_2\big) \cdot \|A - B\widetilde{K}_{n+1}\|_2 + \|B\|_2 \cdot \|\widetilde{K}_{n+1}\|_2 + 1\big) \\ &\leq \text{poly}_2\big(\|K_n\|_2\big) \cdot \|\widehat{\Upsilon}_{K_n} - \Upsilon_{K_n}\|_{\text{F}} \\ &\leq \varepsilon/8 \cdot \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_2^{-1} \cdot \|\Psi_\epsilon\|_{\text{F}}^{-1}, \end{aligned} \tag{B.6.23}$$

where the second inequality comes from (B.6.22), and the last inequality comes from (B.6.20) and (B.6.21). Combining (B.6.12) and (B.6.23), it holds with probability at least $1 - \varepsilon^{15}$ that

$$\big|J_1(\widetilde{K}_{n+1}) - J_1(K_{n+1})\big| \leq \gamma \cdot \sigma_{\min}(\Psi_\epsilon) \cdot \sigma_{\min}(R) \cdot \|\Phi_{K^*}\|_2^{-1} \cdot \varepsilon/4,$$

which concludes the proof of the lemma. □

B.6.4. Proof of Lemma B.4.4

Proof. Note that $\Upsilon_{K^*}^{22} K^* - \Upsilon_{K^*}^{21}$ is the natural gradient of J_1 at the minimizer K^*, which implies that

$$\Upsilon_{K^*}^{22} K^* - \Upsilon_{K^*}^{21} = 0. \tag{B.6.24}$$

By Lemma B.4.1, it holds that

$$\begin{aligned}
J_1(K) - J_1(K^*) &= \operatorname{tr}(P_K \Psi_\epsilon - P_{K^*}\Psi_\epsilon) = \mathbb{E}_{y\sim\mathcal{N}(0,\Psi_\epsilon)}(y^\top P_K y - y^\top P_{K^*} y) \\
&= \mathbb{E}\biggl\{\sum_{t\geq 0}\Bigl[2y_t^\top (K-K^*)(\Upsilon^{22}_{K^*}K^* - \Upsilon^{21}_{K^*})y_t + y_t^\top (K-K^*)^\top \Upsilon^{22}_{K^*}(K-K^*)y_t\Bigr]\biggr\}
\end{aligned}$$

(B.6.25)
$$= \mathbb{E}\biggl\{\sum_{t\geq 0} y_t^\top (K-K^*)^\top \Upsilon^{21}_{K^*}(K-K^*)y_t\biggr\},$$

where we use (B.6.24) in the last line. Here the expectations are taken following the transition $y_{t+1} = (A-BK)y_t$ with initial state $y_0 \sim \mathcal{N}(0,\Psi_\epsilon)$. Also, we have

$$\begin{aligned}
&\mathbb{E}\biggl\{\sum_{t\geq 0} y_t^\top (K-K^*)^\top \Upsilon^{22}_{K^*}(K-K^*)y_t\biggr\} \\
&\qquad = \operatorname{tr}\bigl[\Phi_K (K-K^*)^\top \Upsilon^{22}_{K^*}(K-K^*)\bigr] \\
&\qquad \geq \|\Phi_K\|_2 \cdot \|\Upsilon^{22}_{K^*}\|_2 \cdot \operatorname{tr}\bigl[(K-K^*)^\top (K-K^*)\bigr] \\
&\qquad \geq \sigma_{\min}(\Psi_\epsilon)\cdot\sigma_{\min}(R)\cdot\|K-K^*\|_{\mathrm{F}}^2, \qquad \text{(B.6.26)}
\end{aligned}$$

where we use the fact that $\Phi_K = (A-BK)\Phi_K(A-BK) + \Psi_\epsilon \succeq \Psi_\epsilon$ and $\Upsilon^{22}_{K^*} = R + B^\top P_{K^*}B \succeq R$ in the last line. Combining (B.6.25) and (B.6.26), we have

$$J_1(K) - J_1(K^*) \geq \sigma_{\min}(\Psi_\epsilon)\cdot\sigma_{\min}(R)\cdot\|K-K^*\|_{\mathrm{F}}^2.$$

We conclude the proof of the lemma. □

B.6.5. Proof of Lemma B.4.5

Proof. Following from Proposition 3.2.3, we have

$$
\begin{aligned}
& J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1}) \\
& \qquad \leq \gamma^b \cdot \nabla_b J_2(K_N, \widetilde{b}_{h+1})^\top \bigl[\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\bigr] \\
& \qquad\qquad + (\gamma^b)^2 \cdot \nu_{K_N}/2 \cdot \bigl\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\bigr\|_2^2, \\
& J_2(K_N, \widetilde{b}_{h+1}) - J_2(K_N, b_{h+1}) \\
& \qquad \leq -\gamma^b \cdot \nabla_b J_2(K_N, \widetilde{b}_{h+1})^\top \bigl[\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\bigr] \\
& \qquad\qquad - (\gamma^b)^2 \cdot \iota_{K_N}/2 \cdot \bigl\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\bigr\|_2^2,
\end{aligned}
\tag{B.6.27}
$$

where ν_{K_N} and ι_{K_N} are defined in Proposition 3.2.3. Also, following from Proposition B.2.3, it holds that

$$
\bigl\|\nabla_b J_2(K_N, \widetilde{b}_{h+1})\bigr\|_2 \leq \mathrm{poly}_1\bigl(\|K_N\|_{\mathrm{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\bigr) \cdot \bigl[1 - \rho(A - BK_N)\bigr]^{-1}. \tag{B.6.28}
$$

Combining (B.6.27), (B.6.28), and the fact that $\nu_{K_N} \le \iota_{K_N} \le [1-\rho(A-BK_N)]^{-2} \cdot \text{poly}_2(\|K_N\|_2)$, we know that

(B.6.29)

$$\begin{aligned}
&\big|J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1})\big| \\
&\qquad \le (\gamma^b)^2 \cdot \text{poly}_2\big(\|K_N\|_2\big) \cdot \big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2^2 \cdot \big[1-\rho(A-BK_N)\big]^{-2} \\
&\qquad\qquad + \gamma^b \cdot \text{poly}_1\big(\|K_N\|_{\text{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\big) \cdot \big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2 \\
&\qquad\qquad \cdot \big[1-\rho(A-BK_N)\big]^{-1}.
\end{aligned}$$

Note that from the definition of $\widehat{\nabla}_b J_2(K_N, b_h)$ and $\nabla_b J_2(K_N, b_h)$ in (B.4.31) and (B.4.33), respectively, it holds by triangle inequality that

$$\begin{aligned}
&\big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2 \\
&\qquad \le \|\widehat{\Upsilon}^{22}_{K_N} - \Upsilon^{22}_{K_N}\|_2 \cdot \|K_N\|_2 \cdot \|\widehat{\mu}_{K_N, b_h}\|_2 + \|\Upsilon^{22}_{K_N}\|_2 \cdot \|K_N\|_2 \cdot \|\widehat{\mu}_{K_N, b_h} - \mu_{K_N, b_h}\|_2 \\
&\qquad\qquad + \|\widehat{\Upsilon}^{22}_{K_N} - \Upsilon^{22}_{K_N}\|_2 \cdot \|b_h\|_2 + \|\widehat{\Upsilon}^{21}_{K_N} - \Upsilon^{21}_{K_N}\|_2 \cdot \|\widehat{\mu}_{K_N, b_h}\|_2 + \|\widehat{q}_{K_N, b_h} - q_{K_N, b_h}\|_2 \\
&\qquad\qquad + \|\Upsilon^{21}_{K_N}\|_2 \cdot \|\widehat{\mu}_{K_N, b_h} - \mu_{K_N, b_h}\|_2.
\end{aligned}$$

By Theorem B.2.8, combining the fact that $J_2(K_N, b_h) \le J_2(K_N, b_0)$ and the fact that $\|\mu_{K_N, b}\|_2 \le J(K_N, b_0)/\sigma_{\min}(Q)$, we know that with probability at least $1-(T^b_n)^{-4}-(\widetilde{T}^b_n)^{-6}$, it holds for any $\rho \in (\rho(A - BK_N), 1)$ that

(B.6.30)

$$\big\|\nabla_b J_2(K_N, b_h) - \widehat{\nabla}_b J_2(K_N, b_h)\big\|_2$$

$$\leq \lambda_{K_N}^{-1} \cdot \text{poly}_3\big(\|K_N\|_{\text{F}}, \|b_h\|_2, \|\mu\|_2, J_2(K_N, b_0)\big) \cdot \bigg[\frac{\log^3 T_n^b}{(T_n^b)^{1/4}(1-\rho)^2} + \frac{\log^{1/2} \widetilde{T}_n^b}{(\widetilde{T}_n^b)^{1/8} \cdot (1-\rho)}\bigg].$$

Following from the choices of γ^b, T_n^b, and $\widetilde{T}_n^b$ in the statement of Theorem B.2.4, it holds that

$$\gamma^b \cdot \text{poly}_1\big(\|K_N\|_{\text{F}}, \|b_h\|_2, \|\mu\|_2, J(K_N, b_0)\big) \cdot \lambda_{K_N}^{-1} \cdot \text{poly}_3\big(\|K_N\|_{\text{F}}, \|b_h\|_2, \|\mu\|_2, J_2(K_N, b_0)\big)$$

$$\cdot \bigg[\frac{\log^3 T_n^b}{(T_n^b)^{1/4}(1-\rho)^2} + \frac{\log^{1/2} \widetilde{T}_n^b}{(\widetilde{T}_n^b)^{1/8} \cdot (1-\rho)}\bigg] \cdot \big[1 - \rho(A - BK_N)\big]^{-1} + \big[1 - \rho(A - BK_N)\big]^{-2}$$

$$\cdot \text{poly}_3\big(\|K_N\|_{\text{F}}, \|b_h\|_2, \|\mu\|_2, J_2(K_N, b_0)\big) \cdot \bigg[\frac{\log^6 T_n^b}{(T_n^b)^{1/2}(1-\rho)^4} + \frac{\log \widetilde{T}_n^b}{(\widetilde{T}_n^b)^{1/4} \cdot (1-\rho)^2}\bigg]$$

$$\cdot (\gamma^b)^2 \cdot \text{poly}_2\big(\|K_N\|_2\big) \cdot \lambda_{K_N}^{-1}$$

$$\leq \nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2.$$

Further combining (B.6.29) and (B.6.30), it holds with probability at least $1 - \varepsilon^{15}$ that

$$\big|J_2(K_N, b_{h+1}) - J_2(K_N, \widetilde{b}_{h+1})\big| \leq \nu_{K_N} \cdot \gamma^b \cdot \varepsilon/2.$$

We then finish the proof of the lemma. □

B.6.6. Proof of Lemma B.4.6

Proof. We show that $\zeta_{K,b} \in \mathcal{V}_\zeta$ and $\xi(\zeta) \in \mathcal{V}_\xi$ for any $\zeta \in \mathcal{V}_\zeta$ separately.

Part 1. First we show that $\zeta_{K,b} \in \mathcal{V}_\zeta$. Note that from Definition B.2.7, we know that $\zeta^1_{K,b} = J(K,b)$ satisfies that $0 \leq \zeta^1_{K,b} \leq J(K_0, b_0)$. It remains to show that $\zeta^2_{K,b} = \alpha_{K,b}$ satisfies that $\|\zeta^2_{K,b}\|_2 \leq M_\zeta$. By the definition of $\alpha_{K,b}$ in (B.2.6), we know that

$$
\begin{aligned}
\|\alpha_{K,b}\|_2^2 &\leq \|\Upsilon_K\|_{\mathrm{F}}^2 + \|\Upsilon_K\|_2^2 \cdot \big(\|\mu_{K,b}\|_2^2 + \|\mu^u_{K,b}\|_2^2\big) \\
&\quad + \big(\|A\|_2 + \|B\|_2\big)^2 \cdot \big(\|P_K\|_2 \cdot \|\overline{A}\mu + d\|_2 + \|f_{K,b}\|_2\big)^2
\end{aligned}
\tag{B.6.31}
$$

where $f_{K,b} = (I - A + BK)^{-\top}[(A - BK)^\top P_K(Bb + \overline{A}\mu + d) - K^\top Rb]$ and for notational simplicity, we denote by $\mu^u_{K,b} = -K\mu_{K,b} + b$. We only need to bound Υ_K, $\mu_{K,b}$, $\mu^u_{K,b}$, P_K, and $f_{K,b}$. Note that by Proposition B.2.2, the expected total cost $J(K,b)$ takes the form of

$$
J(K,b) = \mathrm{tr}(P_K\Psi_\epsilon) + \mu_{K,b}^\top Q\mu_{K,b} + (\mu^u_{K,b})^\top R\mu^u_{K,b} + \sigma^2 \cdot \mathrm{tr}(R) + \mu^\top\overline{Q}\mu.
$$

Thus, we have

$$
\begin{aligned}
&J(K_0, b_0) \geq J(K,b) \geq \sigma_{\min}(\Psi_\omega) \cdot \mathrm{tr}(P_K) \geq \sigma_{\min}(\Psi_\omega) \cdot \|P_K\|_2, \\
&J(K_0, b_0) \geq J(K,b) \geq \mu_{K,b}^\top Q\mu_{K,b} \geq \sigma_{\min}(Q) \cdot \|\mu_{K,b}\|_2, \\
&J(K_0, b_0) \geq J(K,b) \geq (\mu^u_{K,b})^\top R\mu^u_{K,b} \geq \sigma_{\min}(R) \cdot \|\mu^u_{K,b}\|_2,
\end{aligned}
$$

which imply that

$$
\begin{aligned}
&\|P_K\|_2 \leq J(K_0, b_0)/\sigma_{\min}(\Psi_\omega),\\
&\|\mu_{K,b}\|_2 \leq J(K_0, b_0)/\sigma_{\min}(Q),\\
&\|\mu_{K,b}^u\|_2 \leq J(K_0, b_0)/\sigma_{\min}(R).
\end{aligned}
\tag{B.6.32}
$$

For Υ_K, it holds that

$$
\Upsilon_K = \begin{pmatrix} Q & \mathbf{0} \\ \mathbf{0} & R \end{pmatrix} + \begin{pmatrix} A^\top \\ B^\top \end{pmatrix} P_K \begin{pmatrix} A & B \end{pmatrix},
$$

which gives

$$
\begin{aligned}
&\|\Upsilon_K\|_{\mathrm{F}} \leq (\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}}) + \big(\|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2\big) \cdot \|P_K\|_{\mathrm{F}},\\
&\|\Upsilon_K\|_2 \leq (\|Q\|_2 + \|R\|_2) + \big(\|A\|_2 + \|B\|_2\big)^2 \cdot \|P_K\|_2.
\end{aligned}
$$

Combining (B.6.32) and the fact that $\|P_K\|_{\mathrm{F}} \leq \sqrt{m} \cdot \|P_K\|_2$, we know that

$$
\begin{aligned}
&\|\Upsilon_K\|_{\mathrm{F}} \leq \big(\|Q\|_{\mathrm{F}} + \|R\|_{\mathrm{F}}\big) + \big(\|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2\big) \cdot \sqrt{m} \cdot J(K_0, b_0)/\sigma_{\min}(\Psi_\omega),\\
&\|\Upsilon_K\|_2 \leq \big(\|Q\|_2 + \|R\|_2\big) + \big(\|A\|_2 + \|B\|_2\big)^2 \cdot J(K_0, b_0)/\sigma_{\min}(\Psi_\omega).
\end{aligned}
\tag{B.6.33}
$$

Now, we upper bound the vector $f_{K,b}$. Note that by algebra, the vector $f_{K,b}$ takes the form of

$$
f_{K,b} = -P_K \mu_{K,b} + (I - A + BK)^{-T}(Q\mu_{K,b} - K^\top R \mu_{K,b}^u).
$$

Therefore, we upper bound $f_{K,b}$ as

(B.6.34)
$$\|f_{K,b}\|_2 \le J(K_0, b_0)^2 \cdot \sigma_{\min}^{-1}(\Psi_\omega) \cdot \sigma_{\min}^{-1}(Q) + \big[1 - \rho(A - BK)\big]^{-1} \cdot (\kappa_Q + \kappa_R \cdot \|K\|_{\mathrm{F}})$$

Combining (B.6.31), (B.6.32), (B.6.33), and (B.6.34), it holds that

$$\|\zeta_{K,b}^2\|_2 = \|\alpha_{K,b}\|_2 \le M_{\zeta,1} + M_{\zeta,2} \cdot (1 + \|K\|_{\mathrm{F}}) \cdot [1 - \rho(A - BK)]^{-1}.$$

Therefore, it holds that $\zeta_{K,b} \in \mathcal{V}_\zeta$.

Part 2. Now we show that for any $\zeta \in \mathcal{V}_\zeta$, we have $\xi(\zeta) \in \mathcal{V}_\xi$. Recall that from (B.4.41), it holds that

(B.6.35)
$$\xi^1(\zeta) = \zeta^1 - J(K, b), \quad \xi^2(\zeta) = \mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1 + \Theta_{K,b}\zeta^2 - \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big].$$

Then we have

$$\big|\xi^1(\zeta)\big| = \big|\zeta^1 - J(K,b)\big| \le J(K_0, b_0), \tag{B.6.36}$$

where we use the fact that since $\zeta \in \mathcal{V}_\zeta$, we have $0 \le \zeta^1 \le J(K_0, b_0)$ by Definition B.2.7. Also, by (B.6.35), we have

$$\big\|\xi^2(\zeta)\big\|_2 \le \underbrace{\Big\|\mathbb{E}_{\pi_{K,b}}\big[\psi(x,u)\big]\zeta^1\Big\|_2}_{B_1} + \underbrace{\|\Theta_{K,b}\|_2 \cdot \|\zeta^2\|_2}_{B_2} + \underbrace{\Big\|\mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]\Big\|_2}_{B_3}. \tag{B.6.37}$$

Note that we upper bound B_1 as

$$B_1 \leq J(K_0, b_0) \cdot \Big\| \mathbb{E}_{\pi_{K,b}} \big[\psi(x, u) \big] \Big\|_2. \tag{B.6.38}$$

Following from the definition of $\psi(x, u)$ in (B.2.5), we know that

$$\Big\| \mathbb{E}_{\pi_{K,b}} \big[\psi(x, u) \big] \Big\|_2 \leq \|\Sigma_z\|_{\mathrm{F}}, \tag{B.6.39}$$

where Σ_z is defined as

$$\Sigma_z = \mathrm{Cov}\left[\begin{pmatrix} x \\ u \end{pmatrix} \right] = \begin{pmatrix} \Phi_K & -\Phi_K K^\top \\ -K\Phi_K & K\Phi_K K^\top + \sigma^2 I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma^2 I \end{pmatrix} + \begin{pmatrix} I \\ -K \end{pmatrix} \Phi_K \begin{pmatrix} I \\ -K \end{pmatrix}^\top.$$

Combining (B.6.38) and (B.6.39), we have

$$B_1 \leq J(K_0, b_0) \cdot \|\Sigma_z\|_{\mathrm{F}}. \tag{B.6.40}$$

By Proposition B.2.6, we upper bound B_2 as

$$B_2 \leq 4(1 + \|K\|_{\mathrm{F}}^2)^3 \cdot \|\Phi_K\|_2^2 \cdot (M_{\zeta,1} + M_{\zeta,2}) \cdot \big[1 - \rho(A - BK)\big]^{-1}, \tag{B.6.41}$$

where we use the fact that $\zeta \in \mathcal{V}_\zeta$ and Definition B.2.7. As for the term B_3 in (B.6.37), we utilize the following lemma to provide an upper bound.

Lemma B.6.2. The vector $\mathbb{E}_{\pi_{K,b}}[c(x,u)\psi(x,u)]$ takes the following form,

$$\mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big] = \begin{pmatrix} 2\mathrm{svec}\big[\Sigma_z \mathrm{diag}(Q,R)\Sigma_z + \langle \Sigma_z, \mathrm{diag}(Q,R)\rangle \Sigma_z\big] \\ \Sigma_z \begin{pmatrix} 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix} \end{pmatrix}$$

$$+ \big[\mu_{K,b}^\top Q \mu_{K,b} + (\mu_{K,b}^u)^\top R \mu_{K,b}^u + \mu^\top \overline{Q}\mu\big] \begin{pmatrix} \mathrm{svec}(\Sigma_z) \\ \mathbf{0}_m \\ \mathbf{0}_k \end{pmatrix}.$$

Here the matrix Σ_z takes the form of

$$\Sigma_z = \begin{pmatrix} \Phi_K & -\Phi_K K^\top \\ -K\Phi_K & K\Phi_K K^\top + \sigma^2 \cdot I \end{pmatrix}.$$

Proof. See §B.6.11 for a detailed proof. □

From Lemma B.6.2 and (B.6.32), it holds that

$$B_3 \leq 3\big[\|Q\|_\mathrm{F} + \|R\|_\mathrm{F} + J(K_0,b_0)\cdot \|Q\|_2/\sigma_{\min}(Q) + J(K_0,b_0)\cdot \|R\|_2/\sigma_{\min}(R)\big]\cdot \|\Sigma_z\|_2^2. \tag{B.6.42}$$

Moreover, by the definition of Σ_z in (B.5.25), combining the triangle inequality, we have the following bounds for $\|\Sigma_z\|_\mathrm{F}$ and $\|\Sigma_z\|_2$,

$$\|\Sigma_z\|_\mathrm{F} \leq 2(d + \|K\|_\mathrm{F}^2)\cdot \|\Phi_K\|_2, \qquad \|\Sigma_z\|_2 \leq 2(1 + \|K\|_\mathrm{F}^2)\cdot \|\Phi_K\|_2. \tag{B.6.43}$$

Also, we have

$$J(K_0, b_0) \geq J(K, b) \geq \operatorname{tr}\big[(Q + K^\top RK)\Phi_K\big] \geq \|\Phi_K\|_2 \cdot \sigma_{\min}(Q),$$

which gives the upper bound for Φ_K as follows,

$$\|\Phi_K\|_2 \leq J(K_0, b_0)/\sigma_{\min}(Q). \tag{B.6.44}$$

Therefore, combining (B.6.37), (B.6.40), (B.6.41), (B.6.42), (B.6.43), and (B.6.44), we know that

$$\begin{aligned} \big\|\xi^2(\zeta)\big\|_2 \leq C \cdot (M_{\zeta,1} + M_{\zeta,2}) \cdot J(K_0, b_0)^2/\sigma^2_{\min}(Q) \\ \cdot \big(1 + \|K\|_{\mathrm{F}}^2\big)^3 \cdot \big[1 - \rho(A - BK)\big]^{-1}. \end{aligned} \tag{B.6.45}$$

By (B.6.36) and (B.6.45), we know that $\xi(\zeta) \in \mathcal{V}_\xi$ for any $\zeta \in \mathcal{V}_\zeta$. We conclude the proof of the lemma. □

B.6.7. Proof of Lemma B.4.7

Proof. Assume that $\widetilde{z}_0 \sim \mathcal{N}(\mu_\dagger, \Sigma_\dagger)$. Following from the fact that

$$\widetilde{z}_{t+1} = L\widetilde{z}_t + \nu + \delta_t,$$

it holds that

$$\widetilde{z}_t \sim \mathcal{N}\Bigg(L^t\mu_\dagger + \sum_{i=0}^{t-1} L^i \cdot \nu,\ (L^\top)^t\Sigma_\dagger L^t + \sum_{i=0}^{t-1}(L^\top)^i\Psi_\delta L^i\Bigg), \tag{B.6.46}$$

where

$$\Psi_\delta = \begin{pmatrix} \Psi_\omega & K\Psi_\omega \\ K\Psi_\omega & K\Psi_\omega K^\top + \sigma^2 I \end{pmatrix}.$$

From (B.4.47), we know that μ_z takes the form of

$$\mu_z = (I - L)^{-1}\nu = \sum_{j=0}^{\infty} L^j \nu. \tag{B.6.47}$$

Therefore, combining (B.6.46) and (B.6.47), we have

$$\mathbb{E}(\widehat{\mu}_z) = \mu_z + \frac{1}{\widetilde{T}} \sum_{t=1}^{\widetilde{T}} L^t \mu_\dagger - \frac{1}{\widetilde{T}} \sum_{t=1}^{\widetilde{T}} \sum_{i=t}^{\infty} L^i \nu. \tag{B.6.48}$$

We denote by

$$\mu_{\widetilde{T}} = \sum_{t=1}^{\widetilde{T}} L^t \mu_\dagger - \sum_{t=1}^{\widetilde{T}} \sum_{i=t}^{\infty} L^i \nu.$$

Meanwhile, it holds that

$$\begin{aligned} &\bigg\| \sum_{t=1}^{\widetilde{T}} L^t \mu_\dagger - \sum_{t=1}^{\widetilde{T}} \sum_{i=t}^{\infty} L^i \nu \bigg\|_2 \\ &\qquad \le \sum_{t=1}^{\widetilde{T}} \rho(L)^t \cdot \|\mu_\dagger\|_2 + \sum_{t=1}^{\widetilde{T}} \sum_{i=t}^{\infty} \rho(L)^i \cdot \|\nu\|_2 \\ &\qquad \le \big[1 - \rho(L)\big]^{-1} \cdot \|\mu_\dagger\|_2 + \big[1 - \rho(L)\big]^{-2} \cdot \|\nu\|_2 \\ &\qquad \le M_\mu \cdot (1 - \rho)^{-2} \cdot \|\mu_z\|_2, \end{aligned} \tag{B.6.49}$$

where M_μ is a positive absolute constant.

For the covariance, note that for any random variables $X \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \Sigma_2)$, we know that $Z = X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma)$, where $\|\Sigma\|_{\mathrm{F}} \le 2\|\Sigma_1\|_{\mathrm{F}} + 2\|\Sigma_2\|_{\mathrm{F}}$. Combining (B.6.46), we know that $\widehat{\mu}_z \sim \mathcal{N}(\mathbb{E}\widehat{\mu}_z, \widetilde{\Sigma}_{\widetilde{T}}/\widetilde{T})$, where $\widetilde{\Sigma}_{\widetilde{T}}$ satisfies that

$$\begin{aligned}\widetilde{T}/2 \cdot \|\widetilde{\Sigma}_{\widetilde{T}}\|_{\mathrm{F}} &\le \sum_{t=1}^{\widetilde{T}} \rho(L)^{2t} \cdot \|\Sigma_\dagger\|_{\mathrm{F}} + \sum_{t=1}^{\widetilde{T}} \sum_{i=0}^{t-1} \rho(L)^{2i} \cdot \|\Psi_\delta\|_{\mathrm{F}} \\ &\le \bigl[1 - \rho(L)^2\bigr]^{-1} \cdot \|\Sigma_\dagger\|_{\mathrm{F}} + \widetilde{T} \cdot \bigl[1 - \rho(L)^2\bigr]^{-1} \cdot \|\Psi_\delta\|_{\mathrm{F}},\end{aligned}$$

which implies that

$$\|\widetilde{\Sigma}_{\widetilde{T}}\|_{\mathrm{F}} \le M_\Sigma \cdot (1-\rho)^{-1} \cdot \|\Sigma_z\|_{\mathrm{F}}, \tag{B.6.50}$$

where M_Σ is a positive absolute constant. Combining (B.6.48), (B.6.49), and (B.6.50), we obtain that

$$\widehat{\mu}_z \sim \mathcal{N}\biggl(\mu_z + \frac{1}{\widetilde{T}}\mu_{\widetilde{T}},\ \frac{1}{\widetilde{T}}\widetilde{\Sigma}_{\widetilde{T}}\biggr),$$

where $\|\mu_{\widetilde{T}}\|_2 \le M_\mu \cdot (1-\rho)^{-2} \cdot \|\mu_z\|_2$ and $\|\widetilde{\Sigma}_{\widetilde{T}}\|_{\mathrm{F}} \le M_\Sigma \cdot (1-\rho)^{-1} \cdot \|\Sigma_z\|_{\mathrm{F}}$. Moreover, by the Gaussian tail inequality, it holds with probability at least $1 - \widetilde{T}^{-6}$ that

$$\|\widehat{\mu}_z - \mu_z\|_2 \le \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}} \cdot (1-\rho)^{-2} \cdot \mathrm{poly}\bigl(\|\Phi_K\|_2, \|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2\bigr).$$

Then we finish the proof of the lemma. □

B.6.8. Proof of Lemma B.4.8

Proof. We continue using the notations given in §B.4.3. We define

$$\widehat{F}(\zeta,\xi) = \Big\{\mathbb{E}(\widehat{\psi})\zeta^1 + \mathbb{E}\big[(\widehat{\psi}-\widehat{\psi}')\widehat{\psi}^\top\big]\zeta^2 - \mathbb{E}(c\widehat{\psi})\Big\}^\top \xi^2 + \big[\zeta^1 - \mathbb{E}(c)\big]\cdot\xi^1 - 1/2\cdot\|\xi\|_2^2,$$

where $\widehat{\psi} = \widehat{\psi}(x,u)$ is the estimated feature vector. Here the expectation is only taken over the trajectory generated by the state transition and the policy $\pi_{K,b}$, conditioning on the randomness induced when calculating the estimated feature vectors. Thus, the function $\widehat{F}(\zeta,\xi)$ is still random, where the randomness comes from the estimated feature vectors. Note that $|F(\zeta,\xi)-\widetilde{F}(\zeta,\xi)| \le |F(\zeta,\xi)-\widehat{F}(\zeta,\xi)| + |\widehat{F}(\zeta,\xi)-\widetilde{F}(\zeta,\xi)|$. Thus, we only need to upper bound $|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)|$ and $|\widehat{F}(\zeta,\xi)-\widetilde{F}(\zeta,\xi)|$.

Part 1. First we upper bound $|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)|$. Note that by algebra, we have

$$\begin{aligned} &\big|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big| \\ &\quad = \bigg|\Big\{\mathbb{E}(\psi-\widehat{\psi})\zeta^1 + \mathbb{E}\big[(\psi-\psi')\psi^\top - (\widehat{\psi}-\widehat{\psi}')\widehat{\psi}^\top\big]\zeta^2 - \mathbb{E}\big[c(\psi-\widehat{\psi})\big]\Big\}^\top\xi^2\bigg| \\ &\quad \le \mathbb{E}\big(\|\psi-\widehat{\psi}\|_2\big)\cdot\Big[|\zeta^1| + \mathbb{E}\big(\|\psi-\psi'\|_2 + 2\|\widehat{\psi}\|_2\big)\cdot\|\zeta^2\|_2 + \mathbb{E}(c)\Big]\cdot\|\xi^2\|_2, \end{aligned} \tag{B.6.51}$$

where the expectation is only taken over the trajectory generated by the state transition and the policy $\pi_{K,b}$. From Lemma B.4.7, it holds that

$$\mathbb{P}\big(\|\widehat{\mu}_z - \mu_z + 1/\widetilde{T}\cdot\mu_{\widetilde{T}}\|_2 \le C_1\big) \ge 1-\widetilde{T}^{-6}. \tag{B.6.52}$$

Therefore, combining (B.6.52), it holds with probability at least $1-\widetilde{T}^{-6}$ that

$$\mathbb{E}\big(\|\psi-\psi'\|_2+2\|\widehat{\psi}\|_2\big)\le \operatorname{poly}\big(\|\Phi_K\|_2,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big), \tag{B.6.53}$$

where the expectation is conditioned on the randomness induced when calculating the estimated feature vectors. Also, we know that

$$\mathbb{E}(c)\le \operatorname{poly}\big(\|\Phi_K\|_2,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big). \tag{B.6.54}$$

Therefore, combining (B.6.51), (B.6.53), (B.6.54), and Definition B.2.7, it holds with probability at least $1-\widetilde{T}^{-6}$ that

$$\big|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big|\le \mathbb{E}\big(\|\psi-\widehat{\psi}\|_2\big)\cdot \operatorname{poly}\big(\|\Phi_K\|_2,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big). \tag{B.6.55}$$

Following from the definitions of $\psi(x,u)$ in (B.2.5) and $\widehat{\psi}(x,u)$ in (B.2.14), we upper bound $\|\psi(x,u)-\widehat{\psi}(x,u)\|_2$ for any x and u as

$$\begin{aligned}\|\psi(x,u)-\widehat{\psi}(x,u)\|_2^2&=\|\widehat{\mu}_z-\mu_z\|_2^2+\big\|z(\widehat{\mu}_z-\mu_z)^\top+(\widehat{\mu}_z-\mu_z)z^\top\big\|_{\mathrm{F}}^2+\|\mu_z\mu_z^\top-\widehat{\mu}_z\widehat{\mu}_z^\top\|_{\mathrm{F}}^2\\&\le \operatorname{poly}\big(\|\Phi_K\|_2,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big)\cdot\|\widehat{\mu}_z-\mu_z\|_2^2, \qquad \text{(B.6.56)}\end{aligned}$$

where μ_z is defined in (B.4.47), $\widehat{\mu}_z$ is defined in (B.4.48), and $z=(x^\top,u^\top)^\top$. Also, by Lemma B.4.7, we know that

$$\|\widehat{\mu}_z-\mu_z\|_2\le \frac{\log\widetilde{T}}{\widetilde{T}^{1/4}}\cdot(1-\rho)^{-2}\cdot \operatorname{poly}\big(\|\Phi_K\|_2,\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big), \tag{B.6.57}$$

which holds with probability at least $1-\widetilde{T}^{-6}$. Combining (B.6.55), (B.6.56), and (B.6.57), it holds with probability at least $1-\widetilde{T}^{-6}$ that

$$\big|F(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big| \le \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}}\cdot(1-\rho)^{-2}\cdot \mathrm{poly}\big(\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big). \tag{B.6.58}$$

Part 2. We now upper bound $|\widehat{F}(\zeta,\xi)-\widetilde{F}(\zeta,\xi)|$ in the sequel. By definitions, we have

$$\begin{aligned}
&\big|\widetilde{F}(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big| \\
&\quad=\bigg|\Big\{\mathbb{E}(\widetilde{\psi}-\widehat{\psi})\zeta^1+\mathbb{E}\big[(\widetilde{\psi}-\widetilde{\psi}')\widetilde{\psi}^\top-(\widehat{\psi}-\widehat{\psi}')\widehat{\psi}^\top\big]\zeta^2-\mathbb{E}(\widetilde{c}\widetilde{\psi}-\widehat{c}\widehat{\psi})\Big\}^\top\xi^2+\mathbb{E}(\widehat{c}-\widetilde{c})\xi^1\bigg| \\
&\text{(B.6.59)} \\
&\quad\le\bigg|\Big\{\mathbb{E}(\widehat{\psi})\zeta^1+\mathbb{E}(\widehat{\psi}\widehat{\psi}^\top)\zeta^2-\mathbb{E}(\widehat{c}\widehat{\psi})\Big\}^\top\xi^2+\mathbb{E}(\widehat{c})\xi^1\bigg|\cdot\mathbb{1}_{\mathcal{E}^c} \\
&\qquad\quad+\Big|\big[\mathbb{E}(\widehat{\psi}'\widehat{\psi}^\top)\zeta^2\big]^\top\xi^2\Big|\cdot\mathbb{1}_{(\mathcal{E}'\cap\mathcal{E})^c},
\end{aligned}$$

where we define the event $\mathcal{E}'$ as

$$\mathcal{E}'=\bigg(\bigcap_{t\in[T]}\Big\{\big|\|z_t'-\mu_z+1/\widetilde{T}\cdot\mu_{\widetilde{T}}\|_2^2-\mathrm{tr}(\widetilde{\Sigma}_z)\big|\le C_1\cdot\log T\cdot\|\widetilde{\Sigma}_z\|_2\Big\}\bigg)\bigcap\mathcal{E}_2,$$

where $\mathcal{E}_2$ is defined in (B.4.55). Combining the fact that $\mathbb{P}(\mathcal{E}_2)\ge 1-\widetilde{T}^{-6}$ and Lemma B.7.3, it holds that $\mathbb{P}(\mathcal{E}')\ge 1-T^{-5}-\widetilde{T}^{-6}$. Following a similar argument as in **Part 1**, it holds from (B.6.59) that

$$\big|\widetilde{F}(\zeta,\xi)-\widehat{F}(\zeta,\xi)\big|\le\bigg(\frac{1}{T}+\frac{1}{\widetilde{T}^{1/4}}\bigg)\cdot\mathrm{poly}\big(\|K\|_{\mathrm{F}},\|b\|_2,\|\mu\|_2,J(K_0,b_0)\big) \tag{B.6.60}$$

for sufficiently large T and $\widetilde{T}$.

Now, combining (B.6.58) and (B.6.60), by triangle inequality, it holds with probability at least $1 - \widetilde{T}^{-6}$ that

$$\big|F(\zeta,\xi) - \widetilde{F}(\zeta,\xi)\big| \le \biggl(\frac{1}{2T} + \frac{\log \widetilde{T}}{\widetilde{T}^{1/4}}\biggr) \cdot (1-\rho)^{-2} \cdot \mathrm{poly}\bigl(\|K\|_{\mathrm{F}}, \|b\|_2, \|\mu\|_2, J(K_0, b_0)\bigr).$$

We finish the proof of the lemma. □

B.6.9. Proof of Lemma B.5.2

Proof. Recall that the feature vector $\psi(x,u)$ takes the following form

$$\psi(x,u) = \begin{pmatrix} \mathrm{svec}\bigl[(z-\mu_z)(z-\mu_z)^\top\bigr] \\ z-\mu_z \end{pmatrix}.$$

We then have

$$\psi(x,u) - \psi(x',u') = \begin{pmatrix} \mathrm{svec}\bigl[yy^\top - (Ly+\delta)(Ly+\delta)^\top\bigr] \\ y - (Ly+\delta) \end{pmatrix}, \tag{B.6.61}$$

where we denote by $y = z - \mu_z$, and (x',u') is the state-action pair after (x,u) following the state transition and the policy $\pi_{K,b}$. Therefore, for any symmetric matrices M, N and any vectors m, n, it holds from (B.2.7) and (B.6.61) that

$$\begin{pmatrix} \mathrm{svec}(M) \\ m \end{pmatrix}^\top \Theta_{K,b} \begin{pmatrix} \mathrm{svec}(N) \\ n \end{pmatrix}$$

$$= \mathbb{E}_{y,\delta}\Biggl\{ \begin{pmatrix} \mathrm{svec}(M) \\ m \end{pmatrix}^\top \begin{pmatrix} \mathrm{svec}(yy^\top) \\ y \end{pmatrix} \begin{pmatrix} \mathrm{svec}\bigl[yy^\top - (Ly+\delta)(Ly+\delta)^\top\bigr] \\ y - (Ly+\delta) \end{pmatrix}^\top \begin{pmatrix} \mathrm{svec}(N) \\ n \end{pmatrix} \Biggr\}$$

$$= \mathbb{E}_{y,\delta}\Bigl\{ \bigl(\langle M, yy^\top\rangle + m^\top y\bigr) \cdot \bigl[\langle N, yy^\top - (Ly+\delta)(Ly+\delta)^\top\rangle + n^\top(y - Ly - \delta)\bigr] \Bigr\}$$

(B.6.62)

$$= \underbrace{\mathbb{E}_y\bigl(\langle yy^\top, M\rangle \cdot \langle yy^\top - Lyy^\top L^\top - \Psi_\delta, N\rangle\bigr)}_{A_1} + \underbrace{\mathbb{E}_y\bigl(\langle yy^\top, M\rangle \cdot n^\top(y - Ly)\bigr)}_{A_2}$$

$$+ \underbrace{\mathbb{E}_y\bigl(m^\top y \cdot \langle yy^\top - Lyy^\top L^\top - \Psi_\delta, N\rangle\bigr)}_{A_3} + \underbrace{\mathbb{E}_y\bigl[m^\top y \cdot n^\top(y - Ly)\bigr]}_{A_4},$$

where the expectations are taken over $y \sim \mathcal{N}(0, \Sigma_z)$ and $\delta \sim \mathcal{N}(0, \Psi_\delta)$. We evaluate the terms A_1, A_2, A_3, and A_4 in the sequel.

For the terms A_2 and A_3 in (B.6.62), by the fact that $y = z - \mu_z \sim \mathcal{N}(0, \Sigma_z)$, we know that these two terms vanish. For A_4, it holds that

(B.6.63) $$A_4 = \mathbb{E}_y\bigl[m^\top y \cdot (y - Ly)^\top n\bigr] = \mathbb{E}_y\bigl[m^\top yy^\top (I - L)^\top n\bigr] = m^\top \Sigma_z (I - L)^\top n.$$

For A_1, by algebra, we have

$$\begin{aligned} A_1 &= \mathbb{E}_y\big(\langle yy^\top, M\rangle \cdot \langle yy^\top - Lyy^\top L^\top - \Psi_\delta, N\rangle\big) \\ &= \mathbb{E}_y\big(\langle yy^\top, M\rangle \cdot \langle yy^\top - Lyy^\top L^\top, N\rangle\big) - \mathbb{E}_y\big(\langle yy^\top, M\rangle \cdot \langle \Psi_\delta, N\rangle\big) \\ &= \mathbb{E}_y\big[y^\top My \cdot y^\top(N - L^\top NL)y\big] - \langle \Sigma_z, M\rangle \cdot \langle \Psi_\delta, N\rangle \\ &= \mathbb{E}_{u\sim\mathcal{N}(0,I)}\big[u^\top \Sigma_z^{1/2} M \Sigma_z^{1/2} u \cdot u^\top \Sigma_z^{1/2}(N - L^\top NL)\Sigma_z^{1/2} u\big] - \langle \Sigma_z, M\rangle \cdot \langle \Psi_\delta, N\rangle. \qquad \text{(B.6.64)} \end{aligned}$$

Now, by applying Lemma B.7.1 to the first term on the RHS of (B.6.64), we know that

$$\begin{aligned} A_1 &= 2\,\mathrm{tr}\big[\Sigma_z^{1/2} M \Sigma_z^{1/2} \cdot \Sigma_z^{1/2}(N - L^\top NL)\Sigma_z^{1/2}\big] \\ &\qquad + \mathrm{tr}(\Sigma_z^{1/2} M \Sigma_z^{1/2}) \cdot \mathrm{tr}\big[\Sigma_z^{1/2}(N - L^\top NL)\Sigma_z^{1/2}\big] - \langle \Sigma_z, M\rangle \cdot \langle \Psi_\delta, N\rangle \\ &= 2\big\langle M, \Sigma_z(N - L^\top NL)\Sigma_z\big\rangle + \big\langle \Sigma_z, M\big\rangle \cdot \langle \Sigma_z - L\Sigma_z L^\top - \Psi_\delta, N\rangle \\ &= 2\big\langle M, \Sigma_z(N - L^\top NL)\Sigma_z\big\rangle, \end{aligned}$$

where we use the fact that $\Sigma_z = L\Sigma_z L^\top + \Psi_\delta$ in the last equality. By using the property of the operator svec$(\cdot)$ and the definition of the symmetric Kronecker product, we obtain that

$$\begin{aligned} A_1 &= 2\mathrm{svec}(M)^\top \mathrm{svec}\big[\Sigma_z(N - L^\top NL)\Sigma_z\big] \\ &= 2\mathrm{svec}(M)^\top\big[\Sigma_z \otimes_s \Sigma_z - (\Sigma_z L^\top) \otimes_s (\Sigma_z L^\top)\big]\mathrm{svec}(N) \\ &= 2\mathrm{svec}(M)^\top\big[(\Sigma_z \otimes_s \Sigma_z)(I - L \otimes_s L)^\top\big]\mathrm{svec}(N). \qquad \text{(B.6.65)} \end{aligned}$$

Combining (B.6.62), (B.6.63), and (B.6.65), we obtain that

$$\begin{pmatrix} \operatorname{svec}(M) \\ m \end{pmatrix}^\top \Theta_{K,b} \begin{pmatrix} \operatorname{svec}(N) \\ n \end{pmatrix}$$

$$= \operatorname{svec}(M)^\top \bigl[2(\Sigma_z \otimes_s \Sigma_z)(I - L \otimes_s L)^\top\bigr] \operatorname{svec}(N) + m^\top \Sigma_z (I - L)^\top n$$

$$= \begin{pmatrix} \operatorname{svec}(M) \\ m \end{pmatrix}^\top \begin{pmatrix} 2(\Sigma_z \otimes_s \Sigma_z)(I - L \otimes_s L)^\top & 0 \\ 0 & \Sigma_z (I - L)^\top \end{pmatrix} \begin{pmatrix} \operatorname{svec}(N) \\ n \end{pmatrix}.$$

Thus, the matrix $\Theta_{K,b}$ takes the following form,

$$\Theta_{K,b} = \begin{pmatrix} 2(\Sigma_z \otimes_s \Sigma_z)(I - L \otimes_s L)^\top & 0 \\ 0 & \Sigma_z (I - L)^\top \end{pmatrix},$$

which concludes the proof of the lemma. □

B.6.10. Proof of Lemma B.5.3

Proof. From the definition of $\widetilde{\Theta}_{K,b}$ in (B.2.9), it holds that

$$\|\widetilde{\Theta}_{K,b}^{-1}\|_2^2 \le 1 + \|\Theta_{K,b}^{-1}\|_2^2 + \|\Theta_{K,b}^{-1}\widetilde{\sigma}_z\|_2^2, \tag{B.6.66}$$

where $\widetilde{\sigma}_z$ is defined as

$$\widetilde{\sigma}_z = \mathbb{E}_{\pi_{K,b}}\bigl[\psi(x,u)\bigr] = \begin{pmatrix} \operatorname{svec}(\Sigma_z) \\ \mathbf{0}_{k+m} \end{pmatrix}.$$

We bound the RHS of (B.6.66) in the sequel. For the term $\Theta_{K,b}^{-1}\widetilde{\sigma}_z$, combining Lemma B.5.2, we have

$$\begin{aligned}
\Theta_{K,b}^{-1}\widetilde{\sigma}_z &= \begin{pmatrix} 1/2\cdot(I-L\otimes_s L)^{-\top}(\Sigma_z\otimes_s\Sigma_z)^{-1}\cdot\mathrm{svec}(\Sigma_z) \\ \mathbf{0}_{k+m}\end{pmatrix} \\
&= \begin{pmatrix} 1/2\cdot(I-L\otimes_s L)^{-\top}(\Sigma_z^{-1}\otimes_s\Sigma_z^{-1})\cdot\mathrm{svec}(\Sigma_z) \\ \mathbf{0}_{k+m}\end{pmatrix} \\
&= \begin{pmatrix} 1/2\cdot(I-L\otimes_s L)^{-\top}\cdot\mathrm{svec}(\Sigma_z^{-1}) \\ \mathbf{0}_{k+m}\end{pmatrix}, \qquad \text{(B.6.67)}
\end{aligned}$$

where we use the property of the symmetric Kronecker product in the second and last line. By taking the spectral norm on both sides of (B.6.67), it holds that

$$\begin{aligned}
\|\Theta_{K,b}^{-1}\widetilde{\sigma}_z\|_2 &= 1/2\cdot\big\|(I-L\otimes_s L)^{-\top}\cdot\mathrm{svec}(\Sigma_z^{-1})\big\|_2 \\
&\le 1/2\cdot\big\|(I-L\otimes_s L)^{-\top}\big\|_2\cdot\big\|\mathrm{svec}(\Sigma_z^{-1})\big\|_2 \\
&\le 1/2\cdot\big[1-\rho^2(L)\big]^{-1}\cdot\|\Sigma_z^{-1}\|_{\mathrm{F}} \\
&\le 1/2\cdot\sqrt{k+m}\cdot\big[1-\rho^2(L)\big]^{-1}\cdot\|\Sigma_z^{-1}\|_2 \\
&= 1/2\cdot\sqrt{k+m}\cdot\big[1-\rho^2(L)\big]^{-1}\cdot\sigma_{\min}^{-1}(\Sigma_z), \qquad \text{(B.6.68)}
\end{aligned}$$

where in the third line we use Lemma B.7.2 to the matrix $L\otimes_s L$. Similarly, we upper bound $\|\Theta_{K,b}^{-1}\|_2$ in the sequel

$$\|\Theta_{K,b}^{-1}\|_2 \le \min\Big\{1/2\cdot\big[1-\rho^2(L)\big]^{-1}\sigma_{\min}^{-2}(\Sigma_z),\, \big[1-\rho(L)\big]^{-1}\sigma_{\min}^{-1}(\Sigma_z)\Big\}. \qquad \text{(B.6.69)}$$

Thus, combining (B.6.66), (B.6.68), and (B.6.69), we obtain that

$$
\begin{aligned}
\|\widetilde{\Theta}_{K,b}^{-1}\|_2^2 \le 1 + 1/2 \cdot \sqrt{k+m} \cdot \bigl[1-\rho^2(L)\bigr]^{-1} \cdot \sigma_{\min}^{-1}(\Sigma_z) \\
+ \min\Bigl\{1/2 \cdot \bigl[1-\rho^2(L)\bigr]^{-1}\sigma_{\min}^{-2}(\Sigma_z), \bigl[1-\rho(L)\bigr]^{-1}\sigma_{\min}^{-1}(\Sigma_z)\Bigr\}. \qquad \text{(B.6.70)}
\end{aligned}
$$

Now it remains to characterize $\sigma_{\min}(\Sigma_z)$. For any vectors $s \in \mathbb{R}^m$ and $r \in \mathbb{R}^k$, we have

$$
\begin{aligned}
\begin{pmatrix} s \\ r \end{pmatrix}^\top \Sigma_z \begin{pmatrix} s \\ r \end{pmatrix} &= \mathbb{E}_{x\sim\mathcal{N}(\mu_{K,b},\Phi_K),u\sim\pi_{K,b}(\cdot\,|\,x)}\Bigl\{\bigl[s^\top(x-\mu_{K,b}) + r^\top(u + K\mu_{K,b} - b)\bigr]^2\Bigr\} \\
&= \mathbb{E}_{x\sim\mathcal{N}(\mu_{K,b},\Phi_K),\eta\sim\mathcal{N}(0,I)}\Bigl\{\bigl[(s-K^\top r)^\top(x-\mu_{K,b}) + \sigma r^\top\eta\bigr]^2\Bigr\} \\
&= \mathbb{E}_{x\sim\mathcal{N}(\mu_{K,b},\Phi_K)}\Bigl\{\bigl[(s-K^\top r)^\top(x-\mu_{K,b})\bigr]^2\Bigr\} + \mathbb{E}_{\eta\sim\mathcal{N}(0,I)}\bigl[(\sigma r^\top\eta)^2\bigr]. \qquad \text{(B.6.71)}
\end{aligned}
$$

The first term on the RHS of (B.6.71) is lower bounded as

$$
\begin{aligned}
\mathbb{E}_{x\sim\mathcal{N}(\mu_{K,b},\Phi_K)}\Bigl\{\bigl[(s-K^\top r)^\top(x-\mu_{K,b})\bigr]^2\Bigr\} &= (s-K^\top r)^\top\Phi_K(s-K^\top r) \\
&\ge \|s-K^\top r\|_2^2 \cdot \sigma_{\min}(\Phi_K) \ge \|s-K^\top r\|_2^2 \cdot \sigma_{\min}(\Psi_\omega), \qquad \text{(B.6.72)}
\end{aligned}
$$

where the last inequality comes from the fact that $\sigma_{\min}(\Phi_K) \ge \sigma_{\min}(\Psi_\omega)$ by (3.2.3). The second term on the RHS of (B.6.71) takes the form of

$$
\mathbb{E}_{\eta\sim\mathcal{N}(0,I)}\bigl[(\sigma r^\top\eta)^2\bigr] = \sigma^2\|r\|_2^2. \qquad \text{(B.6.73)}
$$

Therefore, combining (B.6.71), (B.6.72), and (B.6.73), we have

$$\begin{aligned}\begin{pmatrix} s \\ r \end{pmatrix}^{\top} \Sigma_z \begin{pmatrix} s \\ r \end{pmatrix} &\geq \|s - K^{\top} r\|_2^2 \cdot \sigma_{\min}(\Psi_\omega) + \sigma^2 \|r\|_2^2 \\ &\geq \sigma_{\min}(\Psi_\omega) \cdot \|s\|_2^2 + \bigl[\sigma^2 - \|K\|_2^2 \cdot \sigma_{\min}(\Psi_\omega)\bigr] \cdot \|r\|_2^2.\end{aligned}$$

From this, we know that

$$\sigma_{\min}(\Sigma_z) \geq \min\bigl\{\sigma_{\min}(\Psi_\omega), \sigma^2 - \|K\|_2^2 \cdot \sigma_{\min}(\Psi_\omega)\bigr\}. \tag{B.6.74}$$

Thus, combining (B.6.70) and (B.6.74), we know that $\|\widetilde{\Theta}_{K,b}^{-1}\|_2$ is upper bounded by a constant $\widetilde{\lambda}_K$, where $\widetilde{\lambda}_K$ only depends on $\|K\|_2$ and $\rho(L) = \rho(A - BK)$. This finishes the proof of the lemma. □

B.6.11. Proof of Lemma B.6.2

Proof. First, note that the cost function $c(x, u)$ takes the following form,

$$c(x, u) = \psi(x, u)^{\top} \begin{pmatrix} \operatorname{svec}\bigl[\operatorname{diag}(Q, R)\bigr] \\ 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix} + \bigl[\mu_{K,b}^{\top} Q \mu_{K,b} + (\mu_{K,b}^u)^{\top} R \mu_{K,b}^u + \mu^{\top} \overline{Q} \mu\bigr].$$

For any matrix V and vectors v_x, v_u, it holds that

$$\mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big]^\top \begin{pmatrix} \mathrm{svec}(V) \\ v_x \\ v_u \end{pmatrix}$$

(B.6.75)

$$= \underbrace{\mathbb{E}_{\pi_{K,b}}\left\{ \psi(x,u)^\top \begin{pmatrix} \mathrm{svec}\big[\mathrm{diag}(Q,R)\big] \\ 2Q\mu_{K,b} \\ 2R\mu^u_{K,b} \end{pmatrix} \psi(x,u)^\top \begin{pmatrix} \mathrm{svec}(V) \\ v_x \\ v_u \end{pmatrix} \right\}}_{D_1}$$

$$+ \underbrace{\mathbb{E}_{\pi_{K,b}}\left\{ \psi(x,u)^\top (\mu_{K,b}^\top Q\mu_{K,b} + (\mu^u_{K,b})^\top R\mu^u_{K,b} + \mu^\top \overline{Q}\mu) \begin{pmatrix} \mathrm{svec}(V) \\ v_x \\ v_u \end{pmatrix} \right\}}_{D_2}.$$

In the sequel, we calculate D_1 and D_2 respectively.

Calculation of D_1. Note that by the definition of $\psi(x,u)$ in (B.2.5), it holds that

$$D_1 = \mathbb{E}_{\pi_{K,b}}\left\{\left[(z-\mu_z)^\top \mathrm{diag}(Q,R)(z-\mu_z) + (z-\mu_z)^\top \begin{pmatrix} 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix}\right]\right.$$

$$\left.\cdot \left[(z-\mu_z)^\top V(z-\mu_z) + (z-\mu_z)^\top \begin{pmatrix} v_x \\ v_u \end{pmatrix}\right]\right\}$$

$$= \mathbb{E}_{\pi_{K,b}}\big[(z-\mu_z)^\top \mathrm{diag}(Q,R)(z-\mu_z)\cdot(z-\mu_z)^\top V(z-\mu_z)\big] \tag{B.6.76}$$

$$+ \mathbb{E}_{\pi_{K,b}}\left[\begin{pmatrix} 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix}^\top (z-\mu_z)(z-\mu_z)^\top \begin{pmatrix} v_x \\ v_u \end{pmatrix}\right].$$

Here $z = (x^\top, u^\top)^\top$ and $\mu_z = \mathbb{E}_{\pi_{K,b}}(z)$. For the first term on the RHS of (B.6.76), note that $z - \mu_z \sim \mathcal{N}(0, \Sigma_z)$. Therefore, by Lemma B.7.1, we obtain that

$$\mathbb{E}_{\pi_{K,b}}\big[(z-\mu_z)^\top \mathrm{diag}(Q,R)(z-\mu_z)\cdot(z-\mu_z)^\top V(z-\mu_z)\big]$$

$$= 2\big\langle \Sigma_z \mathrm{diag}(Q,R)\Sigma_z, V\big\rangle + \big\langle \Sigma_z, \mathrm{diag}(Q,R)\big\rangle \cdot \langle \Sigma_z, V\rangle$$

$$= \mathrm{svec}\Big[2\Sigma_z \mathrm{diag}(Q,R)\Sigma_z + \big\langle \Sigma_z, \mathrm{diag}(Q,R)\big\rangle \cdot \Sigma_z\Big]^\top \mathrm{svec}(V). \tag{B.6.77}$$

Meanwhile, the second term on the RHS of (B.6.76) takes the form of

$$\mathbb{E}_{\pi_{K,b}}\left[\begin{pmatrix} 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix}^\top (z-\mu_z)(z-\mu_z)^\top \begin{pmatrix} v_x \\ v_u \end{pmatrix}\right] = \left[\Sigma_z \begin{pmatrix} 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix}\right]^\top \begin{pmatrix} v_x \\ v_u \end{pmatrix}. \tag{B.6.78}$$

Combining (B.6.76), (B.6.77), and (B.6.78), we obtain that

$$D_1 = \begin{pmatrix} 2\text{svec}\big[\Sigma_z \text{diag}(Q,R)\Sigma_z + \langle \Sigma_z, \text{diag}(Q,R)\rangle \Sigma_z\big] \\ \Sigma_z \begin{pmatrix} 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix} \end{pmatrix}^\top \begin{pmatrix} \text{svec}(V) \\ v_x \\ v_u \end{pmatrix}. \tag{B.6.79}$$

Calculation of D_2. By the definition of the feature vector $\psi(x,u)$ in (B.2.5), we know that

$$D_2 = (\mu_{K,b}^\top Q\mu_{K,b} + (\mu_{K,b}^u)^\top R\mu_{K,b}^u + \mu^\top \overline{Q}\mu) \begin{pmatrix} \text{svec}(\Sigma_z) \\ \mathbf{0}_m \\ \mathbf{0}_k \end{pmatrix}^\top \begin{pmatrix} \text{svec}(V) \\ v_x \\ v_u \end{pmatrix}. \tag{B.6.80}$$

Now, combining (B.6.75), (B.6.79), and (B.6.80), it holds that

$$\begin{aligned} \mathbb{E}_{\pi_{K,b}}\big[c(x,u)\psi(x,u)\big] = &\begin{pmatrix} 2\text{svec}\big[\Sigma_z \text{diag}(Q,R)\Sigma_z + \langle \Sigma_z, \text{diag}(Q,R)\rangle \Sigma_z\big] \\ \Sigma_z \begin{pmatrix} 2Q\mu_{K,b} \\ 2R\mu_{K,b}^u \end{pmatrix} \end{pmatrix} \\ &+ \big[\mu_{K,b}^\top Q\mu_{K,b} + (\mu_{K,b}^u)^\top R\mu_{K,b}^u + \mu^\top \overline{Q}\mu\big] \begin{pmatrix} \text{svec}(\Sigma_z) \\ \mathbf{0}_m \\ \mathbf{0}_k \end{pmatrix}, \end{aligned}$$

which concludes the proof of the lemma. □

B.7. Auxiliary Results

Lemma B.7.1. Assume that the random variable $w \sim \mathcal{N}(0, I)$, and let U and V be two symmetric matrices, then it holds that

$$\mathbb{E}(w^\top U w \cdot w^\top V w) = 2\,\mathrm{tr}(UV) + \mathrm{tr}(U) \cdot \mathrm{tr}(V).$$

Proof. See Magnus et al. (1978) and Magnus (1979) for a detailed proof. □

Lemma B.7.2. Let M, N be commuting symmetric matrices, and let $\alpha_1, \ldots, \alpha_n$, $\beta_1, \ldots, \beta_n$ denote their eigenvalues with $v_1, \ldots, v_n$ a common basis of orthogonal eigenvectors. Then the $n(n+1)/2$ eigenvalues of $M \otimes_s N$ are given by $(\alpha_i \beta_j + \alpha_j \beta_i)/2$, where $1 \le i \le j \le n$.

Proof. See Lemma 2 in Alizadeh et al. (1998) for a detailed proof. □

Lemma B.7.3. For any integer $m > 0$, let $A \in \mathbb{R}^{m \times m}$ and $\eta \sim \mathcal{N}(0, I_m)$. Then, there exists some absolute constant $C > 0$ such that for any $t \ge 0$, we have

$$\mathbb{P}\Big[\big|\eta^\top A \eta - \mathbb{E}(\eta^\top A \eta)\big| > t\Big] \le 2 \cdot \exp\Big[-C \cdot \min\big(t^2 \|A\|_{\mathrm{F}}^{-2},\ t\|A\|_2^{-1}\big)\Big].$$

Proof. See Rudelson et al. (2013) for a detailed proof. □

APPENDIX C

Supplemental Materials in Chapter 4

C.1. Analysis of Maximum Likelihood Estimation

We denote by

$$L_1(\Theta) = -\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\log\Theta(S_t, Z_t)\right]$$

the population counterpart of $\widehat{L}_1$. We define

$$H^2(\Theta_1, \Theta_2) = \frac{1}{2}\cdot\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\int\left(\sqrt{\Theta_1(S_t, z)} - \sqrt{\Theta_2(S_t, z)}\right)^2 \mathrm{d}z\right]. \tag{C.1.1}$$

We have the following supporting result.

Lemma C.1.1. Under Assumption (b), for any $\Theta_1, \Theta_2 \in \mathcal{F}_1$, it holds with probability at least $1-\delta$ for any $c/(NT)^2 \leq \delta \leq 1$ that

$$\begin{aligned}&\left|(L_1(\Theta_1) - L_1(\Theta_2)) - \left(\widehat{L}_1(\Theta_1) - \widehat{L}_1(\Theta_2)\right)\right| \\ &\qquad \leq c\cdot\frac{\log C_{\Theta^*}}{NT\kappa}\log\frac{1}{\delta}\log(NT) + c\cdot\sqrt{\frac{C_{\Theta^*}}{NT\kappa}H^2(\Theta_1, \Theta_2)\log\frac{1}{\delta}\log(NT)},\end{aligned}$$

where $H^2(\Theta_1, \Theta_2)$ is defined in (C.1.1).

Proof. By Theorem C.7.7, it holds with probability at least $1-\delta$ that

$$\begin{aligned}&\left|(L_1(\Theta_1)-L_1(\Theta_2))-\left(\widehat{L}_1(\Theta_1)-\widehat{L}_1(\Theta_2)\right)\right|\\ &\le c\cdot\frac{\log C_{\Theta^*}}{NT\kappa}\log\frac{1}{\delta}\log(NT)+c\cdot\sqrt{\frac{1}{NT\kappa}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\log\frac{\Theta_1(S_t,Z_t)}{\Theta_2(S_t,Z_t)}\right)^2\right]\log\frac{1}{\delta}\log(NT)}.\end{aligned}\tag{C.1.2}$$

Now, it suffices to upper bound the variance term on the RHS of the above inequality. Note that $\log x\le 2(\sqrt{x}-1)$ for any $x>0$. Thus, for any $s\in\mathcal{S}$, we have

$$\begin{aligned}&\int\Theta^*(s,z)\left(\log\frac{\Theta_1(s,z)}{\Theta_2(s,z)}\right)^2\mathrm{d}z\\ &\quad\le 4\int\Theta^*\max\left\{\left(\sqrt{\frac{\Theta_2}{\Theta_1}}-1\right)^2,\left(\sqrt{\frac{\Theta_1}{\Theta_2}}-1\right)^2\right\}\mathrm{d}z\\ &\quad=4\int\max\left\{\frac{\Theta^*}{\Theta_1}\left(\sqrt{\Theta_2}-\sqrt{\Theta_1}\right)^2,\frac{\Theta^*}{\Theta_2}\left(\sqrt{\Theta_1}-\sqrt{\Theta_2}\right)^2\right\}\mathrm{d}z\\ &\quad\le 4C_{\Theta^*}\int\left(\sqrt{\Theta_1(s,z)}-\sqrt{\Theta_2(s,z)}\right)^2\mathrm{d}z,\end{aligned}\tag{C.1.3}$$

which implies that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\log\frac{\Theta_1(S_t,Z_t)}{\Theta_2(S_t,Z_t)}\right)^2\right]\le 8C_{\Theta^*}H^2(\Theta_1,\Theta_2).\tag{C.1.4}$$

By plugging (C.1.4) into (C.1.2), we conclude the proof of the lemma. □

C.1.1. Proof of Theorem 4.4.5

Proof. Proof of the first statement. It suffices to show that with probability at least $1-\delta$, we have

$$\widehat{L}_1(\Theta^*) - \widehat{L}_1(\widehat{\Theta}) \leq \alpha_1.$$

By Corollary C.7.10, it holds with probability at least $1-\delta$ that

$$H^2(\Theta^*, \widehat{\Theta}) \leq c \cdot \frac{d}{NT\kappa} \log \frac{\theta_{\max}}{\delta}, \tag{C.1.5}$$

where $c > 0$ is an absolute constant, which may vary from lines to lines. Thus, by Lemma C.1.1, it holds with probability at least $1-\delta$ that

$$\begin{aligned}
&\Big| \Big(L_1(\Theta^*) - L_1(\widehat{\Theta}) \Big) - \Big(\widehat{L}_1(\Theta^*) - \widehat{L}_1(\widehat{\Theta}) \Big) \Big| \\
&\quad \leq c \cdot \frac{\log C_{\Theta^*}}{NT\kappa} d \log \frac{\theta_{\max}}{\delta} \log(NT) + c \cdot \sqrt{\frac{C_{\Theta^*}}{NT\kappa} H^2(\Theta^*, \widehat{\Theta}) d \log \frac{\theta_{\max}}{\delta} \log(NT)} \\
&\quad \leq c \cdot \frac{C_{\Theta^*} d}{NT\kappa} \log \frac{\theta_{\max}}{\delta} \log(NT), \qquad \text{(C.1.6)}
\end{aligned}$$

where we use a covering argument and (C.1.5) in the first and last inequalities, respectively. Further, by a similar idea as in (C.1.3), we upper bound $|L_1(\Theta^*) - L_1(\widehat{\Theta})|$ as follows,

$$\begin{aligned}
|L_1(\Theta^*) - L_1(\widehat{\Theta})| &= \left| \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \int \Theta^*(S_t, z) \log \frac{\Theta^*(S_t, z)}{\widehat{\Theta}(S_t, z)} \mathrm{d}z \right] \right| \\
&\leq 2C_{\Theta^*} H^2(\Theta^*, \widehat{\Theta}) \leq c \cdot \frac{C_{\Theta^*} d}{NT\kappa} \log \frac{\theta_{\max}}{\delta}, \qquad \text{(C.1.7)}
\end{aligned}$$

where we use (C.1.5) and Corollary C.7.10 in the first and last inequalities, respectively. Now, by combining (C.1.6) and (C.1.7), it holds with probability at least $1-\delta$ that

$$\widehat{L}_1(\Theta^*) - \widehat{L}_1(\widehat{\Theta}) \leq c \cdot \frac{C_{\Theta^*} d}{NT\kappa} \log \frac{\theta_{\max}}{\delta} \log(NT) = \alpha_1,$$

which concludes the proof of the first statement.

Proof of the second statement. By Lemma C.1.1, with probability at least $1-\delta$, for any $\Theta \in \mathsf{conf}^1_{\alpha_1}$, we have

$$\begin{aligned} &\left|(L_1(\Theta^*) - L_1(\Theta)) - \left(\widehat{L}_1(\Theta^*) - \widehat{L}_1(\Theta)\right)\right| \\ &\leq c \cdot \frac{\log C_{\Theta^*}}{NT\kappa} d \log \frac{\theta_{\max}}{\delta} \log(NT) + c \cdot \sqrt{\frac{C_{\Theta^*}}{NT\kappa} H^2(\Theta^*, \Theta) d \log \frac{\theta_{\max}}{\delta} \log(NT)}, \end{aligned} \tag{C.1.8}$$

where we use a covering argument. In the meanwhile, by the first statement, we have $\Theta^* \in \mathsf{conf}^1_{\alpha_1}$ with probability at least $1-\delta$. Thus, we have

$$\left|\widehat{L}_1(\Theta^*) - \widehat{L}_1(\Theta)\right| \leq \left|\widehat{L}_1(\Theta^*) - \widehat{L}_1(\widehat{\Theta})\right| + \left|\widehat{L}_1(\widehat{\Theta}) - \widehat{L}_1(\Theta)\right| \leq 2\alpha_1, \tag{C.1.9}$$

where we use the fact that $\Theta \in \mathsf{conf}^1_{\alpha_1}$. By combining (C.1.8) and (C.1.9), with probability at least $1-\delta$, it holds for any $\Theta \in \mathsf{conf}^1_{\alpha_1}$ that

$$\begin{aligned} &L_1(\Theta) - L_1(\Theta^*) \\ &\quad \leq c \cdot \frac{\log C_{\Theta^*}}{NT\kappa} d \log \frac{\theta_{\max}}{\delta} \log(NT) + c \cdot \sqrt{\frac{C_{\Theta^*}}{NT\kappa} H^2(\Theta^*, \Theta) d \log \frac{\theta_{\max}}{\delta} \log(NT)}. \end{aligned} \tag{C.1.10}$$

On the other hand, it holds for any $s \in \mathcal{S}$ that

$$\begin{aligned}
-\int \Theta^*(s,z)\log\frac{\Theta(s,z)}{\Theta^*(s,z)}\mathrm{d}z &\geq -2\int \Theta^*(s,z)\left(\sqrt{\frac{\Theta(s,z)}{\Theta^*(s,z)}}-1\right)\mathrm{d}z \\
&= \int \left(\Theta^*(s,z)+\Theta(s,z)-2\sqrt{\Theta(s,z)\Theta^*(s,z)}\right)\mathrm{d}z \\
&= \int \left(\sqrt{\Theta^*(s,z)}+\sqrt{\Theta(s,z)}\right)^2\mathrm{d}z,
\end{aligned}$$

which implies that

$$L_1(\Theta) - L_1(\Theta^*) \geq 2H^2(\Theta^*, \Theta). \tag{C.1.11}$$

By combining (C.1.10) and (C.1.11), we have

$$H^2(\Theta^*,\Theta) \leq c\cdot\frac{\log C_{\Theta^*}}{NT\kappa}d\log\frac{\theta_{\max}}{\delta}\log(NT) + c\cdot\sqrt{\frac{C_{\Theta^*}}{NT\kappa}H^2(\Theta^*,\Theta)d\log\frac{\theta_{\max}}{\delta}\log(NT)},$$

which implies that

$$H^2(\Theta^*,\Theta) \leq c\cdot\frac{C_{\Theta^*}d}{NT\kappa}\log\frac{\theta_{\max}}{\delta}\log(NT).$$

Now, by Lemma C.7.8, with probability at least $1-\delta$, it holds for any $\Theta \in \mathsf{conf}^1_{\alpha_1}$ that

$$\sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\Theta(S_t,\cdot)-\Theta^*(S_t,\cdot)\|_1^2\right]} \leq c\cdot\sqrt{\frac{C_{\Theta^*}d}{NT\kappa}\log\frac{\theta_{\max}}{\delta}\log(NT)},$$

which concludes the proof of the second statement. □

C.2. Proofs of Results in §4.2

We provide proofs of results in §4.2. We first present proofs for §4.2.2, then we present proofs for §4.2.1.

C.2.1. Proof of Lemma 4.2.7

Proof. By Assumption (b), we know that

$$\mathbb{E}\left[\frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} w^\pi(S_t) R_t \,\middle|\, S_t\right]$$

is well-defined. Further, we observe for any $t \in \{0, 1, \ldots, T-1\}$ that

$$\begin{aligned}
&\mathbb{E}\left[\frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} w^\pi(S_t) R_t \,\middle|\, S_t\right] \\
&\quad = \sum_{a \in \mathcal{A}} \mathbb{E}\left[\frac{Z_t^\top a \pi(a \mid S_t) d^\pi(S_t)\,\mathbb{1}\{A_t = a\}}{\Delta^*(S_t, a)\Theta^*(S_t, Z_t) d^b(S_t)} R(S_t, U_t, a, S_{t+1}, U_{t+1}) \,\middle|\, S_t\right] \\
&\quad = \sum_{a \in \mathcal{A}} \mathbb{E}\left[\frac{Z_t^\top a \pi(a \mid S_t) d^\pi(S_t)\,\mathbb{1}\{A_t = a\}}{\Delta^*(S_t, a)\Theta^*(S_t, Z_t) d^b(S_t)} r(S_t, U_t, a) \,\middle|\, S_t\right] \\
&\quad = \sum_{a \in \mathcal{A}} \mathbb{E}\left[\frac{Z_t^\top a \pi(a \mid S_t) d^\pi(S_t) \mathbb{P}(A_t = a \mid S_t, U_t, Z_t)}{\Delta^*(S_t, a) d^b(S_t)\Theta^*(S, Z)} r(S_t, U_t, a) \,\middle|\, S_t\right] \\
&\quad = \sum_{a \in \mathcal{A}} \mathbb{E}\left[\frac{\pi(a \mid S_t) d^\pi(S_t) \mathbb{P}(A_t = a \mid S_t, U_t, Z_t = a)}{\Delta^*(S_t, a) d^b(S_t)} r(S_t, U_t, a) \,\middle|\, S_t\right] \\
&\qquad - \sum_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}, z \neq a} \frac{1}{K-1} \mathbb{E}\left[\frac{\pi(a \mid S_t) d^\pi(S_t) \mathbb{P}(A_t = a \mid S_t, U_t, Z_t = z)}{\Delta^*(S_t, a) d^b(S_t)} r(S_t, U_t, a) \,\middle|\, S_t\right] \\
&\quad = \sum_{a \in \mathcal{A}} \frac{\pi(a \mid S_t) d^\pi(S_t)}{d^b(S_t)} \mathbb{E}_{U_t}[r(S_t, U_t, a) \mid S_t],
\end{aligned}$$

where in the first equality, we use the definition of $w^\pi(s)$; in the second equality, we use Assumption (a); in the third equality, we use Assumption (b); in the forth equality, we use

Assumption (c); while in the fifth equality, we use Assumption (d). Now, by Assumption 4.2.6, we know that $\mathbb{E}_{U_t}[r(S_t, U_t, a) \mid S_t] = \widetilde{r}(S_t, a)$, where the function $\widetilde{r}$ is independent of t. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} w^\pi(S_t) R_t\right] \\
&\quad = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\sum_{a\in\mathcal{A}}\frac{\pi(a \mid S_t) d^\pi(S_t)}{d^b(S_t)}\widetilde{r}(S_t, a)\right] \\
&\quad = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{a\in\mathcal{A}}\int \frac{\pi(a \mid s) d^\pi(s)}{d^b(s)}\widetilde{r}(s, a) p_t^b(s) \mathrm{d}s \\
&\quad = \sum_{a\in\mathcal{A}}\int \pi(a \mid s) d^\pi(s) \widetilde{r}(s, a) \mathrm{d}s = J(\pi),
\end{aligned}$$

which concludes the proof of the lemma. □

C.2.2. Proof of Lemma 4.2.8

Proof. Similar to the proof of Lemma 4.2.7 in §C.2.1, we observe that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^\top A_t d^\pi(S_t)\pi(A_t \mid S_t)}{\Delta(S_t, A_t) d^b(S_t) P(Z_t \mid S_t)} f(S_t)\right] = \int f(s') d^\pi(s') \mathrm{d}s'. \tag{C.2.1}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^\top A_t d^\pi(S_t)\pi(A_t \mid S_t)}{\Delta(S_t, A_t) d^b(S_t) P(Z_t \mid S_t)} f(S_{t+1})\right] \\
&\quad = \int f(s') d^\pi(s) \pi(a \mid s) \mathbb{P}(S' = s \mid S = s, A = a) \mathrm{d}s' \mathrm{d}a \mathrm{d}s.
\end{aligned} \tag{C.2.2}$$

Meanwhile, by the definition of $d^\pi(s,a)$, we have

$$\begin{aligned}
d^\pi(s') &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_t^\pi(s') \\
&= (1-\gamma)\nu(s') + (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t+1}p_{t+1}^\pi(s') \\
&= (1-\gamma)\left(\nu(s') + \gamma\sum_{t=0}^{\infty}\gamma^t\int \mathbb{P}(S_{t+1}=s' \,|\, S_t=s, A_t=a)\pi(a\,|\,s)p_t^\pi(s)\mathrm{d}s\mathrm{d}a\right) \\
&= (1-\gamma)\nu(s') + \gamma\int \mathbb{P}(S'=s' \,|\, S=s, A=a)\pi(a\,|\,s)d^\pi(s)\mathrm{d}s\mathrm{d}a, \qquad \text{(C.2.3)}
\end{aligned}$$

where we use the assumption that $S_{t+1}\,|\,(S_t,A_t)$ is time-homogeneous. Combining (C.2.1) and (C.2.2), we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^\top A_t d^\pi(S_t)\pi(A_t\,|\,S_t)}{\Delta(S_t,A_t)d^b(S_t)P(Z_t\,|\,S_t)}\left(f(S_t)-\gamma f(S_{t+1})\right)\right] \\
&\quad = \int f(s')\left(d^\pi(s') - \gamma\int d^\pi(s)\pi(a\,|\,s)\mathbb{P}(S'=s' \,|\, S=s, A=a)\mathrm{d}s\mathrm{d}a\right)\mathrm{d}s' \\
&\quad = (1-\gamma)\int f(s')\nu(s')\mathrm{d}s' = (1-\gamma)\mathbb{E}_{S\sim\nu}\left[f(S)\right],
\end{aligned}$$

where we use (C.2.3) in the forth equality. This concludes the proof of the lemma. □

C.2.3. Proof of Lemma 4.2.4

Proof. Similar to the proof of Lemma 4.2.7 in §C.2.1, we observe that

$$\begin{aligned}
&\mathbb{E}\left[\frac{Z_0^\top A_0\pi(A_0\,|\,S_0)}{\Delta^*(S_0,A_0)P(Z_0\,|\,S_0)}R_0 \,\middle|\, S_0=s\right] \\
&\qquad = \mathbb{E}_{U_0}\left[\sum_{a\in\mathcal{A}}\pi(a\,|\,S_0)r(S_0,U_0,a)\,\middle|\, S_0=s\right] = \mathbb{E}_\pi[R_0\,|\,S_0=s], \qquad \text{(C.2.4)}
\end{aligned}$$

where $\mathbb{E}_\pi[\cdot]$ denotes that the expectation is taken w.r.t. $A_0 \sim \pi(\cdot \,|\, S_0)$. Similarly, by Assumption 4.2.3, we observe that

$$\begin{aligned}
&\mathbb{E}\left[\frac{Z_0^\top A_0 \pi(A_0 \,|\, S_0)}{\Delta^*(S_0, A_0) P(Z_0 \,|\, S_0)} \frac{Z_1^\top A_1 \pi(A_1 \,|\, S_1)}{\Delta^*(S_1, A_1) P(Z_1 \,|\, S_1)} R_1 \,\middle|\, S_0 = s\right] \\
&\quad = \mathbb{E}\left[\sum_{a \in \mathcal{A}} \pi(a \,|\, S_0) \mathbb{E}\left[\frac{Z_1^\top A_1 \pi(A_1 \,|\, S_1)}{\Delta^*(S_1, A_1) P(Z_1 \,|\, S_1)} R_1 \,\middle|\, S_0, U_0, A_0 = a\right] \,\middle|\, S_0 = s\right] \\
&\quad = \mathbb{E}\left[\sum_{a \in \mathcal{A}} \pi(a \,|\, S_0) \mathbb{E}\left[\mathbb{E}\left[\frac{Z_1^\top A_1 \pi(A_1 \,|\, S_1)}{\Delta^*(S_1, A_1) P(Z_1 \,|\, S_1)} R_1 \,\middle|\, S_1, U_1\right] \,\middle|\, S_0, U_0, A_0 = a\right] \,\middle|\, S_0 = s\right] \\
&\quad = \mathbb{E}\left[\sum_{a \in \mathcal{A}} \pi(a \,|\, S_0) \mathbb{E}\left[\mathbb{E}_\pi[R_1 \,|\, S_1, U_1] \,|\, S_0, U_0, A_0 = a\right] \,\middle|\, S_0 = s\right] = \mathbb{E}_\pi[R_1 \,|\, S_0 = s].
\end{aligned}$$

Now, by induction, it holds for any $t \geq 0$ that

$$\mathbb{E}\left[R_t \left(\prod_{j=0}^{t} \frac{Z_j^\top A_j \pi(A_j \,|\, S_j)}{\Delta^*(S_j, A_j) P(Z_j \,|\, S_j)}\right) \,\middle|\, S_0 = s\right] = \mathbb{E}_\pi[R_t \,|\, S_0 = s],$$

which implies that

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \left(\prod_{j=0}^{t} \frac{Z_j^\top A_j \pi(A_j \,|\, S_j)}{\Delta^*(S_j, A_j) P(Z_j \,|\, S_j)}\right) \,\middle|\, S_0 = s\right].$$

To show that the IV-aided Bellman equation holds, by a similar argument as in (C.2.4), we observe that

$$\begin{aligned}
&\mathbb{E}\left[\frac{Z_0^\top A_0 \pi(A_0 \,|\, S_0)}{\Delta^*(S_0, A_0) P(Z_0 \,|\, S_0)} \gamma V^\pi(S_1) \,\Big|\, S_0 = s\right] \\
&= \mathbb{E}\left[\frac{Z_0^\top A_0 \pi(A_0 \,|\, S_0)}{\Delta^*(S_0, A_0) P(Z_0 \,|\, S_0)} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t+1} R_{t+1} \left(\prod_{j=1}^{t+1} \frac{Z_j^\top A_j \pi(A_j \,|\, S_j)}{\Delta^*(S_j, A_j) P(Z_j \,|\, S_j)}\right) \,\Big|\, S_1\right] \,\Big|\, S_0 = s\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R_t \left(\prod_{j=0}^{t} \frac{Z_j^\top A_j \pi(A_j \,|\, S_j)}{\Delta^*(S_j, A_j) P(Z_j \,|\, S_j)}\right) \,\Big|\, S_0 = s\right].
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{Z_0^\top A_0 \pi(A_0 \,|\, S_0)}{\Delta^*(S_0, A_0) P(Z_0 \,|\, S_0)} (R_0 + \gamma V^\pi(S_1)) \,\Big|\, S_0 = s\right] \\
&\quad = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t+1} R_{t+1} \left(\prod_{j=0}^{t+1} \frac{Z_j^\top A_j \pi(A_j \,|\, S_j)}{\Delta^*(S_j, A_j) P(Z_j \,|\, S_j)}\right) \,\Big|\, S_0 = s\right] = V^\pi(s).
\end{aligned}$$

Similar results also hold for any $t \geq 0$. This concludes the proof of the lemma. □

C.2.4. Proof of Corollary 4.2.5

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[f(S_t) \frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t) \Theta^*(S_t, Z_t)} \left(R_t + \gamma V^\pi(S_{t+1})\right)\right] \\
&\quad = \mathbb{E}\left[f(S_t) \mathbb{E}\left[\frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t) \Theta^*(S_t, Z_t)} \left(R_t + \gamma V^\pi(S_{t+1})\right) \,\Big|\, S_t\right]\right] \\
&\quad = \mathbb{E}\left[f(S_t) V^\pi(S_t)\right],
\end{aligned}$$

where the last equality comes from Lemma 4.2.4. By summing all $t \in \{0, 1, \ldots, T-1\}$, we conclude the proof of the corollary. □

C.3. Proofs of Results in §4.4.1

C.3.1. Proof of Theorem 4.4.9

Proof. By the definition of $J(\pi)$ in (4.1.2), we proceed as follows,

$$
\begin{aligned}
&J(\pi^*) - J(\widehat{\pi}_{\mathsf{vf}}) \\
&\quad = (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0) - V^{\widehat{\pi}_{\mathsf{vf}}}(S_0)\right] \\
&\quad \leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right] - \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\widehat{\pi})} (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right]
\end{aligned}
$$

(C.3.1)

$$
\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right] - \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right],
$$

where in the first inequality, we use Lemma 4.4.7; while in the last inequality, we use the optimality of $\widehat{\pi}_{\mathsf{vf}}$. It suffices to characterize the RHS of the above. We proceed (C.3.1) as follows,

$$
J(\pi^*) - J(\widehat{\pi}_{\mathsf{vf}})
$$

(C.3.2)

$$
\leq \max_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} \left|(1-\gamma)\mathbb{E}_{S_0\sim\nu}[v(S_0)] - (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right]\right|.
$$

Meanwhile, by Lemmas 4.2.7 and 4.2.8, we have

$$(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right] = J(\pi^*) = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} w^{\pi^*}(S_t)\frac{Z_t^\top A_t\pi^*(A_t\,|\,S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}R_t\right],$$

(C.3.3)

$$(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} w^{\pi^*}(S_t)\frac{Z_t^\top A_t\pi^*(A_t\,|\,S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}\left(v(S_t)-\gamma v(S_{t+1})\right)\right].$$

Now, by plugging (C.3.3) into the RHS of (C.3.2), we obtain

$$J(\pi^*) - J(\widehat{\pi}_{\mathsf{vf}}) \le \max_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\ \max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)}\left|\Phi^{\pi^*}_{\mathsf{vf}}(v, w^{\pi^*};\Delta^*,\Theta^*)\right|.$$

By continuing the above inequality, we have

$$\begin{aligned}
&J(\pi^*) - J(\widehat{\pi}_{\mathsf{vf}})\\
&\quad\le \max_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\ \max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)}\ \max_{g\in\mathcal{W}}\left|\Phi^{\pi^*}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*)\right|\\
&\quad= \max_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\ \max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)}\ \max_{g\in\mathcal{W}}\max\left\{\Phi^{\pi^*}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*), -\Phi^{\pi^*}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*)\right\}\\
&\quad= \max_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\ \max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)}\ \max_{g\in\mathcal{W}}\max\left\{\Phi^{\pi^*}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*), \Phi^{\pi^*}_{\mathsf{vf}}(v,-g;\Delta^*,\Theta^*)\right\}\\
&\quad= \max_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\ \max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)}\ \max_{g\in\mathcal{W}}\Phi^{\pi^*}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*)\\
&\quad\le c\cdot\frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)L_\Pi\sqrt{\frac{1}{NT\kappa}\cdot\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\cdot\log\frac{1}{\delta}\log(NT)},
\end{aligned}$$

where in the first inequality, we use Assumption 4.4.6; in the third equality, we use the fact that $\mathcal{W}$ is symmetric; while in the last inequality, we use Lemma 4.4.8. This concludes the proof of the theorem. □

C.3.2. Proof of Lemma 4.4.7

Proof. By Assumption 4.4.6, we know that $V^\pi \in \mathcal{V}$. Thus, to show that $V^\pi \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi)$ with a high probability, it suffices to show that

$$\max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(V^\pi, g; \Delta^*, \Theta^*) - \max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(\widehat{v}^\pi_{\Delta^*,\Theta^*}, g; \Delta^*, \Theta^*) \leq \alpha_{\mathsf{vf}}. \tag{C.3.4}$$

In the follows, we show that (C.3.4) holds with a high probability. For the simplicity of notations, we denote by $\Phi^\pi_{\mathsf{vf}}(v, g; *) = \Phi^\pi_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*)$ and $\widehat{v}^\pi_* = \widehat{v}^\pi_{\Delta^*,\Theta^*}$ for any (π, v, g). Note that

$$\begin{aligned}
&\max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(V^\pi, g; *) - \max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(\widehat{v}^\pi_*, g; *) \\
&\quad = \max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(V^\pi, g; *) - \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(V^\pi, g; *) + \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(V^\pi, g; *) - \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(\widehat{v}^\pi_*, g; *) \\
&\qquad + \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(\widehat{v}^\pi_*, g; *) - \max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(\widehat{v}^\pi_*, g; *) \\
&\quad \leq \max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(V^\pi, g; *) - \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(V^\pi, g; *) + \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(\widehat{v}^\pi_*, g; *) - \max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(\widehat{v}^\pi_*, g; *) \\
&\quad \leq 2 \max_{v\in\mathcal{V}} \left| \max_{g\in\mathcal{W}} \widehat{\Phi}^\pi_{\mathsf{vf}}(v, g; *) - \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(v, g; *) \right| \\
&\quad \leq 2 \max_{v\in\mathcal{V}} \max_{g\in\mathcal{W}} \left| \widehat{\Phi}^\pi_{\mathsf{vf}}(v, g; *) - \Phi^\pi_{\mathsf{vf}}(v, g; *) \right|,
\end{aligned} \tag{C.3.5}$$

where in the first inequality, we use $\max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(V^\pi, g; *) = 0$ while $\max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(v, g; *) \geq 0$ for any v. In the meanwhile, by Theorem C.7.6, with probability at least $1 - \delta$, it holds for any $(\pi, v, g) \in \Pi \times \mathcal{V} \times \mathcal{W}$ that

$$\left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; *) - \Phi^{\pi}_{\mathsf{vf}}(v, g; *)\right| \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{W},\mathcal{V},\Pi}}{NT\kappa} \cdot \log\frac{1}{\delta} \log(NT)}, \tag{C.3.6}$$

where we use Assumption (b) and $\|g\|_\infty \le C_*$ for any $g \in \mathcal{W}$. Now, combining (C.3.5) and (C.3.6), with probability at least $1-\delta$, we have

$$\max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(V^\pi, g; *) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_*, g; *) \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{W},\mathcal{V},\Pi}}{NT\kappa} \cdot \log\frac{1}{\delta} \log(NT)} = \alpha_{\mathsf{vf}},$$

which implies that $V^\pi \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi)$ for any $\pi \in \Pi$. This concludes the proof of the lemma. □

C.3.3. Proof of Lemma 4.4.8

Proof. Since $v \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi)$, there exists a pair $(\widetilde{\Delta}, \widetilde{\Theta}) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ such that $v \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$. For the simplicity of notations, we denote by

$$\widetilde{v} \in \operatorname*{argmin}_{v\in\mathcal{V}} \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta}), \tag{C.3.7}$$

i.e., $\widetilde{v} = \widehat{v}^{\pi}_{\widetilde{\Delta},\widetilde{\Theta}}$, which is defined in (4.3.2). By the definition of $\widetilde{v}$ and $v \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$, we know that

$$\max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta}) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widetilde{v}, g; \widetilde{\Delta}, \widetilde{\Theta}) \le \alpha_{\mathsf{vf}}. \tag{C.3.8}$$

Note that

$$
\begin{aligned}
&\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*) \\
&\quad= \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*) + \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*) \\
&\qquad - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\widetilde{\Delta},\widetilde{\Theta}) + \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\widetilde{\Delta},\widetilde{\Theta}) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widetilde{v},g;\widetilde{\Delta},\widetilde{\Theta}) \\
&\qquad + \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widetilde{v},g;\widetilde{\Delta},\widetilde{\Theta}) - \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v},g;\widetilde{\Delta},\widetilde{\Theta}) + \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v},g;\widetilde{\Delta},\widetilde{\Theta}) \\
&\quad\le 2\underbrace{\max_{(v,g,\Delta,\Theta)\in(\mathcal{V},\mathcal{W},\mathcal{F}_0,\mathcal{F}_1)} \left|\Phi^{\pi}_{\mathsf{vf}}(v,g;\Delta,\Theta) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\Delta,\Theta)\right|}_{\text{Term (I)}} + \underbrace{\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v},g;\widetilde{\Delta},\widetilde{\Theta})}_{\text{Term (II)}} \\
&\qquad + \underbrace{\max_{g\in\mathcal{W}} \left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\widetilde{\Delta},\widetilde{\Theta})\right|}_{\text{Term (III)}} + \alpha_{\mathsf{vf}},
\end{aligned}
\tag{C.3.9}
$$

where we use (C.3.8) in the last inequality. Now we upper bound terms (I), (II), and (III) on the RHS of (C.3.9).

Upper Bounding Term (I). By Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(v,g,\Delta,\Theta,\pi)\in(\mathcal{V},\mathcal{W},\mathcal{F}_0,\mathcal{F}_1,\Pi)$ that

$$
\left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\Delta,\Theta) - \Phi^{\pi}_{\mathsf{vf}}(v,g;\Delta,\Theta)\right| \le c\cdot\frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}}{NT\kappa}\log\frac{1}{\delta}\log(NT)},
$$

which implies that with probability at least $1-\delta$, we have

$$
\text{Term (I)} \le c\cdot\frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}}{NT\kappa}\log\frac{1}{\delta}\log(NT)}. \tag{C.3.10}
$$

Upper Bounding Term (II). We introduce the following lemma to help upper bound term (II).

Lemma C.3.1. Suppose α_0 and α_1 are defined in Assumption 4.4.4. With probability at least $1-\delta$, for any $(\Delta, \Theta, \pi) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1} \times \Pi$, we have

$$\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{\Delta,\Theta}, g; \Delta, \Theta) \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0+\xi_1) L_\Pi \sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}}{NT\kappa} \log\frac{1}{\delta} \log(NT)},$$

where $\widehat{v}^{\pi}_{\Delta,\Theta}$ is defined in (4.3.2), ξ_0 and ξ_1 are constants defined in Assumption 4.4.4.

Proof. See §C.3.4 for a detailed proof. □

By the definition of $\widetilde{v}$ in (C.3.7) and Lemma C.3.1, with probability at least $1-\delta$, we have

$$\text{Term (II)} \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0+\xi_1) L_\Pi \sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}}{NT\kappa} \log\frac{1}{\delta} \log(NT)}. \tag{C.3.11}$$

Upper Bounding Term (III). Note that

$$\begin{aligned} &\left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta})\right| \\ &\leq \left|\left(\widehat{\mathbb{E}} - \mathbb{E}\right)\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(R_t + \gamma v(S_{t+1}))\right]\right| \\ &\quad + \left|\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(R_t + \gamma v(S_{t+1}))\right]\right|. \end{aligned} \tag{C.3.12}$$

For the first term on the RHS of (C.3.12), by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(v, g, \pi) \in \mathcal{V} \times \mathcal{W} \times \Pi$ that

$$\left|\left(\widehat{\mathbb{E}}-\mathbb{E}\right)\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(R_t + \gamma v(S_{t+1}))\right]\right|$$

(C.3.13)

$$\leq c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi}}{NT\kappa} \log\frac{1}{\delta} \log(NT)}.$$

For the second term on the RHS of (C.3.12), with probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)} - \frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\widetilde{\Delta}(S_t,A_t)\widetilde{\Theta}(S_t,Z_t)}\right)(R_t+\gamma v(S_{t+1}))\right]\right| \\
&\quad \le \frac{C_*}{1-\gamma}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left|\frac{1}{\widetilde{\Delta}(S_t,A_t)\widetilde{\Theta}(S_t,Z_t)} - \frac{1}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}\right|\right] \\
&\quad = \frac{C_*}{1-\gamma}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left|\frac{\Theta^*(S_t,Z_t)-\widetilde{\Theta}(S_t,Z_t)}{\widetilde{\Delta}(S_t,A_t)\widetilde{\Theta}(S_t,Z_t)\Theta^*(S_t,Z_t)} - \frac{\Delta^*(S_t,A_t)-\widetilde{\Delta}(S_t,A_t)}{\Delta^*(S_t,A_t)\widetilde{\Delta}(S_t,A_t)\Theta^*(S_t,Z_t)}\right|\right] \\
&\quad \le \frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\left(C_{\Theta^*}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\|\Delta^*(S_t,\cdot)-\widetilde{\Delta}(S_t,\cdot)\right\|_1\right]\right. \\
&\qquad \left. + C_{\Delta^*}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\|\Theta^*(S_t,\cdot)-\widetilde{\Theta}(S_t,\cdot)\right\|_1\right]\right) \\
&\quad \le \frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\left(C_{\Theta^*}\sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\|\Delta^*(S_t,\cdot)-\widetilde{\Delta}(S_t,\cdot)\right\|_1^2\right]}\right. \\
&\qquad \left. + C_{\Delta^*}\sqrt{\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\|\Theta^*(S_t,\cdot)-\widetilde{\Theta}(S_t,\cdot)\right\|_1^2\right]}\right)
\end{aligned}$$

(C.3.14)

$$\le \frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\left(\xi_0 C_{\Theta^*}\sqrt{\frac{C_{\Delta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0}\log\frac{1}{\delta}\log(NT)} + \xi_1 C_{\Delta^*}\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_1}\log\frac{1}{\delta}\log(NT)}\right),$$

where in the first inequality, we use the fact that $\|v\|_\infty \le 1/(1-\gamma)$ and $\|g\|_\infty \le C_*$; in the third inequality, we use Cauchy-Schwarz inequality; while in the last inequality, we use Assumption 4.4.4 with the fact that $(\widetilde{\Delta},\widetilde{\Theta}) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$. Now, by plugging (C.3.13) and (C.3.14) into (C.3.12), with probability at least $1-\delta$, it holds for any

$v \in \cup_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0}\times \mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi)$, $g \in \mathcal{W}$, and $\pi \in \Pi$ that

$$\begin{aligned}&\left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\widetilde{\Delta},\widetilde{\Theta})\right| \\ &\qquad \le c\cdot \frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\cdot \log\frac{1}{\delta}\log(NT)}. \end{aligned} \tag{C.3.15}$$

Now, by plugging (C.3.10), (C.3.11), and (C.3.15) into (C.3.9), with probability at least $1-\delta$, it holds for any $v \in \cup_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0}\times \mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi)$ and $\pi \in \Pi$ that

$$\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v,g;\Delta^*,\Theta^*) \le c\cdot \frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)L_{\Pi}\sqrt{\frac{1}{NT\kappa}\cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\cdot \log\frac{1}{\delta}\log(NT)},$$

which concludes the proof of the lemma. □

C.3.4. Proof of Lemma C.3.1

Proof. Note that

$$\begin{aligned}&\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{\Delta,\Theta},g;\Delta,\Theta) \\ &\quad = \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{\Delta,\Theta},g;\Delta,\Theta) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{\Delta,\Theta},g;\Delta,\Theta) + \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{\Delta,\Theta},g;\Delta,\Theta) \\ &\qquad - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(V^{\pi},g;\Delta,\Theta) + \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(V^{\pi},g;\Delta,\Theta) - \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(V^{\pi},g;\Delta,\Theta) \\ &\qquad + \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(V^{\pi},g;\Delta,\Theta) - \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(V^{\pi},g;\Delta^*,\Theta^*) \\ &\quad \le 2\max_{v\in\mathcal{V}}\max_{g\in\mathcal{W}}\left|\Phi^{\pi}_{\mathsf{vf}}(v,g;\Delta,\Theta) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v,g;\Delta,\Theta)\right| \\ &\qquad + \max_{g\in\mathcal{W}}\left|\Phi^{\pi}_{\mathsf{vf}}(V^{\pi},g;\Delta,\Theta) - \Phi^{\pi}_{\mathsf{vf}}(V^{\pi},g;\Delta^*,\Theta^*)\right|, \end{aligned} \tag{C.3.16}$$

where we use the fact that $\widehat{v}^{\pi}_{\Delta,\Theta} \in \operatorname{argmin}_{v\in\mathcal{V}} \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta)$ in the last inequality. In the meanwhile, by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(v, g, \pi) \in \mathcal{V}\times\mathcal{W}\times\Pi$ that

$$\left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta) - \Phi^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta)\right| \le c \cdot \frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\sqrt{\frac{\mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi}}{NT\kappa}\log\frac{1}{\delta}\log(NT)}. \tag{C.3.17}$$

Also, we upper bound the second term on the RHS of (C.3.16) with probability at least $1-\delta$ by a similar argument as in (C.3.14),

$$\begin{aligned}&|\Phi^{\pi}_{\mathsf{vf}}(V^{\pi}, g; \Delta, \Theta) - \Phi^{\pi}_{\mathsf{vf}}(V^{\pi}, g; \Delta^*, \Theta^*)|\\ &\le \frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\left(\xi_0 C_{\Theta^*}\sqrt{\frac{C_{\Delta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0}\log\frac{1}{\delta}\log(NT)} + \xi_1 C_{\Delta^*}\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_1}\log\frac{1}{\delta}\log(NT)}\right),\end{aligned} \tag{C.3.18}$$

where in the first inequality, we use the fact that $\|V^{\pi}\|_\infty \le 1/(1-\gamma)$ and $\|g\|_\infty \le C_*$; in the third inequality, we use Cauchy Schwarz inequality; while in the last inequality, we use Assumption 4.4.4 with $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$. Now, by plugging (C.3.17) and (C.3.18) into (C.3.16), with probability at least $1-\delta$, it holds for any $(\Delta, \Theta, \pi) \in \mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}\times\Pi$ that

$$\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{\Delta,\Theta}, g; \Delta, \Theta) \le c \cdot \frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)L_{\Pi}\sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}}{NT\kappa}\log\frac{1}{\delta}\log(NT)},$$

which concludes the proof of the lemma. □

C.4. Proofs of Results in §4.4.2

C.4.1. Proof of Theorem 4.4.13

Proof. Before the proof of the theorem, we first introduce some supporting results as follows. We define the population counterpart of $\widehat{L}_{\mathrm{mis}}(w, \pi; \Delta, \Theta)$ as

$$L_{\mathrm{mis}}(w, \pi; \Delta, \Theta) = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \frac{Z_t^\top A_t \pi(A_t \mid S_t)}{\Delta(S_t, A_t)\Theta(S_t, Z_t)} w(S_t) R_t\right]$$

for any (w, π, Δ, Θ).

Lemma C.4.1. It holds for any $(\pi, w) \in \Pi \times \mathcal{W}$ that

$$L_{\mathrm{mis}}(w^\pi, \pi; \Delta^*, \Theta^*) - L_{\mathrm{mis}}(w, \pi; \Delta^*, \Theta^*) = \Phi^\pi_{\mathrm{mis}}(w, V^\pi; \Delta^*, \Theta^*),$$

where V^π is the state-value function defined in (4.1.2).

Proof. See §C.4.4 for a detailed proof. □

Lemma C.4.2. Suppose that $(\alpha_0, \alpha_1, \alpha_{\mathrm{mis}})$ is defined in Lemmas 4.4.11. With probability at least $1 - \delta$, it holds for any $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ that

$$\left| \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta^*, \Theta^*) - \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta, \Theta) \right|$$
$$\leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}} \log\frac{1}{\delta}\log(NT)} = \varepsilon^*_L.$$

Proof. See §C.4.5 for a detailed proof. □

Lemma C.4.3. With probability at least $1-\delta$, it holds for any $(w, \Delta, \Theta, \pi) \in \mathcal{W} \times \mathcal{F}_0 \times \mathcal{F}_1 \times \Pi$ that

$$\begin{aligned}&\Big| L_{\mathrm{mis}}(w, \pi; \Delta, \Theta) - \widehat{L}_{\mathrm{mis}}(w, \pi; \Delta, \Theta) \Big| \\ &\qquad \leq c \cdot C_{\Delta^*} C_{\Theta^*} C_* \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \Pi} \log \frac{1}{\delta} \log(NT)} = \widehat{\varepsilon}_L.\end{aligned}$$

Proof. See §C.4.6 for a detailed proof. □

Now we start the proof of the theorem. By the definition of $J(\pi)$, it holds with probability at least $1-\delta$ that

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}_{\mathrm{mis}}) &= L_{\mathrm{mis}}(w^{\pi^*}, \pi^*; \Delta^*, \Theta^*) - L_{\mathrm{mis}}(w^{\widehat{\pi}_{\mathrm{mis}}}, \widehat{\pi}_{\mathrm{mis}}; \Delta^*, \Theta^*) \\ &\leq L_{\mathrm{mis}}(w^{\pi^*}, \pi^*; \Delta^*, \Theta^*) - \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \widehat{\pi}_{\mathrm{mis}})} L_{\mathrm{mis}}(w, \widehat{\pi}_{\mathrm{mis}}; \Delta^*, \Theta^*) \\ &\leq L_{\mathrm{mis}}(w^{\pi^*}, \pi^*; \Delta^*, \Theta^*) \\ &\quad - \min_{(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \widehat{\pi}_{\mathrm{mis}})} L_{\mathrm{mis}}(w, \widehat{\pi}_{\mathrm{mis}}; \Delta, \Theta) \\ &\leq L_{\mathrm{mis}}(w^{\pi^*}, \pi^*; \Delta^*, \Theta^*) \\ &\quad - \min_{(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \widehat{\pi}_{\mathrm{mis}})} \widehat{L}_{\mathrm{mis}}(w, \widehat{\pi}_{\mathrm{mis}}; \Delta, \Theta) + \widehat{\varepsilon}_L \\ &\leq L_{\mathrm{mis}}(w^{\pi^*}, \pi^*; \Delta^*, \Theta^*) \\ &\quad - \min_{(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi^*)} \widehat{L}_{\mathrm{mis}}(w, \pi^*; \Delta, \Theta) + \widehat{\varepsilon}_L, \end{aligned} \tag{C.4.1}$$

where we use Lemma 4.4.11 in the first inequality; we use Assumption 4.4.4 that $(\Delta^*, \Theta^*) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ with probability at least $1-\delta$ in the second inequality; we use Lemma

C.4.3 in the third inequality; while we use the optimality of $\widehat{\pi}_{\mathrm{mis}}$ in the last inequality. Now, by applying Lemmas C.4.2 and C.4.3, we obtain from (C.4.1) that

$$\begin{aligned}
J(\pi^*) - J(\widehat{\pi}_{\mathrm{mis}}) &\leq L_{\mathrm{mis}}(w^{\pi^*}, \pi^*; \Delta^*, \Theta^*) - \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta^*, \Theta^*) + \varepsilon_L^* + 2\widehat{\varepsilon}_L \\
&\leq \max_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} \left|\Phi^{\pi^*}_{\mathrm{mis}}(w, V^{\pi^*}; \Delta^*, \Theta^*)\right| + \varepsilon_L^* + 2\widehat{\varepsilon}_L \\
&\leq \max_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} \max_{f \in \mathcal{V}} \left|\Phi^{\pi^*}_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*)\right| + \varepsilon_L^* + 2\widehat{\varepsilon}_L \\
&\leq \max_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} \max_{f \in \mathcal{V}} \max \left\{\Phi^{\pi^*}_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*), -\Phi^{\pi^*}_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*)\right\} \\
&\qquad + \varepsilon_L^* + 2\widehat{\varepsilon}_L \\
&\leq \max_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} \max_{f \in \mathcal{V}} \max \left\{\Phi^{\pi^*}_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*), \Phi^{\pi^*}_{\mathrm{mis}}(w, -f; \Delta^*, \Theta^*)\right\} \\
&\qquad + \varepsilon_L^* + 2\widehat{\varepsilon}_L \\
&\leq \max_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} \max_{f \in \mathcal{V}} \Phi^{\pi^*}_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) + \varepsilon_L^* + 2\widehat{\varepsilon}_L, \qquad \text{(C.4.2)}
\end{aligned}$$

where in the second inequality, we use Lemma C.4.1; in the third inequality, we use Assumption 4.4.10; while in the last inequality, we use the fact that $\mathcal{V}$ is symmetric. Now, by Lemma 4.4.12 and plugging the definition of $\widehat{\varepsilon}_L$ and ε_L^* into (C.4.2), it holds with probability at least $1 - \delta$ that

$$J(\pi^*) - J(\widehat{\pi}_{\mathrm{mis}}) \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \cdot \log \frac{NT}{\delta}},$$

which concludes the proof of the theorem. □

C.4.2. Proof of Lemma 4.4.11

Proof. First, by Assumption 4.4.10, we know that $w^\pi \in \mathcal{W}$. For notation simplicity, we denote by $\Phi^\pi_{\mathrm{mis}}(w, f; *) = \Phi^\pi_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*)$ and $\widehat{w}^\pi_* = \widehat{w}^\pi_{\Delta^*,\Theta^*}$ for any (π, w, f). Note that

$$
\begin{aligned}
&\max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(w^\pi, f; *) - \max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(\widehat{w}^\pi_*, f; *) \\
&\quad = \max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(w^\pi, f; *) - \max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w^\pi, f; *) + \max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w^\pi, f; *) - \max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(\widehat{w}^\pi_*, f; *) \\
&\qquad + \max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(\widehat{w}^\pi_*, f; *) - \max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(\widehat{w}^\pi_*, f; *) \\
&\quad \le \max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(w^\pi, f; *) - \max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w^\pi, f; *) + \max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(\widehat{w}^\pi_*, f; *) - \max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(\widehat{w}^\pi_*, f; *) \\
&\quad \le 2 \max_{w\in\mathcal{W}} \left| \max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; *) - \max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w, f; *) \right| \\
&\quad \le 2 \max_{w\in\mathcal{W}} \max_{f\in\mathcal{V}} \left| \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; *) - \Phi^\pi_{\mathrm{mis}}(w, f; *) \right|, \qquad \text{(C.4.3)}
\end{aligned}
$$

where in the first inequality, we use the fact that $w^\pi = \mathrm{argmin}_{w\in\mathcal{W}} \max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w, f; *)$; while in the second inequality, we use $w^\pi \in \mathcal{W}$ by Assumption 4.4.10. In the meanwhile, by Theorem C.7.6, with probability at least $1 - \delta$, it holds for any $(w, f, \pi) \in \mathcal{W} \times \mathcal{V} \times \Pi$ that

$$
\left| \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; *) - \Phi^\pi_{\mathrm{mis}}(w, f; *) \right| \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi} \log \frac{1}{\delta} \log(NT)}, \qquad \text{(C.4.4)}
$$

where we use Assumption (b). Now, combining (C.4.3) and (C.4.4), with probability at least $1-\delta$, we have

$$\max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w^{\pi}, f; *) - \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *) \\ \leq c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi} \log\frac{1}{\delta} \log(NT)} = \alpha_{\mathrm{mis}},$$

which implies that $w^{\pi} \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi)$. This concludes the proof of the lemma. □

C.4.3. Proof of Lemma 4.4.12

Proof. Since $w \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi)$, there exists a pair $(\widetilde{\Delta}, \widetilde{\Theta}) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ such that $w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$. For the simplicity of notations, we denote by

$$\widetilde{w} \in \operatorname*{argmin}_{w\in\mathcal{W}} \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta}), \tag{C.4.5}$$

i.e., $\widetilde{w} = \widehat{w}^{\pi}_{\widetilde{\Delta},\widetilde{\Theta}}$, which is defined in (4.3.5). By the definition of $\widetilde{w}$ and $w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$, with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ and $w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$ that

$$\max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta}) - \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widetilde{w}, f; \widetilde{\Delta}, \widetilde{\Theta}) \leq \alpha_{\mathrm{mis}}. \tag{C.4.6}$$

Further, we observe that

$$\max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w,f;\Delta^*,\Theta^*)$$

(C.4.7)
$$\le \underbrace{\max_{(w,f,\Delta,\Theta)\in(\mathcal{W},\mathcal{V},\mathcal{F}_0,\mathcal{F}_1)} \left|\Phi^{\pi}_{\mathrm{mis}}(w,f;\Delta,\Theta) - \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w,f;\Delta,\Theta)\right|}_{\text{Term (I)}} + \underbrace{\max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w},f;\widetilde{\Delta},\widetilde{\Theta})}_{\text{Term (II)}}$$
$$+ \underbrace{\max_{f\in\mathcal{V}} \left|\widehat{\Phi}^{\pi}_{\mathrm{mis}}(w,f;\Delta^*,\Theta^*) - \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w,f;\widetilde{\Delta},\widetilde{\Theta})\right|}_{\text{Term (III)}} + \alpha_{\mathrm{mis}},$$

where we use (C.4.6) in the last inequality. Now we upper bound terms (I), (II), and (III) on the RHS of (C.4.7).

Upper Bounding Term (I). By Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w,f,\Delta,\Theta,\pi)\in(\mathcal{W},\mathcal{V},\mathcal{F}_0,\mathcal{F}_1,\Pi)$ that

$$\left|\widehat{\Phi}^{\pi}_{\mathrm{mis}}(w,f;\Delta,\Theta) - \Phi^{\pi}_{\mathrm{mis}}(w,f;\Delta,\Theta)\right| \le c\cdot\frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{1}{\delta}\log(NT)},$$

which implies that with probability at least $1-\delta$, we have

$$\text{Term (I)} \le c\cdot\frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{1}{\delta}\log(NT)}. \tag{C.4.8}$$

Upper Bounding Term (II). We introduce the following lemma to help upper bound term (II).

Lemma C.4.4. Suppose (α_0, α_1) is defined in Assumption 4.4.4. With probability at least $1-\delta$, for any $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ and $\pi \in \Pi$, we have

$$\max_{f \in \mathcal{V}} \Phi^\pi_{\mathrm{mis}}(\widehat{w}^\pi_{\Delta,\Theta}, f; \Delta, \Theta) \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \log \frac{1}{\delta} \log(NT)},$$

where $\widehat{w}^\pi_{\Delta,\Theta}$ is defined in (4.3.5), ξ_0 and ξ_1 are constants defined in Assumption 4.4.4.

Proof. See §C.4.7 for a detailed proof. □

By the definition of $\widetilde{w}$ in (C.4.5) and Lemma C.4.4, with probability at least $1-\delta$, we have

$$\text{Term (II)} \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \log \frac{1}{\delta} \log(NT)}. \tag{C.4.9}$$

Upper Bounding Term (III). Note that

$$\begin{aligned}
&\left| \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) - \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta}) \right| \\
&\leq \left| \left(\widehat{\mathbb{E}} - \mathbb{E} \right) \left[\frac{1}{T} \sum_{t=0}^{T-1} \left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t) w(S_t)}{\Delta^*(S_t, A_t) \Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t) w(S_t)}{\widetilde{\Delta}(S_t, A_t) \widetilde{\Theta}(S_t, Z_t)} \right) (f(S_t) - \gamma f(S_{t+1})) \right] \right| \\
&\quad + \left| \mathbb{E} \left[\frac{1}{T} \sum_{t=0}^{T-1} \left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t) w(S_t)}{\Delta^*(S_t, A_t) \Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t) w(S_t)}{\widetilde{\Delta}(S_t, A_t) \widetilde{\Theta}(S_t, Z_t)} \right) (f(S_t) - \gamma f(S_{t+1})) \right] \right|.
\end{aligned} \tag{C.4.10}$$

For the first term on the RHS of (C.4.10), by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w, f, \pi) \in \mathcal{W} \times \mathcal{V} \times \Pi$ that

$$\left|\left(\widehat{\mathbb{E}} - \mathbb{E}\right)\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t) w(S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t) w(S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(f(S_t) - \gamma f(S_{t+1}))\right]\right|$$

(C.4.11)

$$\leq c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma}\sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi}}{NT\kappa}\log\frac{1}{\delta}\log(NT)}.$$

For the second term on the RHS of (C.4.10), by a similar argument as in (C.3.14), it holds with probability at least $1-\delta$ that

$$\left|\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t) w(S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t) w(S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(f(S_t) - \gamma f(S_{t+1}))\right]\right|$$

(C.4.12)

$$\leq \frac{2C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma}\left(\xi_0 C_{\Theta^*}\sqrt{\frac{C_{\Delta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0}\log\frac{1}{\delta}\log(NT)} + \xi_1 C_{\Delta^*}\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_1}\log\frac{1}{\delta}\log(NT)}\right),$$

where in the first inequality, we use the fact that $\|f\|_\infty \leq 1/(1-\gamma)$ and $\|w\|_\infty \leq C_*$; in the third inequality, we use Cauchy Schwarz inequality; while in the last inequality, we use Assumption 4.4.4 with the fact that $(\widetilde{\Delta}, \widetilde{\Theta}) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$. Now, by plugging (C.4.11) and (C.4.12) into (C.4.10), with probability at least $1-\delta$, it holds for any $w \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi)$ and $(f, \pi) \in \mathcal{V} \times \Pi$ that

$$\left|\widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) - \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta})\right|$$

$$\leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \cdot \log\frac{1}{\delta}\log(NT)}. \tag{C.4.13}$$

Now, by plugging (C.4.8), (C.4.9), and (C.4.13) into (C.4.7), with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ and $w \in \cup_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0}\times \mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta,\Theta,\pi)$ that

$$\max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w,f;\Delta^*,\Theta^*) \le c\cdot \frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\cdot \log\frac{1}{\delta}\log(NT)},$$

which concludes the proof of the lemma. □

C.4.4. Proof of Lemma C.4.1

Proof. Since $\Phi^\pi_{\mathrm{mis}}(w^\pi, V^\pi;\Delta^*,\Theta^*)=0$, we have

$$\begin{aligned}
&\Phi^\pi_{\mathrm{mis}}(w,V^\pi;\Delta^*,\Theta^*)\\
&\quad= \Phi^\pi_{\mathrm{mis}}(w,V^\pi;\Delta^*,\Theta^*) - \Phi^\pi_{\mathrm{mis}}(w^\pi,V^\pi;\Delta^*,\Theta^*)\\
&\quad= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}\left(w^\pi(S_t)-w(S_t)\right)\left(V^\pi(S_t)-\gamma V^\pi(S_{t+1})\right)\right]\\
&\quad= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(w^\pi(S_t)-w(S_t)\right)\mathbb{E}_\pi\left[V^\pi(S_t)-\gamma V^\pi(S_{t+1})\,|\,S_t\right]\right]\\
&\quad= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(w^\pi(S_t)-w(S_t)\right)\mathbb{E}_\pi[R_t\,|\,S_t]\right]\\
&\quad= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}\left(w^\pi(S_t)-w(S_t)\right)R_t\right]\\
&\quad= L_{\mathrm{mis}}(w^\pi,\pi;\Delta^*,\Theta^*) - L_{\mathrm{mis}}(w,\pi;\Delta^*,\Theta^*),
\end{aligned}$$

which concludes the proof of the lemma. □

C.4.5. Proof of Lemma C.4.2

Proof. With a slight abuse of notations, we denote by

$$w_0 \in \operatorname*{argmin}_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta^*, \Theta^*), \qquad w_1 \in \operatorname*{argmin}_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta, \Theta).$$

Then we have

$$\begin{aligned}
&\left| \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta^*, \Theta^*) - \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta, \Theta) \right| \\
&\quad = \left| L_{\mathrm{mis}}(w_0, \pi^*; \Delta^*, \Theta^*) - L_{\mathrm{mis}}(w_1, \pi^*; \Delta, \Theta) \right| \\
&\quad \le \left| L_{\mathrm{mis}}(w_0, \pi^*; \Delta^*, \Theta^*) - L_{\mathrm{mis}}(w^{\pi^*}, \pi^*; \Delta^*, \Theta^*) \right| \\
&\qquad + \left| L_{\mathrm{mis}}(w^{\pi^*}, \pi^*; \Delta^*, \Theta^*) - L_{\mathrm{mis}}(w_1, \pi^*; \Delta^*, \Theta^*) \right| \\
&\qquad + \left| L_{\mathrm{mis}}(w_1, \pi^*; \Delta^*, \Theta^*) - L_{\mathrm{mis}}(w_1, \pi^*; \Delta, \Theta) \right| \\
&\quad = \underbrace{\left| \Phi^{\pi^*}_{\mathrm{mis}}(w_0, V^{\pi^*}; \Delta^*, \Theta^*) \right|}_{\text{Term (I)}} + \underbrace{\left| \Phi^{\pi^*}_{\mathrm{mis}}(w_1, V^{\pi^*}; \Delta^*, \Theta^*) \right|}_{\text{Term (II)}} \\
&\qquad + \underbrace{\left| L_{\mathrm{mis}}(w_1, \pi^*; \Delta^*, \Theta^*) - L_{\mathrm{mis}}(w_1, \pi^*; \Delta, \Theta) \right|}_{\text{Term (III)}}.
\end{aligned} \tag{C.4.14}$$

We upper bound terms (I), (II), and (III) on the RHS of (C.4.14), respectively.

Upper Bounding Term (I). Note that with probability at least $1-\delta$, we have

$$\begin{aligned}
\left|\Phi^{\pi^*}_{\mathrm{mis}}(w_0, V^{\pi^*}; \Delta^*, \Theta^*)\right| &\leq \max_{f\in\mathcal{V}} \left|\Phi^{\pi^*}_{\mathrm{mis}}(w_0, f; \Delta^*, \Theta^*)\right| \\
&= \max_{f\in\mathcal{V}} \max\left\{\Phi^{\pi^*}_{\mathrm{mis}}(w_0, f; \Delta^*, \Theta^*), -\Phi^{\pi^*}_{\mathrm{mis}}(w_0, f; \Delta^*, \Theta^*)\right\} \\
&= \max_{f\in\mathcal{V}} \max\left\{\Phi^{\pi^*}_{\mathrm{mis}}(w_0, f; \Delta^*, \Theta^*), \Phi^{\pi^*}_{\mathrm{mis}}(w_0, -f; \Delta^*, \Theta^*)\right\} \\
&= \max_{f\in\mathcal{V}} \Phi^{\pi^*}_{\mathrm{mis}}(w_0, f; \Delta^*, \Theta^*) \\
&\leq c\cdot\frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V}}\log\frac{1}{\delta}\log(NT)}, \qquad \text{(C.4.15)}
\end{aligned}$$

where in the first inequality, we use the fact that $V^{\pi^*}\in\mathcal{V}$; in the third equality, we use the fact that $\mathcal{V}$ is symmetric; in the last inequality, by noting that $w_0\in\cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta,\Theta,\pi)$, we use Lemma 4.4.12. This upper bounds term (I) on the RHS of (C.4.14).

Upper Bounding Term (II). Similar to (C.4.15), note that $w_1\in\cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta,\Theta,\pi)$, it holds with probability at least $1-\delta$ that

(C.4.16)

$$\left|\Phi^{\pi^*}_{\mathrm{mis}}(w_1, V^{\pi^*}; \Delta^*, \Theta^*)\right| \leq c\cdot\frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V}}\log\frac{1}{\delta}\log(NT)},$$

which upper bounds term (II) on the RHS of (C.4.14).

Upper Bounding Term (III). Note that with probability at least $1-\delta$, it holds for any $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ that

$$
\begin{aligned}
&|L_{\mathrm{mis}}(w_1, \pi^*; \Delta^*, \Theta^*) - L_{\mathrm{mis}}(w_1, \pi^*; \Delta, \Theta)| \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^\top A_t \pi^*(A_t \,|\, S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi^*(A_t \,|\, S_t)}{\Delta(S_t, A_t)\Theta(S_t, Z_t)}\right) w_1(S_t) R_t\right]
\end{aligned}
$$

(C.4.17)

$$
\le C_{\Delta^*} C_{\Theta^*} C_* \left(\xi_0 C_{\Theta^*}\sqrt{\frac{C_{\Delta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0}\log\frac{1}{\delta}\log(NT)} + \xi_1 C_{\Delta^*}\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_1}\log\frac{1}{\delta}\log(NT)}\right),
$$

where we use Cauchy-Schwarz inequality and Assumption 4.4.4 in the last inequality.

Now, by plugging (C.4.15), (C.4.16), and (C.4.17) into (C.4.14), with probability at least $1-\delta$, it holds for any $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ that

$$
\begin{aligned}
&\left|\min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta^*, \Theta^*) - \min_{w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi^*)} L_{\mathrm{mis}}(w, \pi^*; \Delta, \Theta)\right| \\
&\qquad \le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}}\log\frac{1}{\delta}\log(NT)},
\end{aligned}
$$

which concludes the proof of the lemma. □

C.4.6. Proof of Lemma C.4.3

Proof. By Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w, \Delta, \Theta, \pi) \in \mathcal{W} \times \mathcal{F}_0 \times \mathcal{F}_1 \times \Pi$ that

$$
\left|L_{\mathrm{mis}}(w, \pi; \Delta, \Theta) - \widehat{L}_{\mathrm{mis}}(w, \pi; \Delta, \Theta)\right| \le c \cdot C_{\Delta^*} C_{\Theta^*} C_* \sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \Pi}\log\frac{1}{\delta}\log(NT)},
$$

which concludes the proof of the lemma. □

C.4.7. Proof of Lemma C.4.4

Proof. Note that

$$\begin{aligned}
&\max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{\Delta,\Theta}, f; \Delta, \Theta) \\
&\quad = \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{\Delta,\Theta}, f; \Delta, \Theta) - \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{\Delta,\Theta}, f; \Delta, \Theta) + \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{\Delta,\Theta}, f; \Delta, \Theta) \\
&\qquad - \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta, \Theta) + \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta, \Theta) - \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta, \Theta) \\
&\qquad + \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta, \Theta) - \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta^*, \Theta^*) \\
&\quad \le 2 \max_{w\in\mathcal{W}} \max_{f\in\mathcal{V}} \left| \Phi^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta) - \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta) \right| \\
&\qquad + \max_{f\in\mathcal{V}} \left| \Phi^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta, \Theta) - \Phi^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta^*, \Theta^*) \right|, \qquad \text{(C.4.18)}
\end{aligned}$$

where we use the fact that $\widehat{w}^{\pi}_{\Delta,\Theta} \in \mathrm{argmin}_{w\in\mathcal{W}} \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta)$ in the last inequality. In the meanwhile, by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w, f, \pi) \in \mathcal{W} \times \mathcal{V} \times \Pi$ that

$$\left| \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta) - \Phi^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta) \right| \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi} \log\frac{1}{\delta} \log(NT)}. \qquad \text{(C.4.19)}$$

Also, we upper bound the second term on the RHS of (C.4.18) with probability at least $1-\delta$ as follows,

$$
\begin{aligned}
&|\Phi^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta, \Theta) - \Phi^{\pi}_{\mathrm{mis}}(w^{\pi}, f; \Delta^*, \Theta^*)| \\
&= \left| \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \left(\frac{Z_t^{\top} A_t \pi(A_t \,|\, S_t) w(S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^{\top} A_t \pi(A_t \,|\, S_t) w(S_t)}{\Delta(S_t, A_t)\Theta(S_t, Z_t)} \right) (f(S_t) - \gamma f(S_{t+1})) \right] \right| \\
&\le \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \left(\xi_0 C_{\Theta^*} \sqrt{\frac{C_{\Delta^*}}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0} \log\frac{1}{\delta} \log(NT)} + \xi_1 C_{\Delta^*} \sqrt{\frac{C_{\Theta^*}}{NT\kappa} \mathfrak{C}_{\mathcal{F}_1} \log\frac{1}{\delta} \log(NT)} \right),
\end{aligned} \tag{C.4.20}
$$

where we use Cauchy-Schwarz inequality and Assumption 4.4.4 in the last inequality.

Now, by plugging (C.4.19) and (C.4.20) into (C.4.18), with probability at least $1-\delta$, it holds for any $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ and $\pi \in \Pi$ that

$$
\max_{f \in \mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{\Delta,\Theta}, f; \Delta, \Theta) \le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \log\frac{1}{\delta} \log(NT)},
$$

which concludes the proof of the lemma. □

C.5. Proof of Results in §4.4.3

C.5.1. Proof of Theorem 4.4.14

Proof. We split the proof into two case: (i) Assumption 4.4.6 holds; (ii) Assumption 4.4.10 holds.

Case (i): Assumption 4.4.6 holds. We introduce the following supporting lemmas.

Lemma C.5.1. For any policy π, with probability at least $1-\delta$ with $c/(NT)^2 \le \delta \le 1$, it holds for any $(w, v, \Delta, \Theta, \pi) \in \mathcal{W} \times \mathcal{V} \times \mathcal{F}_0 \times \mathcal{F}_1 \times \Pi$ that

$$\begin{aligned}&\Big| L_{\mathsf{dr}}(w, v, \pi; \Delta, \Theta) - \widehat{L}_{\mathsf{dr}}(w, v, \pi; \Delta, \Theta) \Big| \\ &\quad \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \log \frac{1}{\delta} \log(NT)} = \widehat{\epsilon}_L.\end{aligned}$$

Proof. See §C.5.3 for a detailed proof. □

Lemma C.5.2. Suppose that $(\alpha_0, \alpha_1, \alpha_{\mathrm{mis}}, \alpha_{\mathsf{vf}})$ is defined in Assumption 4.4.4, Lemmas 4.4.11, and 4.4.7. With probability at least $1-\delta$, it holds for any $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ that

$$\begin{aligned}&\left| \min_{(w,v) \in \mathsf{conf}_{\alpha_{\mathrm{mis}}, \alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta^*, \Theta^*) - \min_{(w,v) \in \mathsf{conf}_{\alpha_{\mathrm{mis}}, \alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) \right| \\ &\quad \le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}} \log \frac{1}{\delta} \log(NT)} = \epsilon^*_L.\end{aligned}$$

Proof. See §C.5.4 for a detailed proof. □

By the definition of L_{dr}, it holds with probability at least $1-\delta$ that

$$\begin{aligned}
J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) &= J(\pi^*) - L_{\mathsf{dr}}(w, V^{\widehat{\pi}_{\mathsf{dr}}}, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) \\
&\le J(\pi^*) - \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\widehat{\pi}_{\mathsf{dr}})} L_{\mathsf{dr}}(w, v, \widehat{\pi}_{\mathsf{dr}}; \Delta, \Theta) \\
&\le J(\pi^*) - \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\widehat{\pi}_{\mathsf{dr}})} \widehat{L}_{\mathsf{dr}}(w, v, \widehat{\pi}_{\mathsf{dr}}; \Delta, \Theta) + \widehat{\epsilon}_L \\
&\le J(\pi^*) - \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} \widehat{L}_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) + \widehat{\epsilon}_L \\
&\le J(\pi^*) - \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) + 2\widehat{\epsilon}_L, \qquad \text{(C.5.1)}
\end{aligned}$$

where in the first inequality, we use Assumption 4.4.4 that $(\Delta^*, \Theta^*) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ with probability at least $1-\delta$, and Lemma 4.4.7 with Assumption 4.4.6 that $V^{\widehat{\pi}_{\mathsf{dr}}} \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \widehat{\pi}_{\mathsf{dr}})$ with probability at least $1-\delta$; in the second inequality, we use Lemma C.5.1; in the third inequality, we use the optimality of $\widehat{\pi}_{\mathsf{dr}}$; while in the last inequality, we use Lemma C.5.1 again. By combining Lemma C.5.2 and (C.5.1), we have

$$\begin{aligned}
&J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \\
&\quad\le J(\pi^*) - \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta^*, \Theta^*) + 2\widehat{\epsilon}_L + \epsilon^*_L \\
&\quad= L_{\mathsf{dr}}(w, V^{\pi^*}, \pi^*; \Delta^*, \Theta^*) - \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta^*, \Theta^*) + 2\widehat{\epsilon}_L + \epsilon^*_L \\
&\quad= \max_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| L_{\mathsf{dr}}(w, V^{\pi^*}, \pi^*; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w, v, \pi^*; \Delta^*, \Theta^*) \right| + 2\widehat{\epsilon}_L + \epsilon^*_L \\
&\text{(C.5.2)} \\
&\quad= \max_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| \Phi^{\pi^*}_{\mathsf{vf}}(v, w^{\pi^*}; \Delta^*, \Theta^*) - \Phi^{\pi^*}_{\mathsf{vf}}(v, w; \Delta^*, \Theta^*) \right| + 2\widehat{\epsilon}_L + \epsilon^*_L,
\end{aligned}$$

where we use the following fact in the last equality,

$$L_{\mathsf{dr}}(w, V^{\pi^*}, \pi^*; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w, v, \pi^*; \Delta^*, \Theta^*) = \Phi_{\mathsf{vf}}^{\pi^*}(v, w^{\pi^*}; \Delta^*, \Theta^*) - \Phi_{\mathsf{vf}}^{\pi^*}(v, w; \Delta^*, \Theta^*).$$

In the meanwhile, note that by Assumption 4.4.6 that $w^{\pi^*} \in \mathcal{W}$, we obtain from (C.5.2) that

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) &\leq 2 \max_{(w,v) \in \mathsf{conf}_{\alpha_{\mathrm{mis}}, \alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi^*)} \left| \Phi_{\mathsf{vf}}^{\pi^*}(v, w; \Delta^*, \Theta^*) \right| + 2\widehat{\epsilon}_L + \epsilon_L^* \\ &\leq c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \log \frac{NT}{\delta}}, \end{aligned} \tag{C.5.3}$$

where we use Lemma 4.4.8 and plug in the definition of ϵ_L and ϵ_L^* in the last inequality. This concludes the proof of case (i).

Case (ii): Assumption 4.4.10 holds. It holds with probability at least $1 - \delta$ that

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) &= J(\pi^*) - L_{\mathsf{dr}}(w^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) \\ &\leq J(\pi^*) - \min_{(\Delta,\Theta) \in \mathsf{conf}_{\alpha_0}^0 \times \mathsf{conf}_{\alpha_1}^1} \min_{(w,v) \in \mathsf{conf}_{\alpha_{\mathrm{mis}}, \alpha_{\mathsf{vf}}}(\Delta, \Theta, \widehat{\pi}_{\mathsf{dr}})} L_{\mathsf{dr}}(w, v, \widehat{\pi}_{\mathsf{dr}}; \Delta, \Theta) \\ &\leq J(\pi^*) - \min_{(\Delta,\Theta) \in \mathsf{conf}_{\alpha_0}^0 \times \mathsf{conf}_{\alpha_1}^1} \min_{(w,v) \in \mathsf{conf}_{\alpha_{\mathrm{mis}}, \alpha_{\mathsf{vf}}}(\Delta, \Theta, \widehat{\pi}_{\mathsf{dr}})} \widehat{L}_{\mathsf{dr}}(w, v, \widehat{\pi}_{\mathsf{dr}}; \Delta, \Theta) + \widehat{\epsilon}_L \\ &\leq J(\pi^*) - \min_{(\Delta,\Theta) \in \mathsf{conf}_{\alpha_0}^0 \times \mathsf{conf}_{\alpha_1}^1} \min_{(w,v) \in \mathsf{conf}_{\alpha_{\mathrm{mis}}, \alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi^*)} \widehat{L}_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) + \widehat{\epsilon}_L, \end{aligned}$$

where we use Assumption 4.4.4 that $(\Delta^*, \Theta^*) \in \mathsf{conf}_{\alpha_0}^0 \times \mathsf{conf}_{\alpha_1}^1$ with probability at least $1 - \delta$ and Assumption 4.4.10 that $w^\pi \in \mathcal{W}$ for any $\pi \in \Pi$ in the first inequality, we use Lemma C.5.1 in the second inequality, and we use the optimality of $\widehat{\pi}_{\mathsf{dr}}$ in the last inequality. Further, by Lemmas C.5.1 and C.5.2, we have

$$
\begin{aligned}
&J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \\
&\quad \leq J(\pi^*) - \min_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0}\times \mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} L_{\mathsf{dr}}(w,v,\pi^*;\Delta,\Theta) + 2\widehat{\epsilon}_L \\
&\quad \leq J(\pi^*) - \min_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) + 2\widehat{\epsilon}_L + \epsilon^*_L \\
&\quad \leq \max_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| L_{\mathsf{dr}}(w^{\pi^*},v,\pi^*;\Delta^*,\Theta^*) - L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*)\right| + 2\widehat{\epsilon}_L + \epsilon^*_L \\
&\quad \leq \max_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| \Phi^{\pi^*}_{\text{mis}}(w,V^{\pi^*};\Delta^*,\Theta^*) - \Phi^{\pi^*}_{\text{mis}}(w,v;\Delta^*,\Theta^*)\right| + 2\widehat{\epsilon}_L + \epsilon^*_L \\
&\quad \leq 2 \max_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| \Phi^{\pi^*}_{\text{mis}}(w,v;\Delta^*,\Theta^*)\right| + 2\widehat{\epsilon}_L + \epsilon^*_L \\
&\quad \leq 2 \max_{w\in \mathsf{conf}^{\text{mis}}_{\alpha_{\text{mis}}}(\Delta^*,\Theta^*,\pi^*)} \max_{v\in\mathcal{V}} \left| \Phi^{\pi^*}_{\text{mis}}(w,v;\Delta^*,\Theta^*)\right| + 2\widehat{\epsilon}_L + \epsilon^*_L
\end{aligned}
$$

where in the third inequality, we use the fact that $J(\pi^*) = L_{\mathsf{dr}}(w^{\pi^*},v,\pi^*;\Delta^*,\Theta^*)$; in the forth inequality, we use the following fact

$$
L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) - L_{\mathsf{dr}}(w^{\pi^*},v,\pi^*;\Delta^*,\Theta^*) = -\Phi^{\pi^*}_{\text{mis}}(w,V^{\pi^*};\Delta^*,\Theta^*) + \Phi^{\pi^*}_{\text{mis}}(w,v;\Delta^*,\Theta^*)
$$

for any $(w,v)\in\mathcal{W}\times\mathcal{V}$; in the fifth inequality, we use the fact that $V^{\pi^*}\in\mathcal{V}$ by Assumption 4.4.10 and $V^{\pi^*}\in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)$ with probability at least $1-\delta$ by Lemma 4.4.7. Now, by Lemma 4.4.12 and the fact that $\mathcal{V}$ is symmetric, we obtain that

$$
J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \leq c\cdot \frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{NT}{\delta}}, \tag{C.5.4}
$$

which concludes the proof of case (ii).

By combining (C.5.3) and (C.5.4), we conclude the proof of the theorem. □

C.5.2. Proof of Theorem 4.4.16

Proof. Recall that

$$\widetilde{v}^\pi \in \operatorname*{argmin}_{v\in\mathcal{V}} \max_{w\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(v, w; \Delta^*, \Theta^*), \qquad \widetilde{w}^\pi \in \operatorname*{argmin}_{w\in\mathcal{W}} \max_{v\in\mathcal{V}} \Phi^\pi_{\text{mis}}(w, v; \Delta^*, \Theta^*).$$

We split the proof into the following two parts.

Part (i). We first introduce the following lemmas.

Lemma C.5.3. Suppose α_{vf} is defined in Lemma 4.4.7 and $c/(NT)^2 \le \delta \le 1$. Then under Assumptions (b) and 4.4.3, with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ that $\widetilde{v}^\pi \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi)$.

Proof. See §C.5.5 for a detailed proof. □

Lemma C.5.4. Suppose that $(\alpha_0, \alpha_1, \alpha_{\mathsf{vf}})$ is defined in Assumption 4.4.4 and Lemma 4.4.7 and $c/(NT)^2 \le \delta \le 1$. Then under Assumptions 4.2.2, 4.2.3, 4.4.3 and 4.4.4, with probability at least $1-\delta$, it holds for any policy $\pi \in \Pi$ and $v \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi)$ that

$$\max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) \le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log\frac{1}{\delta}\log(NT)}$$
$$+ \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(\widetilde{v}^\pi, g; \Delta^*, \Theta^*).$$

Proof. See §C.5.6 for a detailed proof. □

By the definition of L_{dr}, it holds that

$$
\begin{aligned}
&J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \\
&\qquad = J(\pi^*) - L_{\mathsf{dr}}(w, V^{\widehat{\pi}_{\mathsf{dr}}}, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) \\
&\qquad = J(\pi^*) - L_{\mathsf{dr}}(w, \widetilde{v}^{\widehat{\pi}_{\mathsf{dr}}}, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) + L_{\mathsf{dr}}(w, \widetilde{v}^{\widehat{\pi}_{\mathsf{dr}}}, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w, V^{\widehat{\pi}_{\mathsf{dr}}}, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*). \qquad \text{(C.5.5)}
\end{aligned}
$$

Note that

$$
\Big| L_{\mathsf{dr}}(w, \widetilde{v}^{\widehat{\pi}_{\mathsf{dr}}}, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w, V^{\widehat{\pi}_{\mathsf{dr}}}, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) \Big| \le C_* C_{\Delta^*} C_{\Theta^*} \varepsilon_{\mathsf{vf}}^{\mathcal{V}}. \qquad \text{(C.5.6)}
$$

In the meanwhile, by Assumption 4.4.4 and Lemma C.5.3, it holds with probability at least $1 - \delta$ that

$$
\begin{aligned}
L_{\mathsf{dr}}(w, \widetilde{v}^{\widehat{\pi}_{\mathsf{dr}}}, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) &\ge \min_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \ \min_{(w,v)\in \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\widehat{\pi}_{\mathsf{dr}})} L_{\mathsf{dr}}(w, v, \widehat{\pi}_{\mathsf{dr}}; \Delta, \Theta) \\
&\ge \min_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \ \min_{(w,v)\in \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\widehat{\pi}_{\mathsf{dr}})} \widehat{L}_{\mathsf{dr}}(w, v, \widehat{\pi}_{\mathsf{dr}}; \Delta, \Theta) - \widehat{\epsilon}_L \\
&\ge \min_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \ \min_{(w,v)\in \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} \widehat{L}_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) - \widehat{\epsilon}_L \\
&\ge \min_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \ \min_{(w,v)\in \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) - 2\widehat{\epsilon}_L, \qquad \text{(C.5.7)}
\end{aligned}
$$

where in the second inequality, we use Lemma C.5.1; in the third inequality, we use the optimality of $\widehat{\pi}_{\mathsf{dr}}$; in the forth inequality, we again use Lemma C.5.1. Now, by plugging

(C.5.6) and (C.5.7) into (C.5.5), we have

$$
\begin{aligned}
& J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \\
& \le J(\pi^*) - \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} L_{\mathsf{dr}}(w,v,\pi^*;\Delta,\Theta) \qquad \text{(C.5.8)} \\
& \quad + 2\widehat{\epsilon}_L + C_* C_{\Delta^*} C_{\Theta^*} \varepsilon^{\mathcal{V}}_{\mathsf{vf}}. \qquad \text{(C.5.9)}
\end{aligned}
$$

By combining Lemma C.5.2 and (C.5.8), we have

$$
\begin{aligned}
& J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \\
& \le J(\pi^*) - \min_{(w,v)\in\mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) + 2\widehat{\epsilon}_L + \epsilon^*_L + C_* C_{\Delta^*} C_{\Theta^*} \varepsilon^{\mathcal{V}}_{\mathsf{vf}} \\
& = L_{\mathsf{dr}}(w, V^{\pi^*}, \pi^*;\Delta^*,\Theta^*) - \min_{(w,v)\in\mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) \\
& \qquad + 2\widehat{\epsilon}_L + \epsilon^*_L + C_* C_{\Delta^*} C_{\Theta^*} \varepsilon^{\mathcal{V}}_{\mathsf{vf}} \\
& = \max_{(w,v)\in\mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| L_{\mathsf{dr}}(w, V^{\pi^*}, \pi^*;\Delta^*,\Theta^*) - L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) \right| \\
& \qquad + 2\widehat{\epsilon}_L + \epsilon^*_L + C_* C_{\Delta^*} C_{\Theta^*} \varepsilon^{\mathcal{V}}_{\mathsf{vf}} \\
& \text{(C.5.10)} \\
& = \max_{(w,v)\in\mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| \Phi^{\pi^*}_{\mathsf{vf}}(v, w^{\pi^*};\Delta^*,\Theta^*) - \Phi^{\pi^*}_{\mathsf{vf}}(v, w;\Delta^*,\Theta^*) \right| \\
& \qquad + 2\widehat{\epsilon}_L + \epsilon^*_L + C_* C_{\Delta^*} C_{\Theta^*} \varepsilon^{\mathcal{V}}_{\mathsf{vf}},
\end{aligned}
$$

where we use the following fact in the last equality,

$$
L_{\mathsf{dr}}(w, V^{\pi^*}, \pi^*;\Delta^*,\Theta^*) - L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) = \Phi^{\pi^*}_{\mathsf{vf}}(v, w^{\pi^*};\Delta^*,\Theta^*) - \Phi^{\pi^*}_{\mathsf{vf}}(v, w;\Delta^*,\Theta^*).
$$

We upper bound the first term on the RHS of (C.5.10) as follows,

$$\begin{aligned}
&\max_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left|\Phi_{\mathsf{vf}}^{\pi^*}(v,w^{\pi^*};\Delta^*,\Theta^*)-\Phi_{\mathsf{vf}}^{\pi^*}(v,w;\Delta^*,\Theta^*)\right| \\
&\quad\le \max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)}\max_{w\in\mathcal{W}} \left|\Phi_{\mathsf{vf}}^{\pi^*}(v,\widetilde{w}^{\pi^*};\Delta^*,\Theta^*)-\Phi_{\mathsf{vf}}^{\pi^*}(v,w;\Delta^*,\Theta^*)\right| \\
&\qquad + \max_{v\in\mathcal{V}}\left|\Phi_{\mathsf{vf}}^{\pi^*}(v,w^{\pi^*};\Delta^*,\Theta^*)-\Phi_{\mathsf{vf}}^{\pi^*}(v,\widetilde{w}^{\pi^*};\Delta^*,\Theta^*)\right| \\
&\quad\le 2\max_{v\in\mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)}\max_{w\in\mathcal{W}} \left|\Phi_{\mathsf{vf}}^{\pi^*}(v,w;\Delta^*,\Theta^*)\right| \qquad\qquad \text{(C.5.11)} \\
&\qquad + \max_{v\in\mathcal{V}}\left|\Phi_{\mathsf{vf}}^{\pi^*}(v,w^{\pi^*};\Delta^*,\Theta^*)-\Phi_{\mathsf{vf}}^{\pi^*}(v,\widetilde{w}^{\pi^*};\Delta^*,\Theta^*)\right|,
\end{aligned}$$

where in the first inequality, we use triangle inequality; in the second inequality, we use the definition of $\widetilde{w}^{\pi^*}$ that $\widetilde{w}^{\pi^*}\in\mathcal{W}$. By Lemma C.5.4 and Assumption 4.4.15, we obtain from (C.5.11) that

$$\begin{aligned}
&\max_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left|\Phi_{\mathsf{vf}}^{\pi^*}(v,w^{\pi^*};\Delta^*,\Theta^*)-\Phi_{\mathsf{vf}}^{\pi^*}(v,w;\Delta^*,\Theta^*)\right| \\
&\quad\le c\cdot\frac{C_{\Delta^*}^2C_{\Theta^*}^2C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\cdot\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\cdot\log\frac{1}{\delta}\log(NT)} \qquad\qquad \text{(C.5.12)} \\
&\qquad + 2\max_{g\in\mathcal{W}}\Phi_{\mathsf{vf}}^{\pi^*}(\widetilde{v}^{\pi^*},g;\Delta^*,\Theta^*)+C_{\Delta^*}C_{\Theta^*}\varepsilon_{\mathsf{vf}}^{\mathcal{W}}/(1-\gamma).
\end{aligned}$$

Also, we have

$$
\begin{aligned}
&\max_{g\in\mathcal{W}} \Phi_{\mathsf{vf}}^{\pi^*}(\widetilde{v}^{\pi^*}, g; \Delta^*, \Theta^*)\\
&\qquad = \max_{g\in\mathcal{W}} \Phi_{\mathsf{vf}}^{\pi^*}(\widetilde{v}^{\pi^*}, g; \Delta^*, \Theta^*) - \max_{g\in\mathcal{W}} \Phi_{\mathsf{vf}}^{\pi^*}(V^{\pi^*}, g; \Delta^*, \Theta^*) + \max_{g\in\mathcal{W}} \Phi_{\mathsf{vf}}^{\pi^*}(V^{\pi^*}, g; \Delta^*, \Theta^*)\\
&\qquad = \max_{g\in\mathcal{W}} \Phi_{\mathsf{vf}}^{\pi^*}(\widetilde{v}^{\pi^*}, g; \Delta^*, \Theta^*) - \max_{g\in\mathcal{W}} \Phi_{\mathsf{vf}}^{\pi^*}(V^{\pi^*}, g; \Delta^*, \Theta^*)\\
&\qquad = \max_{g\in\mathcal{W}} \left|\Phi_{\mathsf{vf}}^{\pi^*}(\widetilde{v}^{\pi^*}, g; \Delta^*, \Theta^*) - \Phi_{\mathsf{vf}}^{\pi^*}(V^{\pi^*}, g; \Delta^*, \Theta^*)\right|
\end{aligned}
$$

(C.5.13)

$$
\qquad \le C_* C_{\Delta^*} C_{\Theta^*} \varepsilon_{\mathsf{vf}}^{\mathcal{V}},
$$

where we use Assumption 4.4.15 in the last inequality. By plugging (C.5.12) and (C.5.13) into (C.5.10), we have

$$
\begin{aligned}
J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \le c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{NT}{\delta}}\\
+ 3C_{\Delta^*}C_{\Theta^*}\left(C_*\varepsilon_{\mathsf{vf}}^{\mathcal{V}} + \varepsilon_{\mathsf{vf}}^{\mathcal{W}}/(1-\gamma)\right),
\end{aligned} \tag{C.5.14}
$$

where we plug in the definition of ϵ_L and ϵ_L^* in the last inequality.

Part (ii). We first introduce the following lemmas.

Lemma C.5.5. Suppose α_{mis} is defined in Lemma 4.4.11 and and $c/(NT)^2 \le \delta \le 1$. Then under Assumptions (b) and 4.4.3, with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ that $\widetilde{w}^{\pi} \in \mathsf{conf}_{\alpha_{\mathrm{mis}}}^{\mathrm{mis}}(\Delta^*, \Theta^*, \pi)$.

Proof. See §C.5.7 for a detailed proof. □

Lemma C.5.6. Suppose that $(\alpha_0, \alpha_1, \alpha_{\mathrm{mis}})$ is defined in Assumption 4.4.4 and Lemma 4.4.11, and $c/(NT)^2 \le \delta \le 1$. Then under Assumptions 4.2.2, 4.2.6, 4.4.3, and 4.4.4, with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ and $w \in \cup_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0}\times \mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi)$ that

$$\max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) \le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log\frac{1}{\delta}\log(NT)} \\ + \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \Delta^*, \Theta^*).$$

Proof. See §C.5.8 for a detailed proof. □

By the definition of L_{dr}, we have

$$J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}})$$

$$= J(\pi^*) - L_{\mathsf{dr}}(w^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*)$$

(C.5.15)

$$= J(\pi^*) - L_{\mathsf{dr}}(\widetilde{w}^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) + L_{\mathsf{dr}}(\widetilde{w}^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*).$$

By Assumption 4.4.15, we have

$$\text{(C.5.16)} \qquad \left| L_{\mathsf{dr}}(\widetilde{w}^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) \right| \le C_{\Delta^*} C_{\Theta^*} \varepsilon^{\mathcal{W}}_{\mathrm{mis}}/(1-\gamma).$$

In the meanwhile, by Assumption 4.4.4 and Lemma C.5.5, it holds with probability at least $1-\delta$ that

$$\begin{aligned} L_{\mathsf{dr}}(\widetilde{w}^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) &\geq \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\widehat{\pi}_{\mathsf{dr}})} L_{\mathsf{dr}}(w, v, \widehat{\pi}_{\mathsf{dr}}; \Delta, \Theta) \\ &\geq \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\widehat{\pi}_{\mathsf{dr}})} \widehat{L}_{\mathsf{dr}}(w, v, \widehat{\pi}_{\mathsf{dr}}; \Delta, \Theta) - \widehat{\epsilon}_L \\ &\geq \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} \widehat{L}_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) - \widehat{\epsilon}_L \\ &\geq \min_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) - 2\widehat{\epsilon}_L, \end{aligned} \tag{C.5.17}$$

where in the second inequality, we use Lemma C.5.1; in the third inequality, we use the optimality of $\widehat{\pi}_{\mathsf{dr}}$; in the forth inequality, we again use Lemma C.5.1. Further, combining Lemma C.5.2 and (C.5.17), we have

$$L_{\mathsf{dr}}(\widetilde{w}^{\widehat{\pi}_{\mathsf{dr}}}, v, \widehat{\pi}_{\mathsf{dr}}; \Delta^*, \Theta^*) \geq \min_{(w,v)\in\mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta^*, \Theta^*) - 2\widehat{\epsilon}_L - \epsilon^*_L. \tag{C.5.18}$$

Now, by plugging (C.5.16) and (C.5.18) into (C.5.15), we have

$$
\begin{aligned}
& J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \\
& \quad \le J(\pi^*) - \min_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) \\
& \qquad + 2\widehat{\epsilon}_L + \epsilon_L^* + C_{\Delta^*}C_{\Theta^*}\varepsilon_{\text{mis}}^{\mathcal{W}}/(1-\gamma) \\
& \quad \le \max_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| L_{\mathsf{dr}}(w^{\pi^*},v,\pi^*;\Delta^*,\Theta^*) - L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) \right| \\
& \qquad + 2\widehat{\epsilon}_L + \epsilon_L^* + C_{\Delta^*}C_{\Theta^*}\varepsilon_{\text{mis}}^{\mathcal{W}}/(1-\gamma) \\
& \quad = \max_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left| \Phi_{\text{mis}}^{\pi^*}(w,V^{\pi^*};\Delta^*,\Theta^*) - \Phi_{\text{mis}}^{\pi^*}(w,v;\Delta^*,\Theta^*) \right| \\
& \qquad + 2\widehat{\epsilon}_L + \epsilon_L^* + C_{\Delta^*}C_{\Theta^*}\varepsilon_{\text{mis}}^{\mathcal{W}}/(1-\gamma),
\end{aligned}
\tag{C.5.19}
$$

where in the second inequality, we use the fact that $J(\pi^*) = L_{\mathsf{dr}}(w^{\pi^*},v,\pi^*;\Delta^*,\Theta^*)$; in the last equality, we use the following fact

$$
L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) - L_{\mathsf{dr}}(w^{\pi^*},v,\pi^*;\Delta^*,\Theta^*) = -\Phi_{\text{mis}}^{\pi^*}(w,V^{\pi^*};\Delta^*,\Theta^*) + \Phi_{\text{mis}}^{\pi^*}(w,v;\Delta^*,\Theta^*)
$$

for any $(w, v) \in \mathcal{W} \times \mathcal{V}$. We upper bound the first term on the RHS of (C.5.19) as follows,

$$
\begin{aligned}
&\max_{(w,v)\in \mathsf{conf}_{\alpha_{\text{mis}},\alpha_{\text{vf}}}(\Delta^*,\Theta^*,\pi^*)} \left|\Phi_{\text{mis}}^{\pi^*}(w, V^{\pi^*}; \Delta^*, \Theta^*) - \Phi_{\text{mis}}^{\pi^*}(w, v; \Delta^*, \Theta^*)\right| \\
&\quad \le \max_{w\in \mathsf{conf}^{\text{mis}}_{\alpha_{\text{mis}}}(\Delta^*,\Theta^*,\pi^*)} \max_{v\in\mathcal{V}} \left|\Phi_{\text{mis}}^{\pi^*}(w, v; \Delta^*, \Theta^*) - \Phi_{\text{mis}}^{\pi^*}(w, \widetilde{v}^{\pi^*}; \Delta^*, \Theta^*)\right| \\
&\qquad + \max_{w\in\mathcal{W}} \left|\Phi_{\text{mis}}^{\pi^*}(w, V^{\pi^*}; \Delta^*, \Theta^*) - \Phi_{\text{mis}}^{\pi^*}(w, \widetilde{v}^{\pi^*}; \Delta^*, \Theta^*)\right| \\
&\quad \le 2 \max_{w\in \mathsf{conf}^{\text{mis}}_{\alpha_{\text{mis}}}(\Delta^*,\Theta^*,\pi^*)} \max_{v\in\mathcal{V}} \left|\Phi_{\text{mis}}^{\pi^*}(w, v; \Delta^*, \Theta^*)\right| \\
&\qquad + \max_{w\in\mathcal{W}} \left|\Phi_{\text{mis}}^{\pi^*}(w, V^{\pi^*}; \Delta^*, \Theta^*) - \Phi_{\text{mis}}^{\pi^*}(w, \widetilde{v}^{\pi^*}; \Delta^*, \Theta^*)\right| \\
&\quad \le c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log\frac{1}{\delta} \log(NT)} \\
&\qquad + 2 \max_{f\in\mathcal{V}} \Phi_{\text{mis}}^{\pi^*}(\widetilde{w}^{\pi^*}, f; \Delta^*, \Theta^*) + C_* C_{\Delta^*} C_{\Theta^*} \varepsilon_{\text{mis}}^{\mathcal{V}}, \qquad \text{(C.5.20)}
\end{aligned}
$$

where in the first inequality, we use triangle inequality; in the second inequality, we use the definition of $\widetilde{v}^{\pi^*}$ that $\widetilde{v}^{\pi^*} \in \mathcal{V}$; in the last inequality, we use Lemma C.5.6 and Assumption 4.4.15. In the meanwhile, by Assumption 4.4.15, we have

$$
\begin{aligned}
&\max_{f\in\mathcal{V}} \Phi_{\text{mis}}^{\pi^*}(\widetilde{w}^{\pi}, f; \Delta^*, \Theta^*) \\
&\quad \le \max_{f\in\mathcal{V}} \Phi_{\text{mis}}^{\pi^*}(\widetilde{w}^{\pi^*}, f; \Delta^*, \Theta^*) - \max_{f\in\mathcal{V}} \Phi_{\text{mis}}^{\pi^*}(w^{\pi^*}, f; \Delta^*, \Theta^*) + \max_{f\in\mathcal{V}} \Phi_{\text{mis}}^{\pi^*}(w^{\pi^*}, f; \Delta^*, \Theta^*) \\
&\quad = \max_{f\in\mathcal{V}} \Phi_{\text{mis}}^{\pi^*}(\widetilde{w}^{\pi^*}, f; \Delta^*, \Theta^*) - \max_{f\in\mathcal{V}} \Phi_{\text{mis}}^{\pi^*}(w^{\pi^*}, f; \Delta^*, \Theta^*) \\
&\quad \le \max_{f\in\mathcal{V}} \left|\Phi_{\text{mis}}^{\pi^*}(\widetilde{w}^{\pi^*}, f; \Delta^*, \Theta^*) - \Phi_{\text{mis}}^{\pi^*}(w^{\pi^*}, f; \Delta^*, \Theta^*)\right| \\
&\text{(C.5.21)} \\
&\quad \le C_{\Delta^*} C_{\Theta^*} \varepsilon_{\text{mis}}^{\mathcal{W}} / (1-\gamma).
\end{aligned}
$$

Now, by plugging (C.5.20) and (C.5.21) into (C.5.19), we have

$$J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \leq c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{NT}{\delta}} + 3C_{\Delta^*}C_{\Theta^*}\left(C_*\varepsilon_{\text{mis}}^{\mathcal{V}} + \varepsilon_{\text{mis}}^{\mathcal{W}}/(1-\gamma)\right). \tag{C.5.22}$$

By combining (C.5.14) and (C.5.22), we have

$$J(\pi^*) - J(\widehat{\pi}_{\mathsf{dr}}) \leq c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{NT}{\delta}} + 3C_{\Delta^*}C_{\Theta^*}\min\left\{C_*\varepsilon_{\mathsf{vf}}^{\mathcal{V}} + \varepsilon_{\mathsf{vf}}^{\mathcal{W}}/(1-\gamma),\ C_*\varepsilon_{\text{mis}}^{\mathcal{V}} + \varepsilon_{\text{mis}}^{\mathcal{W}}/(1-\gamma)\right\}$$

which conclude the proof. □

C.5.3. Proof of Lemma C.5.1

Proof. By Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w, v, \Delta, \Theta, \pi) \in \mathcal{W} \times \mathcal{V} \times \mathcal{F}_0 \times \mathcal{F}_1 \times \Pi$ that

$$\left|L_{\mathsf{dr}}(w, v, \pi; \Delta, \Theta) - \widehat{L}_{\mathsf{dr}}(w, v, \pi; \Delta, \Theta)\right| \leq c \cdot \frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\log\frac{1}{\delta}\log(NT)},$$

which concludes the proof of the lemma. □

C.5.4. Proof of Lemma C.5.2

Proof. With a slight abuse of notations, we denote by

$$(w_0, v_0) \in \underset{(w,v)\in \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)}{\operatorname{argmin}} L_{\mathsf{dr}}(w, v, \pi^*; \Delta^*, \Theta^*),$$

$$(w_1, v_1) \in \underset{(w,v)\in \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)}{\operatorname{argmin}} L_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta).$$

Then we have

$$\begin{aligned}
&\left| \min_{(w,v)\in \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta^*, \Theta^*) - \min_{(w,v)\in \mathsf{conf}_{\alpha_{\mathrm{mis}},\alpha_{\mathsf{vf}}}(\Delta,\Theta,\pi^*)} L_{\mathsf{dr}}(w, v, \pi^*; \Delta, \Theta) \right| \\
&\quad = \left| L_{\mathsf{dr}}(w_0, v_0, \pi^*; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w_1, v_1, \pi^*; \Delta, \Theta) \right| \\
&\quad \le \left| L_{\mathsf{dr}}(w_0, v_0, \pi^*; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w^{\pi^*}, v_0, \pi^*; \Delta^*, \Theta^*) \right| \\
&\qquad + \left| L_{\mathsf{dr}}(w^{\pi}, v_1, \pi^*; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w_1, v_1, \pi^*; \Delta^*, \Theta^*) \right| \\
&\qquad + \left| L_{\mathsf{dr}}(w_1, v_1, \pi; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w_1, v_1, \pi; \Delta, \Theta) \right|
\end{aligned}$$

(C.5.23)

$$\begin{aligned}
&\quad = \underbrace{\left| \Phi^{\pi^*}_{\mathrm{mis}}(w_0, V^{\pi^*}; \Delta^*, \Theta^*) - \Phi^{\pi^*}_{\mathrm{mis}}(w_0, v_0; \Delta^*, \Theta^*) \right|}_{\text{Term (I)}} \\
&\qquad + \underbrace{\left| \Phi^{\pi^*}_{\mathrm{mis}}(w_1, V^{\pi^*}; \Delta^*, \Theta^*) - \Phi^{\pi^*}_{\mathrm{mis}}(w_1, v_1; \Delta^*, \Theta^*) \right|}_{\text{Term (II)}} \\
&\qquad + \underbrace{\left| L_{\mathsf{dr}}(w_1, v_1, \pi^*; \Delta^*, \Theta^*) - L_{\mathsf{dr}}(w_1, v_1, \pi^*; \Delta, \Theta) \right|}_{\text{Term (III)}},
\end{aligned}$$

where in the first inequality, we use triangle inequality and the fact that $L_{\mathsf{dr}}(w^{\pi^*}, v, \pi^*; \Delta^*, \Theta^*) = J(\pi^*)$ for any function v; while in the last equality, we use the following equality for any

(w,v),

$$L_{\mathsf{dr}}(w,v,\pi^*;\Delta^*,\Theta^*) - L_{\mathsf{dr}}(w^{\pi^*},v,\pi^*;\Delta^*,\Theta^*) = -\Phi^{\pi^*}_{\text{mis}}(w,V^{\pi^*};\Delta^*,\Theta^*) + \Phi^{\pi^*}_{\text{mis}}(w,v;\Delta^*,\Theta^*).$$

We upper bound terms (I), (II), and (III) on the RHS of (C.5.23), respectively.

Upper Bounding Term (I). Note that with probability at least $1-\delta$, we have

$$\begin{aligned}\left|\Phi^{\pi^*}_{\text{mis}}(w_0,V^{\pi^*};\Delta^*,\Theta^*)\right| &\le \max_{f\in\mathcal{V}}\left|\Phi^{\pi^*}_{\text{mis}}(w_0,f;\Delta^*,\Theta^*)\right| \\ &= \max_{f\in\mathcal{V}}\max\left\{\Phi^{\pi^*}_{\text{mis}}(w_0,f;\Delta^*,\Theta^*), -\Phi^{\pi^*}_{\text{mis}}(w_0,f;\Delta^*,\Theta^*)\right\} \\ &= \max_{f\in\mathcal{V}}\max\left\{\Phi^{\pi^*}_{\text{mis}}(w_0,f;\Delta^*,\Theta^*), \Phi^{\pi^*}_{\text{mis}}(w_0,-f;\Delta^*,\Theta^*)\right\} \\ &= \max_{f\in\mathcal{V}}\Phi^{\pi^*}_{\text{mis}}(w_0,f;\Delta^*,\Theta^*) \\ &\le c\cdot\frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V}}\log\frac{1}{\delta}\log(NT)},\end{aligned}$$

where in the first inequality, we use the fact that $V^{\pi^*}\in\mathcal{V}$; in the third equality, we use the fact that $\mathcal{V}$ is symmetric; in the last inequality, by noting that $w_0\in\cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\mathsf{conf}^{\pi}_{\alpha_{\text{mis}}}(\Delta,\Theta)$, we use Lemma 4.4.12. Similarly, we have

$$\left|\Phi^{\pi^*}_{\text{mis}}(w_0,v_0;\Delta^*,\Theta^*)\right| \le c\cdot\frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V}}\log\frac{1}{\delta}\log(NT)},$$

which implies that

$$\text{(C.5.24)}\qquad \text{Term (I)} \le c\cdot\frac{C^2_{\Delta^*}C^2_{\Theta^*}C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V}}\log\frac{1}{\delta}\log(NT)}.$$

Upper Bounding Term (II). Similar to (C.5.24), with probability at least $1-\delta$, we have

$$\text{Term (II)} \leq c \cdot \frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma}(\xi_0+\xi_1)\sqrt{\frac{1}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V}}\log\frac{1}{\delta}\log(NT)}. \tag{C.5.25}$$

Upper Bounding Term (III). Note that with probability at least $1-\delta$, it holds for any $(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}$ that

$$\begin{aligned}
&\text{Term (III)}\\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^\top A_t\pi^*(A_t\,|\,S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)} - \frac{Z_t^\top A_t\pi^*(A_t\,|\,S_t)}{\Delta(S_t,A_t)\Theta(S_t,Z_t)}\right) w_1(S_t)\left(R_t+\gamma v_1(S_{t+1})-v_1(S_t)\right)\right]\\
&\text{(C.5.26)}\\
&\leq \frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\left(\xi_0 C_{\Theta^*}\sqrt{\frac{C_{\Delta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0}\log\frac{1}{\delta}\log(NT)} + \xi_1 C_{\Delta^*}\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_1}\log\frac{1}{\delta}\log(NT)}\right),
\end{aligned}$$

where we use triangle inequality and Assumption 4.4.4 in the last inequality.

Now, by plugging (C.5.24), (C.5.25), and (C.5.26) into (C.5.23), we conclude the proof of the lemma. □

C.5.5. Proof of Lemma C.5.3

Proof. By the definition of $\widetilde{v}^\pi$ in (4.4.2), we know that $\widetilde{v}^\pi \in \mathcal{V}$. Thus, to show that $\widetilde{v}^\pi \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*,\Theta^*,\pi)$ with a high probability, it suffices to show that

$$\max_{g\in\mathcal{W}}\widehat{\Phi}^\pi_{\mathsf{vf}}(\widetilde{v}^\pi, g;\Delta^*,\Theta^*) - \max_{g\in\mathcal{W}}\widehat{\Phi}^\pi_{\mathsf{vf}}(\widehat{v}^\pi_{\Delta^*,\Theta^*}, g;\Delta^*,\Theta^*) \leq \alpha_{\mathsf{vf}}. \tag{C.5.27}$$

In the follows, we show that (C.5.27) holds with a high probability. For the simplicity of notations, we denote by $\Phi^{\pi}_{\mathsf{vf}}(v, g; *) = \Phi^{\pi}_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*)$ and $\widehat{v}^{\pi}_{*} = \widehat{v}^{\pi}_{\Delta^*,\Theta^*}$ for any (π, v, g). Note that

$$\begin{aligned}
&\max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; *) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{*}, g; *) \\
&\quad = \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; *) - \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; *) + \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; *) - \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{*}, g; *) \\
&\qquad\quad + \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{*}, g; *) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{*}, g; *) \\
&\quad \le \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; *) - \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; *) + \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{*}, g; *) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{*}, g; *) \\
&\quad \le 2 \max_{v\in\mathcal{V}} \left| \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; *) - \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v, g; *) \right| \\
&\text{(C.5.28)} \\
&\quad \le 2 \max_{v\in\mathcal{V}} \max_{g\in\mathcal{W}} \left| \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; *) - \Phi^{\pi}_{\mathsf{vf}}(v, g; *) \right|,
\end{aligned}$$

where in the first inequality, we use the fact that $\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; *) \le \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v, g; *)$ for any v due to the definition of $\widetilde{v}^{\pi}$ in (4.4.2); while in the second inequality, we use the fact that $\widetilde{v}^{\pi}, \widehat{v}^{\pi}_{*} \in \mathcal{V}$. In the meanwhile, by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(\pi, v, g) \in \Pi \times \mathcal{V} \times \mathcal{W}$ that

$$\left| \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; *) - \Phi^{\pi}_{\mathsf{vf}}(v, g; *) \right| \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{W},\mathcal{V},\Pi}}{NT\kappa} \cdot \log\frac{1}{\delta} \log(NT)}, \tag{C.5.29}$$

where we use Assumption (b) and $\|g\|_\infty \le C_*$ for any $g \in \mathcal{W}$. Now, combining (C.5.28) and (C.5.29), with probability at least $1-\delta$, we have

$$\max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; *) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widehat{v}^{\pi}_{*}, g; *) \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{W},\mathcal{V},\Pi}}{NT\kappa} \cdot \log\frac{1}{\delta} \log(NT)} = \alpha_{\mathsf{vf}},$$

which implies that $\widetilde{v}^{\pi} \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta^*, \Theta^*, \pi)$ for any $\pi \in \Pi$. This concludes the proof of the lemma. □

C.5.6. Proof of Lemma C.5.4

Proof. Since $v \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi)$, there exists a pair $(\widetilde{\Delta}, \widetilde{\Theta}) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ such that $v \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$. For the simplicity of notations, we denote by

$$v^\dagger \in \operatorname*{argmin}_{v\in\mathcal{V}} \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta}),$$

i.e., $v^\dagger = \widehat{v}^{\pi}_{\widetilde{\Delta},\widetilde{\Theta}}$, which is defined in (4.3.2). By the definition of $v^\dagger$ and $v \in \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$, we know that

$$\max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta}) - \max_{g\in\mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v^\dagger, g; \widetilde{\Delta}, \widetilde{\Theta}) \le \alpha_{\mathsf{vf}}. \tag{C.5.30}$$

Note that

$$\begin{aligned}
&\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) \\
&\qquad \le 2 \underbrace{\max_{(v,g,\Delta,\Theta)\in(\mathcal{V},\mathcal{W},\mathcal{F}_0,\mathcal{F}_1)} \left|\Phi^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta)\right|}_{\text{Term (I)}} + \underbrace{\max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v^\dagger, g; \widetilde{\Delta}, \widetilde{\Theta})}_{\text{Term (II)}} \\
&\qquad\qquad + \underbrace{\max_{g\in\mathcal{W}} \left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta})\right|}_{\text{Term (III)}} + \alpha_{\mathsf{vf}},
\end{aligned} \tag{C.5.31}$$

where we use (C.5.30) in the last inequality. Now we upper bound terms (I), (II), and (III) on the RHS of (C.5.31).

Upper Bounding Term (I). By Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(v, g, \Delta, \Theta, \pi) \in (\mathcal{V}, \mathcal{W}, \mathcal{F}_0, \mathcal{F}_1, \Pi)$ that

$$\left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta) - \Phi^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta)\right| \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi}}{NT\kappa} \log\frac{1}{\delta} \log(NT)},$$

which implies that with probability at least $1-\delta$, we have

$$\text{Term (I)} \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi}}{NT\kappa} \log\frac{1}{\delta} \log(NT)}. \tag{C.5.32}$$

Upper Bounding Term (II). Recall that $\widetilde{v}^{\pi} \in \operatorname{argmin}_{v \in \mathcal{V}} \max_{w \in \mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v, w; \Delta^*, \Theta^*)$. Note that

$$\begin{aligned}
&\max_{g \in \mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(v^{\dagger}, g; \widetilde{\Delta}, \widetilde{\Theta}) \\
&\quad \le 2 \max_{v \in \mathcal{V}} \max_{g \in \mathcal{W}} \left|\Phi^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta}) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta})\right| \\
&\qquad + \max_{g \in \mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v^{\dagger}, g; \widetilde{\Delta}, \widetilde{\Theta}) - \max_{g \in \mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; \widetilde{\Delta}, \widetilde{\Theta}) + \max_{g \in \mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; \widetilde{\Delta}, \widetilde{\Theta}) \\
&\quad \le 2 \max_{v \in \mathcal{V}} \max_{g \in \mathcal{W}} \left|\Phi^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta}) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta})\right| \qquad \text{(C.5.33)} \\
&\qquad + \max_{g \in \mathcal{W}} \left|\Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; \widetilde{\Delta}, \widetilde{\Theta}) - \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; \Delta^*, \Theta^*)\right| + \max_{g \in \mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; \Delta^*, \Theta^*),
\end{aligned}$$

where we use triangle inequality and the fact that $v^{\dagger} \in \operatorname{argmin}_{v \in \mathcal{V}} \max_{g \in \mathcal{W}} \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta, \Theta)$ in the last inequality. In the meanwhile, by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(v, g, \pi) \in \mathcal{V} \times \mathcal{W} \times \Pi$ that

$$\left|\Phi^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta}) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta})\right| \leq c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{\mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi}}{NT\kappa} \log\frac{1}{\delta} \log(NT)}. \tag{C.5.34}$$

Also, we upper bound the second term on the RHS of (C.5.33) with probability at least $1-\delta$ by a similar argument as in (C.3.14),

$$\begin{aligned} &\left|\Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; \widetilde{\Delta}, \widetilde{\Theta}) - \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; \Delta^*, \Theta^*)\right| \\ &\leq \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \left(\xi_0 C_{\Theta^*} \sqrt{\frac{C_{\Delta^*}}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0} \log\frac{1}{\delta} \log(NT)} + \xi_1 C_{\Delta^*} \sqrt{\frac{C_{\Theta^*}}{NT\kappa} \mathfrak{C}_{\mathcal{F}_1} \log\frac{1}{\delta} \log(NT)}\right), \end{aligned} \tag{C.5.35}$$

where in the first inequality, we use the fact that $\|v\|_\infty \leq 1/(1-\gamma)$ and $\|g\|_\infty \leq C_*$; in the third inequality, we use Cauchy Schwarz inequality; while in the last inequality, we use Assumption 4.4.4 with $(\Delta, \Theta) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$. Now, by plugging (C.5.34) and (C.5.35) into (C.5.33), with probability at least $1-\delta$, it holds for any $(\Delta, \Theta, \pi) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1} \times \Pi$ that

$$\text{Term (II)} \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma} (\xi_0 + \xi_1) \sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}}{NT\kappa} \log\frac{1}{\delta} \log(NT)} + \max_{g\in\mathcal{W}} \Phi^{\pi}_{\mathsf{vf}}(\widetilde{v}^{\pi}, g; \Delta^*, \Theta^*). \tag{C.5.36}$$

Upper Bounding Term (III). Note that

$$
\begin{aligned}
&\left|\widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) - \widehat{\Phi}^{\pi}_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta})\right| \\
&\leq \left|\left(\widehat{\mathbb{E}} - \mathbb{E}\right)\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(R_t + \gamma v(S_{t+1}))\right]\right|
\end{aligned}
$$

(C.5.37)

$$
\quad + \left|\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(R_t + \gamma v(S_{t+1}))\right]\right|.
$$

For the first term on the RHS of (C.3.12), by Theorem C.7.6, with probability at least $1 - \delta$, it holds for any $(v, g, \pi) \in \mathcal{V} \times \mathcal{W} \times \Pi$ that

$$
\left|\left(\widehat{\mathbb{E}} - \mathbb{E}\right)\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(R_t + \gamma v(S_{t+1}))\right]\right|
$$

(C.5.38)

$$
\leq c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1 - \gamma}\sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi}}{NT\kappa} \log\frac{1}{\delta}\log(NT)}.
$$

For the second term on the RHS of (C.5.37), by a similar argument as in (C.3.14), with probability at least $1 - \delta$, it holds that

$$
\left|\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t)\left(\frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^\top A_t \pi(A_t \,|\, S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(R_t + \gamma v(S_{t+1}))\right]\right|
$$

(C.5.39)

$$
\leq \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1 - \gamma}\left(\xi_0 C_{\Theta^*}\sqrt{\frac{C_{\Delta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0}\log\frac{1}{\delta}\log(NT)} + \xi_1 C_{\Delta^*}\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_1}\log\frac{1}{\delta}\log(NT)}\right),
$$

where in the first inequality, we use the fact that $\|v\|_\infty \leq 1/(1-\gamma)$ and $\|g\|_\infty \leq C_*$; in the third inequality, we use Cauchy Schwarz inequality; while in the last inequality, we use

Assumption 4.4.4 with the fact that $(\widetilde{\Delta}, \widetilde{\Theta}) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$. Now, by plugging (C.5.38) and (C.5.39) into (C.5.37), it holds with probability at least $1 - \delta$ that

$$\begin{aligned} &\Big|\widehat{\Phi}^\pi_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) - \widehat{\Phi}^\pi_{\mathsf{vf}}(v, g; \widetilde{\Delta}, \widetilde{\Theta})\Big| \\ &\quad \le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \cdot \log\frac{1}{\delta}\log(NT)}. \end{aligned} \tag{C.5.40}$$

Now, by plugging (C.5.32), (C.5.36), and (C.5.40) into (C.5.31), with probability at least $1 - \delta$, it holds for any $v \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathsf{vf}}_{\alpha_{\mathsf{vf}}}(\Delta, \Theta, \pi)$ and $\pi \in \Pi$ that

$$\begin{aligned} \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(v, g; \Delta^*, \Theta^*) &\le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \cdot \log\frac{1}{\delta}\log(NT)} \\ &\quad + \max_{g\in\mathcal{W}} \Phi^\pi_{\mathsf{vf}}(\widetilde{v}^\pi, g; \Delta^*, \Theta^*), \end{aligned}$$

which concludes the proof of the lemma. □

C.5.7. Proof of Lemma C.5.5

Proof. First, by the definition of $\widetilde{w}^\pi$ in (4.4.2), we know that $\widetilde{w}^\pi \in \mathcal{W}$. For notation simplicity, we denote by $\Phi^\pi_{\mathrm{mis}}(w, f; *) = \Phi^\pi_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*)$ and $\widehat{w}^\pi_* = \widehat{w}^\pi_{\Delta^*,\Theta^*}$ for any

(π, w, f). Note that

$$\begin{aligned}
&\max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; *) - \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *) \\
&\quad = \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; *) - \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; *) + \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; *) - \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *) \\
&\qquad + \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *) - \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *) \\
&\quad \le \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; *) - \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; *) + \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *) - \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *) \\
&\quad \le 2 \max_{w\in\mathcal{W}} \left| \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; *) - \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w, f; *) \right| \\
&\quad \le 2 \max_{w\in\mathcal{W}} \max_{f\in\mathcal{V}} \left| \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; *) - \Phi^{\pi}_{\mathrm{mis}}(w, f; *) \right|, \qquad \text{(C.5.41)}
\end{aligned}$$

where in the first inequality, we use the fact that $\max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; *) \le \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *)$ by the definition of $\widetilde{w}^{\pi}$ in (4.4.2); while in the second inequality, we use the fact that $\widetilde{w}^{\pi}, \widehat{w}^{\pi}_{*} \in \mathcal{W}$. In the meanwhile, by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w, f, \pi) \in \mathcal{W} \times \mathcal{V} \times \Pi$ that

$$\left| \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; *) - \Phi^{\pi}_{\mathrm{mis}}(w, f; *) \right| \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi} \log\frac{1}{\delta} \log(NT)}, \tag{C.5.42}$$

where we use Assumption (b). Now, combining (C.5.41) and (C.5.42), with probability at least $1-\delta$, we have

$$\begin{aligned}
&\max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; *) - \max_{f\in\mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widehat{w}^{\pi}_{*}, f; *) \\
&\quad \le c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi} \log\frac{1}{\delta} \log(NT)} = \alpha_{\mathrm{mis}},
\end{aligned}$$

which implies that $\widetilde{w}^\pi \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta^*, \Theta^*, \pi)$. This concludes the proof of the lemma. □

C.5.8. Proof of Lemma C.5.6

Proof. Since $w \in \cup_{(\Delta,\Theta)\in\mathsf{conf}^0_{\alpha_0}\times\mathsf{conf}^1_{\alpha_1}}\mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi)$, there exists a pair $(\widetilde{\Delta}, \widetilde{\Theta}) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$ such that $w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$. For the simplicity of notations, we denote by

$$w^\dagger \in \operatorname*{argmin}_{w\in\mathcal{W}} \max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta}),$$

i.e., $w^\dagger = \widehat{w}^\pi_{\widetilde{\Delta},\widetilde{\Theta}}$, which is defined in (4.3.5). By the definition of $w^\dagger$ and $w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$, with probability at least $1 - \delta$, it holds for any $\pi \in \Pi$ and $w \in \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\widetilde{\Delta}, \widetilde{\Theta}, \pi)$ that

$$\max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta}) - \max_{f\in\mathcal{V}} \widehat{\Phi}^\pi_{\mathrm{mis}}(w^\dagger, f; \widetilde{\Delta}, \widetilde{\Theta}) \le \alpha_{\mathrm{mis}}. \tag{C.5.43}$$

Further, we observe that

$$\begin{aligned}
&\max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) \\
&\quad \le \underbrace{\max_{(w,f,\Delta,\Theta)\in(\mathcal{W},\mathcal{V},\mathcal{F}_0,\mathcal{F}_1)} \left|\Phi^\pi_{\mathrm{mis}}(w, f; \Delta, \Theta) - \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \Delta, \Theta)\right|}_{\text{Term (I)}} + \underbrace{\max_{f\in\mathcal{V}} \Phi^\pi_{\mathrm{mis}}(w^\dagger, f; \widetilde{\Delta}, \widetilde{\Theta})}_{\text{Term (II)}} \\
&\qquad + \underbrace{\max_{f\in\mathcal{V}} \left|\widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) - \widehat{\Phi}^\pi_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta})\right|}_{\text{Term (III)}} + \alpha_{\mathrm{mis}},
\end{aligned} \tag{C.5.44}$$

where we use (C.5.43) in the last inequality. Now we upper bound terms (I), (II), and (III) on the RHS of (C.5.44).

Upper Bounding Term (I). By Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w, f, \Delta, \Theta, \pi) \in (\mathcal{W}, \mathcal{V}, \mathcal{F}_0, \mathcal{F}_1, \Pi)$ that

$$\left|\widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta) - \Phi^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta)\right| \leq c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \log \frac{1}{\delta} \log(NT)},$$

which implies that with probability at least $1-\delta$, we have

$$\text{Term (I)} \leq c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{F}_0, \mathcal{F}_1, \mathcal{W}, \mathcal{V}, \Pi} \log \frac{1}{\delta} \log(NT)}. \tag{C.5.45}$$

Upper Bounding Term (II). Recall that $\widetilde{w}^{\pi} \in \operatorname{argmin}_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w, v; \Delta^*, \Theta^*)$. Note that

$$\begin{aligned}
&\max_{f \in \mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w^{\dagger}, f; \widetilde{\Delta}, \widetilde{\Theta}) \\
&\quad = \max_{f \in \mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w^{\dagger}, f; \widetilde{\Delta}, \widetilde{\Theta}) - \max_{f \in \mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w^{\dagger}, f; \widetilde{\Delta}, \widetilde{\Theta}) + \max_{f \in \mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w^{\dagger}, f; \widetilde{\Delta}, \widetilde{\Theta}) \\
&\qquad - \max_{f \in \mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \widetilde{\Delta}, \widetilde{\Theta}) + \max_{f \in \mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \widetilde{\Delta}, \widetilde{\Theta}) - \max_{f \in \mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \widetilde{\Delta}, \widetilde{\Theta}) \\
&\qquad + \max_{f \in \mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \widetilde{\Delta}, \widetilde{\Theta}) \\
&\quad \leq 2 \max_{w \in \mathcal{W}} \max_{f \in \mathcal{V}} \left| \Phi^{\pi}_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta}) - \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta}) \right| \\
&\qquad + \max_{f \in \mathcal{V}} \left| \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \widetilde{\Delta}, \widetilde{\Theta}) - \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \Delta^*, \Theta^*) \right| + \max_{f \in \mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \Delta^*, \Theta^*),
\end{aligned} \tag{C.5.46}$$

where we use triangle inequality and the fact that $w^{\dagger} \in \operatorname{argmin}_{w \in \mathcal{W}} \max_{f \in \mathcal{V}} \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \Delta, \Theta)$ in the last inequality. In the meanwhile, by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w, f, \pi) \in \mathcal{W} \times \mathcal{V} \times \Pi$ that

$$\left|\Phi^{\pi}_{\text{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta}) - \widehat{\Phi}^{\pi}_{\text{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta})\right| \leq c \cdot \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma} \sqrt{\frac{1}{NT\kappa} \mathfrak{C}_{\mathcal{V},\mathcal{W},\Pi} \log \frac{1}{\delta} \log(NT)}. \tag{C.5.47}$$

Also, we upper bound the second term on the RHS of (C.5.46) with probability at least $1-\delta$ as follows,

$$\begin{aligned}
&\left|\Phi^{\pi}_{\text{mis}}(\widetilde{w}^{\pi}, f; \widetilde{\Delta}, \widetilde{\Theta}) - \Phi^{\pi}_{\text{mis}}(\widetilde{w}^{\pi}, f; \Delta^*, \Theta^*)\right| \\
&= \left|\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^{\top} A_t \pi(A_t \mid S_t) w(S_t)}{\Delta^*(S_t, A_t)\Theta^*(S_t, Z_t)} - \frac{Z_t^{\top} A_t \pi(A_t \mid S_t) w(S_t)}{\widetilde{\Delta}(S_t, A_t)\widetilde{\Theta}(S_t, Z_t)}\right)(f(S_t) - \gamma f(S_{t+1}))\right]\right| \\
&\leq \frac{C_{\Delta^*} C_{\Theta^*} C_*}{1-\gamma}\left(\xi_0 C_{\Theta^*}\sqrt{\frac{C_{\Delta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0} \log\frac{1}{\delta}\log(NT)} + \xi_1 C_{\Delta^*}\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_1} \log\frac{1}{\delta}\log(NT)}\right), \qquad \text{(C.5.48)}
\end{aligned}$$

where we use Cauchy-Schwarz inequality and Assumption 4.4.4 in the last inequality. Now, by plugging (C.5.47) and (C.5.48) into (C.5.46), it holds with probability at least $1-\delta$ that

$$\text{Term (II)} \leq c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}}{NT\kappa}\log\frac{1}{\delta}\log(NT)} + \max_{f \in \mathcal{V}} \Phi^{\pi}_{\text{mis}}(\widetilde{w}^{\pi}, f; \Delta^*, \Theta^*). \tag{C.5.49}$$

Upper Bounding Term (III). Note that

$$
\begin{aligned}
&\left|\widehat{\Phi}^{\pi}_{\mathrm{mis}}(w,f;\Delta^*,\Theta^*)-\widehat{\Phi}^{\pi}_{\mathrm{mis}}(w,f;\widetilde{\Delta},\widetilde{\Theta})\right|\\
&\le \left|\left(\widehat{\mathbb{E}}-\mathbb{E}\right)\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)w(S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}-\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)w(S_t)}{\widetilde{\Delta}(S_t,A_t)\widetilde{\Theta}(S_t,Z_t)}\right)(f(S_t)-\gamma f(S_{t+1}))\right]\right|
\end{aligned}
$$

(C.5.50)

$$
\quad+\left|\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)w(S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}-\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)w(S_t)}{\widetilde{\Delta}(S_t,A_t)\widetilde{\Theta}(S_t,Z_t)}\right)(f(S_t)-\gamma f(S_{t+1}))\right]\right|.
$$

For the first term on the RHS of (C.5.50), by Theorem C.7.6, with probability at least $1-\delta$, it holds for any $(w,f,\pi)\in\mathcal{W}\times\mathcal{V}\times\Pi$ that

$$
\left|\left(\widehat{\mathbb{E}}-\mathbb{E}\right)\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)w(S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}-\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)w(S_t)}{\widetilde{\Delta}(S_t,A_t)\widetilde{\Theta}(S_t,Z_t)}\right)(f(S_t)-\gamma f(S_{t+1}))\right]\right|
$$

(C.5.51)

$$
\le c\cdot\frac{C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\sqrt{\frac{\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}}{NT\kappa}\log\frac{1}{\delta}\log(NT)}.
$$

For the second term on the RHS of (C.5.50), by a similar argument as in (C.3.14), it holds with probability at least $1-\delta$ that

$$
\left|\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)w(S_t)}{\Delta^*(S_t,A_t)\Theta^*(S_t,Z_t)}-\frac{Z_t^\top A_t\pi(A_t\,|\,S_t)w(S_t)}{\widetilde{\Delta}(S_t,A_t)\widetilde{\Theta}(S_t,Z_t)}\right)(f(S_t)-\gamma f(S_{t+1}))\right]\right|
$$

(C.5.52)

$$
\le\frac{2C_{\Delta^*}C_{\Theta^*}C_*}{1-\gamma}\left(\xi_0 C_{\Theta^*}\sqrt{\frac{C_{\Delta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_0}\log\frac{1}{\delta}\log(NT)}+\xi_1 C_{\Delta^*}\sqrt{\frac{C_{\Theta^*}}{NT\kappa}\mathfrak{C}_{\mathcal{F}_1}\log\frac{1}{\delta}\log(NT)}\right),
$$

where in the first inequality, we use the fact that $\|f\|_\infty\le 1/(1-\gamma)$ and $\|w\|_\infty\le C_*$; in the third inequality, we use Cauchy Schwarz inequality; while in the last inequality, we use Assumption 4.4.4 with the fact that $(\widetilde{\Delta}, \widetilde{\Theta}) \in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}$. Now, by plugging (C.5.51) and (C.5.52) into (C.5.50), with probability at least $1-\delta$, it holds for any $w \in \cup_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi)$ and $(f,\pi) \in \mathcal{V} \times \Pi$ that

$$\begin{aligned}&\left|\widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) - \widehat{\Phi}^{\pi}_{\mathrm{mis}}(w, f; \widetilde{\Delta}, \widetilde{\Theta})\right| \\ &\qquad \le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log\frac{1}{\delta}\log(NT)}. \end{aligned} \tag{C.5.53}$$

Now, by plugging (C.5.45), (C.5.49), and (C.5.53) into (C.5.44), with probability at least $1-\delta$, it holds for any $\pi \in \Pi$ and $w \in \cup_{(\Delta,\Theta)\in \mathsf{conf}^0_{\alpha_0} \times \mathsf{conf}^1_{\alpha_1}} \mathsf{conf}^{\mathrm{mis}}_{\alpha_{\mathrm{mis}}}(\Delta, \Theta, \pi)$ that

$$\begin{aligned}\max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(w, f; \Delta^*, \Theta^*) &\le c \cdot \frac{C^2_{\Delta^*} C^2_{\Theta^*} C_*}{1-\gamma}(\xi_0 + \xi_1)\sqrt{\frac{1}{NT\kappa} \cdot \mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi} \cdot \log\frac{1}{\delta}\log(NT)} \\ &\qquad + \max_{f\in\mathcal{V}} \Phi^{\pi}_{\mathrm{mis}}(\widetilde{w}^{\pi}, f; \Delta^*, \Theta^*),\end{aligned}$$

which concludes the proof of the lemma. □

C.6. Proof of Results in §4.5

C.6.1. Proof of Theorem 4.5.1

Proof. We denote by $(\widehat{\pi}, \widehat{\lambda}, \widehat{v})$ the solution of (4.5.1). Note that $(1-\gamma)\mathbb{E}_{S_0\sim\nu}[v(S_0)]$ is a lower bounded real-valued convex functional w.r.t. v, and $\widehat{M}^{\pi^*}_{\mathsf{vf}}$ is also a convex functional. In the meanwhile, we have $\widehat{M}^{\pi^*}_{\mathsf{vf}}(\widehat{v}^{\pi^*}_{\Delta^*,\Theta^*}) = 0$. Thus, by Theorem 1 of §8.6 in

Luenberger (1997), strong duality holds, i.e.,

$$\begin{aligned}
&\max_{\lambda\geq 0}\min_{v\in\mathcal{V}}\left\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right]+\lambda\cdot\left(\widehat{M}^{\pi^*}_{\mathsf{vf}}(v)-\alpha_{\mathsf{vf}}\right)\right\}\\
&\quad=\min_{v\in\mathcal{V}}\max_{\lambda\geq 0}\left\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right]+\lambda\cdot\left(\widehat{M}^{\pi^*}_{\mathsf{vf}}(v)-\alpha_{\mathsf{vf}}\right)\right\}. \qquad \text{(C.6.1)}
\end{aligned}$$

By Lemma 4.4.7, it holds with probability at least $1-\delta$ that $\widehat{M}^{\pi}_{\mathsf{vf}}(V^{\widehat{\pi}})\leq\alpha_{\mathsf{vf}}$. Thus, with probability at least $1-\delta$, we have

$$\begin{aligned}
&J(\pi^*)-J(\widehat{\pi})\\
&\quad=(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)-V^{\widehat{\pi}}(S_0)\right]\\
&\quad\leq(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right]-\left((1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\widehat{\pi}}(S_0)\right]+\widehat{\lambda}\cdot\left(\widehat{M}^{\widehat{\pi}}_{\mathsf{vf}}(V^{\widehat{\pi}})-\alpha_{\mathsf{vf}}\right)\right)\\
&\quad\leq(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right]-\min_{v\in\mathcal{V}}\left\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right]+\widehat{\lambda}\cdot\left(\widehat{M}^{\widehat{\pi}}_{\mathsf{vf}}(v)-\alpha_{\mathsf{vf}}\right)\right\}\\
&\quad=(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right]-\max_{\pi\in\Pi}\max_{\lambda\geq 0}\min_{v\in\mathcal{V}}\left\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right]+\lambda\cdot\left(\widehat{M}^{\pi}_{\mathsf{vf}}(v)-\alpha_{\mathsf{vf}}\right)\right\}\\
&\text{(C.6.2)}\\
&\quad=(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right]-\max_{\lambda\geq 0}\min_{v\in\mathcal{V}}\left\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right]+\lambda\cdot\left(\widehat{M}^{\pi^*}_{\mathsf{vf}}(v)-\alpha_{\mathsf{vf}}\right)\right\},
\end{aligned}$$

where in the second inequality, we use the fact that $V^{\widehat{\pi}}\in\mathcal{V}$; in the third inequality, we use the definition of $\widehat{\pi}$ and $\widehat{\lambda}$; in the last inequality, we use the fact that $\pi^*\in\Pi$. By combining (C.6.1) and (C.6.2), we have

$$J(\pi^*) - J(\widehat{\pi})$$

$$\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right] - \min_{v\in\mathcal{V}}\max_{\lambda\geq 0}\left\{(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right] + \lambda\cdot\left(\widehat{M}_{\mathsf{vf}}^{\pi^*}(v) - \alpha_{\mathsf{vf}}\right)\right\}$$

$$\leq (1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0)\right] - \min_{v\in\mathcal{V}\colon \widehat{M}_{\mathsf{vf}}^{\pi^*}(v)\leq\alpha_{\mathsf{vf}}}(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[v(S_0)\right]$$

(C.6.3)

$$\leq \max_{v\in\mathcal{V}\colon \widehat{M}_{\mathsf{vf}}^{\pi^*}(v)\leq\alpha_{\mathsf{vf}}}\left|(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0) - v(S_0)\right]\right|.$$

Note that by Lemmas 4.2.7 and 4.2.8, we have

$$(1-\gamma)\mathbb{E}_{S_0\sim\nu}\left[V^{\pi^*}(S_0) - v(S_0)\right] = \Phi_{\mathsf{vf}}^{\pi^*}(v, w^{\pi^*}; \Delta^*, \Theta^*). \tag{C.6.4}$$

By plugging (C.6.4) into (C.6.3), we have

$$J(\pi^*) - J(\widehat{\pi}) \leq \max_{v\in\mathcal{V}\colon \widehat{M}_{\mathsf{vf}}^{\pi^*}(v)\leq\alpha_{\mathsf{vf}}}\left|\Phi_{\mathsf{vf}}^{\pi^*}(v, w^{\pi^*}; \Delta^*, \Theta^*)\right|$$

$$\leq c\cdot\frac{C_{\Delta^*}^2 C_{\Theta^*}^2 C_*}{1-\gamma}(\xi_0+\xi_1)L_\Pi\sqrt{\frac{1}{NT\kappa}\cdot\mathfrak{C}_{\mathcal{F}_0,\mathcal{F}_1,\mathcal{W},\mathcal{V},\Pi}\cdot\log\frac{1}{\delta}\log(NT)},$$

where in the last inequality, we use Lemma 4.4.8 with the fact that $w^{\pi^*}\in\mathcal{W}$. This concludes the proof. □

C.7. Auxiliary Results

We introduce auxiliary results used in the paper. We provide the proofs of these results in §C.8. We first introduce the following definition of β-mixing coefficient.

Definition C.7.1. Let $\{Z_t\}_{t\geq 0}$ be a sequence of random variables. For any $i, j \in \mathbb{N}\cup\{\infty\}$, we denote by σ_i^j the sigma algebra generated by $\{Z_k\}_{i\leq k\leq j}$. The β-mixing coefficient of $\{Z_t\}_{t\geq 0}$ is defined as $\beta(t) = \sup_n \mathbb{E}_{B\in\sigma_0^n}[\sup_{A\in\sigma_{n+t}^\infty} |\mathbb{P}(A \,|\, B) - \mathbb{P}(A)|]$.

We introduce the following form of β-mixing coefficient for Markov chains.

Lemma C.7.2. Suppose $\{Z_t\}_{t\geq 0}$ is a Markov chain with initial distribution ζ. It holds that

$$\beta(t) \leq \frac{1}{2}\int \|p_{t'}(\cdot \,|\, z) - p_{\mathrm{stat}}(\cdot)\|_{\mathrm{TV}} \mathrm{d}p_{\mathrm{stat}}(z) + \frac{3}{2}\int \|p_{t'}(\cdot \,|\, z) - p_{\mathrm{stat}}(\cdot)\|_{\mathrm{TV}} \mathrm{d}\zeta(z),$$

where $t' = \lfloor t/2 \rfloor$ and $p_n(\cdot \,|\, z)$ is the the marginal the distribution of Z_n given $Z_0 = z$ for any $n \in [N]$.

Proof. See the proof of Lemma 1 in Meitz and Saikkonen (2021) for a detailed proof.

□

Following from Lemma C.7.2, we can upper bound the β-mixing coefficient for a Markov chain $\{Z_t\}_{t\geq 0}$. Before that, we impose the following assumption on $\{Z_t\}_{t\geq 0}$.

Assumption C.7.3. The Markov chain $\{Z_t\}_{t\geq 0}$ with initial distribution ζ admits a unique stationary distribution p_{stat} over $\mathcal{Z}$ and is geometrically ergodic, i.e., there exists a function $\varphi\colon \mathcal{Z} \to [0, \infty)$ and a constant $\kappa > 0$ such that

$$\|p_{\mathrm{stat}}(\cdot) - p_t(\cdot \,|\, z_0)\|_{\mathrm{TV}} \leq \varphi(z_0) \cdot \exp\left(-2\kappa t\right),$$

where $p_t(\cdot \,|\, z_0)$ is the marginal distribution of Z_t given $Z_0 = z_0$ and there exists a positive absolute constant c such that $\int \varphi(z)\mathrm{d}\zeta(z) \leq c$ and $\int \varphi(z)\mathrm{d}p_{\mathrm{stat}}(z) \leq c$.

Lemma C.7.4. Suppose $\{Z_t\}_{t\geq 0}$ is a Markov chain satisfying Assumption C.7.3. Then we have $\beta(t) \leq c \cdot \exp(-\kappa t)$ for any $t \geq 0$.

Proof. For any $t \geq 0$, by Lemma C.7.2, we have

$$\begin{aligned}
\beta(t) &\leq \frac{1}{2}\int \|p_{t'}(\cdot \,|\, z) - p_{\text{stat}}(\cdot)\|_{\text{TV}} \mathrm{d}p_{\text{stat}}(z) + \frac{3}{2}\int \|p_{t'}(\cdot \,|\, z) - p_{\text{stat}}(\cdot)\|_{\text{TV}} \mathrm{d}\zeta(z) \\
&\leq \frac{1}{2}\int \varphi(z) \cdot \exp\left(-\kappa t\right) \mathrm{d}p_{\text{stat}}(z) + \frac{3}{2}\int \varphi(z) \cdot \exp\left(-\kappa t\right) \mathrm{d}\zeta(z) \\
&\leq c \cdot \exp(-\kappa t),
\end{aligned}$$

where in the second and last inequalities, we use Assumption C.7.3. This concludes the proof of the lemma. □

C.7.1. Concentration Inequality for Geometrically Ergodic Non-Stationary Sequence

We first introduce the following lemma, which is a straight-forward genelization of Berbee's lemma (Berbee, 1979).

Lemma C.7.5. For any $k > 0$ and a random sequence $\{Y_\ell\}_{\ell=1}^k$, there exists a random sequence $\{\widetilde{Y}_\ell\}_{\ell=1}^k$ such that

(1) $\{\widetilde{Y}_\ell\}_{\ell=1}^k$ are independent;

(2) for any $1 \leq \ell \leq k$, $\widetilde{Y}_\ell$ and Y_ℓ have the same distribution;

(3) for any $1 \leq \ell \leq k$, $\mathbb{P}(\widetilde{Y}_\ell \neq Y_\ell) = \beta(\sigma(\{Y_{\ell'}\}_{\ell'=1}^{\ell-1}), \sigma(\{Y_\ell\}))$.

Proof. See Lemma 2.10 in Barrera and Gobet (2021) for a detailed proof. □

We introduce the following Hoeffding's Inequality and Bernstein's Inequality for geometrically ergodic non-stationary sequences.

Theorem C.7.6 (Hoeffding's Inequality for geometrically ergodic non-stationary sequences)**.** We denote by $\{X_t\}_{t\geq 0} \subseteq \mathcal{X}$ a Markov chain satisfying Assumption C.7.3. Then for any function $f\colon \mathcal{X} \to [-f_{\max}, f_{\max}]$, it holds with probability at least $1-\delta$ with $c/(NT)^2 \leq \delta \leq 1$ that

$$\left| \frac{1}{NT} \sum_{i\in[N]} \sum_{t=0}^{T-1} f(X_t^i) - \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} f(X_t) \right] \right| \leq c \cdot f_{\max} \sqrt{\frac{1}{NT\kappa} \log \frac{2}{\delta} \log(NT)},$$

where $\{\{X_t^i\}_{t=0}^{T-1}\}_{i\in[N]}$ consists of N i.i.d. trajectories with length $T > 0$ generated from the same distribution as $\{X_t\}_{t\geq 0}$.

Proof. See §C.8.1 for a detailed proof. □

Theorem C.7.7 (Bernstein's Inequality for geometrically ergodic non-stationary sequences)**.** We denote by $\{X_t\}_{t\geq 0} \subseteq \mathcal{X}$ a Markov chain satisfying Assumption C.7.3. Then for any function $f\colon \mathcal{X} \to [-f_{\max}, f_{\max}]$, it holds with probability at least $1-\delta$ with $c/(NT)^2 \leq \delta \leq 1$ that

$$\begin{aligned} &\frac{1}{NT} \sum_{i\in[N]} \sum_{t=0}^{T-1} f(X_t^i) - \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} f(X_t) \right] \\ &\qquad \leq c_1 \cdot \frac{f_{\max}}{NT\kappa} \log \frac{2}{\delta} \log(NT) + c_2 \cdot \sqrt{\frac{1}{NT\kappa} \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} f(X_t)^2 \right] \log \frac{2}{\delta} \log(NT)}, \end{aligned}$$

where c_1 and c_2 are positive absolute constants, and $\{\{X_t^i\}_{t=0}^{T-1}\}_{i\in[N]}$ consists of N i.i.d. trajectories with length $T > 0$ generated from the same distribution as $\{X_t\}_{t\geq 0}$.

Proof. See §C.8.2 for a detailed proof. □

C.7.2. Empirical Processes for Geometrically Ergodic Non-Stationary Sequence

For any conditional probabilities $p_1(y \,|\, x)$ and $p_2(y \,|\, x)$ such that $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we define the squared Hellinger distance as follows,

$$h^2\left(p_1(\cdot \,|\, x), p_2(\cdot \,|\, x)\right) = \frac{1}{2} \int \left(\sqrt{p_1(y \,|\, x)} - \sqrt{p_2(y \,|\, x)}\right)^2 \mathrm{d}y.$$

We further assume that $\mathcal{Y}$ is a discrete space. We denote by $p^*(y \,|\, x)$ the true conditional probability of $y \in \mathcal{Y}$ given $x \in \mathcal{X}$. Also, let $\{(X_t, Y_t)\}_{t \geq 0} \subset \mathcal{X} \times \mathcal{Y}$ be a Markov chain such that $Y_t \sim p^*(\cdot \,|\, X_t)$ and satisfies Assumption C.7.3. Further, we denote by μ_t the marginal distribution of X_t for any $t \geq 0$. In the meanwhile, with $\mu = 1/T \cdot \sum_{t=0}^{T-1} \mu_t$, we define the generalized squared Hellinger distance over μ as follows,

$$H^2(p_1, p_2) = \mathbb{E}_{X \sim \mu}\left[h^2\left(p_1(\cdot \,|\, X), p_2(\cdot \,|\, X)\right)\right].$$

In the meanwhile, we are given a data set $\{\{(X_t^i, Y_t^i)\}_{t=0}^{T-1}\}_{i \in [N]}$ consisting of N independent trajectories of length T, where $\{(X_t^i, Y_t^i)\}_{t=0}^{T-1}$ is generated from the same distribution as $\{(X_t, Y_t)\}_{t \geq 0}$. We construct the following maximum likelihood estimator for p^*,

$$\widehat{p} \in \operatorname*{argmax}_{p \in \mathcal{P}} \widehat{\mathbb{E}}\left[\log p(Y \,|\, X)\right] = \frac{1}{NT} \sum_{i \in [N]} \sum_{t=0}^{T-1} \log p(Y_t^i \,|\, X_t^i). \tag{C.7.1}$$

We also define

$$g_p(x, y) = \frac{1}{2} \log \frac{p(y \,|\, x) + p^*(y \,|\, x)}{2p^*(y \,|\, x)}$$

for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Now, we are ready to introduce the following lemma.

Lemma C.7.8. We have

$$
\begin{aligned}
&H^2\left(\frac{\widehat{p}+p^*}{2}, p^*\right) \leq \left(\widehat{\mathbb{E}} - \mathbb{E}\right)\left[g_{\widehat{p}}(X, Y)\right], \\
&H^2\left(\frac{p_1+p^*}{2}, \frac{p_2+p^*}{2}\right) \leq \frac{1}{2} H^2(p_1, p_2), \\
&H^2(p, p^*) \leq 16 H^2\left(\frac{p+p^*}{2}, p^*\right), \\
&\|p_1(\cdot \,|\, x) - p_2(\cdot \,|\, x)\|_1 \leq 2\sqrt{2} h\left(p_1(\cdot \,|\, x), p_2(\cdot \,|\, x)\right).
\end{aligned}
$$

Proof. See §C.8.3 for a detailed proof. □

We define the entropy integral as follows,

$$
J_B(\delta, \overline{\mathcal{P}}^{1/2}(\delta)) = \max\left\{\int_{\delta^2/2^{10}}^{\delta} \left(H_B(u, \overline{\mathcal{P}}^{1/2}(\delta))\right)^{1/2} \mathrm{d}u, \delta\right\},
$$

where $H_B(u, \overline{\mathcal{P}}^{1/2}(\delta))$ is the entropy of the space $\overline{\mathcal{P}}^{1/2}(\delta)$ with bracketing, and $\overline{\mathcal{P}}^{1/2}(\delta)$ is defined as follows,

$$
\overline{\mathcal{P}}^{1/2}(\delta) = \{\overline{p}^{1/2} \colon p \in \mathcal{P} \text{ and } H^2(\overline{p}, p^*) \leq \delta^2\}.
$$

Now, we introduce the following theorem, which upper bounds the distance between $\widehat{p}$ and p^*.

Theorem C.7.9. We take $\Psi(\delta) \geq J_B(\delta, \overline{\mathcal{P}}^{1/2}(\delta))$ in such a way that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then for a universal constant c and any $\delta \geq \delta_{NT}$, where δ_{NT} satisfies that $\sqrt{NT}\delta_{NT}^2 \geq c\Psi(\delta_{NT})$, it holds with probability at least $1 - c/\kappa \cdot$

$\exp(-NT\kappa\delta^2/(c^2\log(NT))) - c/(N^2T^2)\cdot\log(4/\delta)$ that

$$H^2(\widehat{p}, p^*) \leq \delta^2,$$

where $\widehat{p}$ is defined in (C.7.1).

Proof. See §C.8.4 for a detailed proof. □

We study the following case, where $\mathcal{P}$ is a parametric class.

Corollary C.7.10 (Parametric Class)**.** Suppose $\mathcal{P} = \{p_\theta : \theta \in \mathbb{R}^d \text{ and } \|\theta\|_2 \leq \theta_{\max}\}$. Then with probability at least $1-\delta$ with $c/(N^2T^2)\cdot\log(NT) \leq \delta \leq 1$, we have

$$H^2(\widehat{p}, p^*) \leq c \cdot \frac{d}{NT\kappa}\log\frac{\theta_{\max}}{\delta}\log(NT),$$

where $c > 0$ is an absolute constant, which may vary from lines to lines.

Proof. Note that

$$J_B(\delta, \overline{\mathcal{P}}^{1/2}(\delta), d) \leq \delta\sqrt{d\log\frac{\theta_{\max}}{\delta}}.$$

By taking $\Psi(\delta) = \delta\sqrt{d\log(\theta_{\max}/\delta)}$, we have

$$\mathbb{P}\left(H^2(\widehat{p}, p^*) \leq c \cdot \frac{d}{NT\kappa}\log\frac{\theta_{\max}}{\delta}\log(NT)\right) \geq 1-\delta$$

with $c/(N^2T^2)\cdot\log(NT) \leq \delta \leq 1$, which concludes the proof of the corollary. □

C.8. Proofs of Auxiliary Results

C.8.1. Proof of Theorem C.7.6

Proof. We take $\tau = \min\{T, 3/\kappa \cdot \log(NT)\}$, and denote by $\mathbb{E}_{\text{stat}}[\cdot]$ the expectation taken with respect to the stationary distribution of $\{X_t\}_{t\geq 0}$. We observe the following decomposition,

$$\begin{aligned}
&\frac{1}{NT}\sum_{i\in[N]}\sum_{t=0}^{T-1} f(X_t^i) - \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)\right] \\
&\quad = \frac{\tau}{T}\underbrace{\left(\frac{1}{N\tau}\sum_{i\in[N]}\sum_{t=0}^{\tau-1} f(X_t^i) - \mathbb{E}\left[\frac{1}{\tau}\sum_{t=0}^{\tau-1} f(X_t)\right]\right)]}_{\text{(I)}} \\
&\qquad + \frac{T-\tau}{T}\underbrace{\left(\frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1} f(X_t^i) - \mathbb{E}_{\text{stat}}\left[f(X)\right]\right)}_{\text{(II)}} \\
&\qquad + \frac{T-\tau}{T}\underbrace{\left(\mathbb{E}_{\text{stat}}\left[f(X)\right] - \mathbb{E}\left[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1} f(X_t)\right]\right)}_{\text{(III)}}.
\end{aligned}$$

In the follows, we upper bound terms (I), (II), and (III), respectively.

Upper Bounding Term (I). Since $\{\{X_t^i\}_{t=0}^{T-1}\}_{i\in[N]}$ are i.i.d. accross each trajectory, by standard Hoeffding's inequality, with probability at least $1-\delta$, we have

$$|\text{(I)}| \leq f_{\max}\sqrt{\frac{2}{N}\log\frac{2}{\delta}}. \tag{C.8.1}$$

Upper Bounding Term (II). We consider an auxiliary Markov chain $\{\{\widetilde{X}_t^i\}_{t=0}^{T-1}\}_{i\in[N]}$, where the i-th trajectory $\{\widetilde{X}_t^i\}_{t=0}^{T-1}$ is sampled such that $\widetilde{X}_0^i \sim p_{\rm stat}$. Here $p_{\rm stat}$ is the stationary distribution of $\{X_t\}_{t\geq 0}$. Similarly, we define the following quantity,

$$(\widetilde{\rm II}) = \frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1} f(\widetilde{X}_t^i) - \mathbb{E}_{\rm stat}\left[f(X)\right].$$

Now, for any $x \geq 0$, we upper bound the difference $\mathbb{P}((\mathrm{II}) \geq x) - \mathbb{P}((\widetilde{\rm II}) \geq x)$ as follows,

(C.8.2)

$$\mathbb{P}\left((\mathrm{II}) \geq x\right) - \mathbb{P}\left((\widetilde{\rm II}) \geq x\right) \leq N\sum_{t=\tau}^{T-1}\mathbb{E}\left[\|p_t(\cdot \,|\, X_0) - p_{\rm stat}(\cdot)\|_{\rm TV}\right] \leq c\cdot NT\exp(-\kappa\tau).$$

Thus, by (C.8.2), to upper bound $\mathbb{P}((\mathrm{II}) \geq x)$, it suffices to upper bound $\mathbb{P}((\widetilde{\rm II}) \geq x)$.

To upper bound $\mathbb{P}((\widetilde{\rm II}) \geq x)$, we take $T - \tau = 2ks$, where k and s are two positive integers for the simplicity of presentation. We partition the set $\{\tau, \tau+1, \ldots, T-1\}$ as follows,

$$J_1 = \{\tau, \tau+1, \ldots, \tau+s-1\}, \quad J_2 = \{\tau+s, \tau+s+1, \ldots, \tau+2s-1\}, \quad \ldots,$$
$$J_{2k-1} = \{T-2s, T-2s+1, \ldots, T-s-1\}, \quad J_{2k} = \{T-s, T-s+1, \ldots, T-1\}.$$

Under such a partition, we see that $\cup_{\ell\in[2k]}J_\ell = \{\tau, \tau+1, \ldots, T-1\}$ and $J_\ell \cap J_{\ell'} = \varnothing$ for any $\ell \neq \ell'$. Also, for any $i \in [N]$, we define

$$Z_\ell^i = (\widetilde{X}_t^i)_{t\in J_\ell}$$

for any $\ell \in [2k]$. Now, for any $i \in [N]$, by Lemma C.7.5, there exists a sequence $\{W_\ell^i\}_{\ell \in [2k]}$, where $W_\ell^i = (\widetilde{Y}_t^i)_{t \in J_\ell}$ such that

(1) $\{W_\ell^i\}_{\ell \in [2k]}$ are independent;

(2) for any $\ell \in [2k]$, W_ℓ^i and Z_ℓ^i have the same distribution;

(3) for any $\ell \in [2k]$, $\mathbb{P}(W_\ell^i \neq Z_\ell^i) = \beta(\sigma(\{Z_{\ell'}^i\}_{\ell' \in [\ell-1]}), \sigma(\{Z_\ell^i\}))$.

Note that the following inclusion relation holds,

$$\left\{\frac{1}{s}\sum_{t \in J_\ell} f(\widetilde{X}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x_\ell\right\} \subseteq \left\{\frac{1}{s}\sum_{t \in J_\ell} f(\widetilde{Y}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x_\ell\right\} \cup \{W_\ell^i \neq Z_\ell^i\}$$

for any $x_\ell \in \mathbb{R}$. Thus, we have

$$\begin{aligned}
\mathbb{P}\left((\widetilde{\text{II}}) \geq x\right) &\leq \mathbb{P}\left(\frac{1}{kN}\sum_{i \in [N], \ell \text{ is odd}} \frac{1}{s}\sum_{t \in J_\ell} f(\widetilde{X}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x\right) \\
&\quad + \mathbb{P}\left(\frac{1}{kN}\sum_{i \in [N], \ell \text{ is even}} \frac{1}{s}\sum_{t \in J_\ell} f(\widetilde{X}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x\right) \\
&\leq \mathbb{P}\left(\frac{1}{kN}\sum_{i \in [N], \ell \text{ is odd}} \frac{1}{s}\sum_{t \in J_\ell} f(\widetilde{Y}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x\right) \\
&\quad + \mathbb{P}\left(\frac{1}{kN}\sum_{i \in [N], \ell \text{ is even}} \frac{1}{s}\sum_{t \in J_\ell} f(\widetilde{Y}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x\right) + \sum_{\ell \in [2k]} \mathbb{P}(W_\ell^i \neq Z_\ell^i) \\
&\leq 2\exp\left(-\frac{kNx^2}{2f_{\max}^2}\right) + 2k\beta(s),
\end{aligned}$$

where we use Hoeffding's inequality in the last line. Similarly, we have

$$\mathbb{P}\left((\widetilde{\text{II}}) \leq -x\right) \leq 2\exp\left(-\frac{kNx^2}{2f_{\max}^2}\right) + 2k\beta(s).$$

Thus, we have

$$\mathbb{P}\left(|(\mathrm{II})| \geq x\right) \leq 4\exp\left(-\frac{kNx^2}{2f_{\max}^2}\right) + 4k\beta(s) + c \cdot NT\exp(-\kappa\tau).$$

Now, by taking $s = 3\log(NT)/\kappa$, it holds with probability at least $1-\delta$ with $c/(NT)^2 \leq \delta \leq 1$ that

$$|(\mathrm{II})| \leq f_{\max}\sqrt{\frac{24}{NT\kappa}\log\frac{4}{\delta}\log(NT)}. \tag{C.8.3}$$

Upper Bounding Term (III). We observe that

$$|(\mathrm{III})| \leq f_{\max} \cdot \sum_{t=\tau}^{T-1} c \cdot \exp(-\kappa t) \leq c \cdot \frac{f_{\max}}{N^2T^2}. \tag{C.8.4}$$

Combining Everything. Now, by combining (C.8.1), (C.8.3), and (C.8.4), it holds with probability at least $1-\delta$ with $c/(NT)^2 \leq \delta \leq 1$ that

$$\frac{1}{NT}\sum_{i\in[N]}\sum_{t=0}^{T-1} f(X_t^i) - \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)\right] \leq c \cdot f_{\max}\sqrt{\frac{48}{NT\kappa}\log\frac{4}{\delta}\log(NT)},$$

which concludes the proof of the theorem. □

C.8.2. Proof of Theorem C.7.7

Proof. The proof follows from the proof of Theorem C.7.6 in §C.8.1. For the completeness of the paper, we present it here. We take $\tau = \min\{T, 3/\kappa \cdot \log(NT)\}$, and denote by $\mathbb{E}_{\mathrm{stat}}[\cdot]$ the expectation taken with respect to the stationary distribution of $\{X_t\}_{t\geq 0}$.

We observe the following decomposition,

$$\begin{aligned}
&\frac{1}{NT}\sum_{i\in[N]}\sum_{t=0}^{T-1} f(X_t^i) - \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)\right] \\
&\quad= \frac{\tau}{T}\underbrace{\left(\frac{1}{N\tau}\sum_{i\in[N]}\sum_{t=0}^{\tau-1} f(X_t^i) - \mathbb{E}\left[\frac{1}{\tau}\sum_{t=0}^{\tau-1} f(X_t)\right]\right)]}_{\text{(I)}} \\
&\qquad+ \frac{T-\tau}{T}\underbrace{\left(\frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1} f(X_t^i) - \mathbb{E}_{\text{stat}}\left[f(X)\right]\right)}_{\text{(II)}} \\
&\qquad+ \frac{T-\tau}{T}\underbrace{\left(\mathbb{E}_{\text{stat}}\left[f(X)\right] - \mathbb{E}\left[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1} f(X_t)\right]\right)}_{\text{(III)}}.
\end{aligned}$$

In the follows, we upper bound terms (I), (II), and (III), respectively.

Upper Bounding Term (I). Since $\{\{X_t^i\}_{t=0}^{T-1}\}_{i\in[N]}$ are i.i.d. across each trajectory, by standard Bernstein's inequality, with probability at least $1-\delta$, we have

$$\begin{aligned}
|\text{(I)}| &\le \frac{2f_{\max}}{3N}\log\frac{2}{\delta} + 4\sqrt{\frac{4}{N}\mathbb{E}\left[\left(\frac{1}{\tau}\sum_{t=0}^{\tau-1} f(X_t)\right)^2\right]\log\frac{2}{\delta}} \\
&= \frac{2f_{\max}}{3N}\log\frac{2}{\delta} + 4\sqrt{\frac{4}{N}\mathbb{E}\left[\frac{1}{\tau}\sum_{t=0}^{\tau-1} f(X_t)^2\right]\log\frac{2}{\delta}}. \qquad \text{(C.8.5)}
\end{aligned}$$

Upper Bounding Term (II). We consider an auxiliary dataset $\{\{\widetilde{X}_t^i\}_{t=0}^{T-1}\}_{i\in[N]}$, where the i-th trajectory $\{\widetilde{X}_t^i\}_{t=0}^{T-1}$ is sampled such that $\widetilde{X}_0^i \sim p_{\text{stat}}$. Here p_{stat} is the stationary distribution of $\{X_t\}_{t\geq 0}$. Similarly, we define the following quantity,

$$(\widetilde{\mathrm{II}}) = \frac{1}{N(T-\tau)} \sum_{i\in[N]} \sum_{t=\tau}^{T-1} f(\widetilde{X}_t^i) - \mathbb{E}_{\mathrm{stat}}\left[f(X)\right].$$

Now, for any $x \geq 0$, we upper bound the difference $\mathbb{P}((\mathrm{II}) \geq x) - \mathbb{P}((\widetilde{\mathrm{II}}) \geq x)$ as follows,

(C.8.6)

$$\mathbb{P}\left((\mathrm{II}) \geq x\right) - \mathbb{P}\left((\widetilde{\mathrm{II}}) \geq x\right) \leq N \sum_{t=\tau}^{T-1} \mathbb{E}\left[\|p_t(\cdot \,|\, X_0) - p_{\mathrm{stat}}(\cdot)\|_{\mathrm{TV}}\right] \leq c \cdot NT \exp(-\kappa\tau).$$

Thus, by (C.8.6), to upper bound $\mathbb{P}((\mathrm{II}) \geq x)$, it suffices to upper bound $\mathbb{P}((\widetilde{\mathrm{II}}) \geq x)$.

To upper bound $\mathbb{P}((\widetilde{\mathrm{II}}) \geq x)$, we take $T - \tau = 2ks$, where k and s are two positive integers for the simplicity of presentation. We partition the set $\{\tau, \tau+1, \ldots, T-1\}$ as follows,

$$J_1 = \{\tau, \tau+1, \ldots, \tau+s-1\}, \quad J_2 = \{\tau+s, \tau+s+1, \ldots, \tau+2s-1\}, \quad \ldots,$$
$$J_{2k-1} = \{T-2s, T-2s+1, \ldots, T-s-1\}, \quad J_{2k} = \{T-s, T-s+1, \ldots, T-1\}.$$

Under such a partition, we see that $\cup_{\ell\in[2k]} J_\ell = \{\tau, \tau+1, \ldots, T-1\}$ and $J_\ell \cap J_{\ell'} = \varnothing$ for any $\ell \neq \ell'$. Also, for any $i \in [N]$, we define

$$Z_\ell^i = (\widetilde{X}_t^i)_{t\in J_\ell}$$

for any $\ell \in [2k]$. Now, for any $i \in [N]$, by Lemma C.7.5, there exists a sequence $\{W_\ell^i\}_{\ell\in[2k]}$, where $W_\ell^i = (\widetilde{Y}_t^i)_{t\in J_\ell}$ such that

(1) $\{W_\ell^i\}_{\ell\in[2k]}$ are independent;

(2) for any $\ell \in [2k]$, W_ℓ^i and Z_ℓ^i have the same distribution;

(3) for any $\ell \in [2k]$, $\mathbb{P}(W_\ell^i \neq Z_\ell^i) = \beta(\sigma(\{Z_{\ell'}^i\}_{\ell' \in [\ell-1]}), \sigma(\{Z_\ell^i\}))$.

Note that the following inclusion relation holds,

$$\left\{\frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{X}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x_\ell\right\} \subseteq \left\{\frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{Y}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x_\ell\right\} \cup \{W_\ell^i \neq Z_\ell^i\}$$

for any $x_\ell \in \mathbb{R}$. Thus, we have

$$\begin{aligned}
\mathbb{P}\left((\widetilde{\text{II}}) \geq x\right) &\leq \mathbb{P}\left(\frac{1}{kN}\sum_{i\in[N],\ell\text{ is odd}} \frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{X}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x\right) \\
&\quad + \mathbb{P}\left(\frac{1}{kN}\sum_{i\in[N],\ell\text{ is even}} \frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{X}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x\right) \\
&\leq \mathbb{P}\left(\frac{1}{kN}\sum_{i\in[N],\ell\text{ is odd}} \frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{Y}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x\right) \\
&\quad + \mathbb{P}\left(\frac{1}{kN}\sum_{i\in[N],\ell\text{ is even}} \frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{Y}_t^i) - \mathbb{E}_{\text{stat}}[f(X)] \geq x\right) + \sum_{\ell\in[2k]} \mathbb{P}(W_\ell^i \neq Z_\ell^i) \\
&\leq \exp\left(-\frac{3k^2N^2x^2}{6\sum_{i\in[N],\ell\text{ is odd}} \mathbb{E}\left[\left(\frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{Y}_t^i)\right)^2\right] + 2f_{\max}Nkx}\right) \\
&\quad + \exp\left(-\frac{3k^2N^2x^2}{6\sum_{i\in[N],\ell\text{ is even}} \mathbb{E}\left[\left(\frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{Y}_t^i)\right)^2\right] + 2f_{\max}Nkx}\right) + 2k\beta(s),
\end{aligned} \tag{C.8.7}$$

where we use Bernstein's inequality in the last line. Note that for any $(i, \ell) \in [N] \times [2k]$, we have

(C.8.8)

$$\mathbb{E}\left[\left(\frac{1}{s}\sum_{t\in J_\ell} f(\widetilde{Y}_t^i)\right)^2\right] = \mathbb{E}_{\text{stat}}\left[f(X)^2\right] \leq \mathbb{E}\left[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1} f(X_t)^2\right] + f_{\max}^2 Tc \cdot \exp(-\kappa\tau).$$

Combining (C.8.7) and (C.8.8), we have

$$\mathbb{P}\left((\widetilde{\text{II}}) \geq x\right) \leq 2\exp\left(-\frac{3k^2N^2x^2}{6kN\left(\mathbb{E}\left[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1} f(X_t)^2\right] + f_{\max}^2\beta(\tau)\right) + 2f_{\max}Nkx}\right) + 2k\beta(s).$$

Similarly, we have

$$\mathbb{P}\left((\widetilde{\text{II}}) \leq -x\right) \leq 2\exp\left(-\frac{3k^2N^2x^2}{6kN\left(\mathbb{E}\left[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1} f(X_t)^2\right] + f_{\max}^2\beta(\tau)\right) + 2f_{\max}Nkx}\right) + 2k\beta(s).$$

Thus, we have

$$\begin{aligned}\mathbb{P}\left(|(\text{II})| \leq x\right) \leq 4\exp&\left(-\frac{3k^2N^2x^2}{6kN\left(\mathbb{E}\left[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1} f(X_t)^2\right] + f_{\max}^2\beta(\tau)\right) + 2f_{\max}Nkx}\right) \\ &+ 4k\beta(s) + c \cdot NT\exp(-\kappa\tau).\end{aligned}$$

Now, by taking $s = 3\log(NT)/\kappa$, it holds with probability at least $1-\delta$ with $c/(NT)^2 \leq \delta \leq 1$ that

$$|(\text{II})| \leq \frac{2f_{\max}}{3NT\kappa}\log\frac{2}{\delta} + 4\sqrt{\frac{4}{NT\kappa}\mathbb{E}\left[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1} f(X_t)^2\right]\log\frac{2}{\delta}}. \tag{C.8.9}$$

Upper Bounding Term (III). We observe that

$$|(\mathrm{III})| \leq f_{\max} \cdot \sum_{t=\tau}^{T-1} c \cdot \exp(-\kappa t) \leq c \cdot \frac{f_{\max}}{N^2 T^2}. \tag{C.8.10}$$

Combining Everything. Now, by combining (C.8.5), (C.8.9), and (C.8.10), it holds with probability at least $1-\delta$ with $c/(NT)^2 \leq \delta \leq 1$ that

$$\begin{aligned}&\frac{1}{NT}\sum_{i\in[N]}\sum_{t=0}^{T-1} f(X_t^i) - \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)\right]\\ &\qquad \leq \frac{24 f_{\max}}{NT\kappa}\log\frac{2}{\delta}\log(NT) + 48\sqrt{\frac{1}{NT\kappa}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)^2\right]\log\frac{2}{\delta}\log(NT)},\end{aligned}$$

which concludes the proof of the theorem. □

C.8.3. Proof of Lemma C.7.8

Proof. The proof follows from the proofs of Lemmas 4.1 and 4.2 in Geer et al. (2000).

First Inequality. By the optimality of $\widehat{p}$, we have

$$\widehat{\mathbb{E}}\left[\log \widehat{p}(Y \mid X)\right] \geq \widehat{\mathbb{E}}\left[\log p^*(Y \mid X)\right],$$

which implies that

$$\int \log\frac{\widehat{p}}{p^*}\mathrm{d}\widetilde{p}^*\mathrm{d}\widetilde{\mu} \geq 0,$$

where $\widetilde{p}^*$ and $\widetilde{\mu}$ are empirical counterparts of p^* and μ. Now, by the concavity of $\log(\cdot)$, we have

$$\log \frac{\widehat{p} + p^*}{2p^*} \geq \frac{1}{2} \log \frac{\widehat{p}}{p^*} + \frac{1}{2} \log \frac{p^*}{p^*} = \frac{1}{2} \log \frac{\widehat{p}}{p^*}.$$

By the above two inequalities, we have

$$\begin{aligned} 0 &\leq \frac{1}{4} \int \log \frac{\widehat{p}}{p^*} \mathrm{d}\widetilde{p}^* \mathrm{d}\widetilde{\mu} \leq \frac{1}{2} \int \log \frac{\widehat{p} + p^*}{2p^*} \mathrm{d}\widetilde{p}^* \mathrm{d}\widetilde{\mu} \\ &= \frac{1}{2} \int \log \frac{\widehat{p} + p^*}{2p^*} \left(\mathrm{d}\widetilde{p}^* \mathrm{d}\widetilde{\mu} - \mathrm{d}p^* \mathrm{d}\mu\right) + \frac{1}{2} \int \log \frac{\widehat{p} + p^*}{2p^*} \mathrm{d}p^* \mathrm{d}\mu \\ &= \left(\widehat{\mathbb{E}} - \mathbb{E}\right) [g_{\widehat{p}}] + \frac{1}{2} \int \log \frac{\widehat{p} + p^*}{2p^*} \mathrm{d}p^* \mathrm{d}\mu. \end{aligned} \tag{C.8.11}$$

In the meanwhile, by the fact that $\log z \leq 2(\sqrt{z} - 1)$ for any $z > 0$, we have

$$\begin{aligned} \frac{1}{2} \int \log \frac{\widehat{p} + p^*}{2p^*} \mathrm{d}p^* \mathrm{d}\mu &\leq \int \left(\sqrt{\frac{\widehat{p} + p^*}{2p^*}} - 1 \right) \mathrm{d}p^* \mathrm{d}\mu \\ &= \int \left(\sqrt{\frac{\widehat{p} + p^*}{2} \cdot p^*} - \frac{1}{2} p^* - \frac{1}{2} \cdot \frac{\widehat{p} + p^*}{2} \right) \mathrm{d}y \mathrm{d}\mu \\ &= - \int \frac{1}{2} \left(\sqrt{\frac{\widehat{p} + p^*}{2}} - \sqrt{p^*} \right)^2 \mathrm{d}y \mathrm{d}\mu \\ &= -H^2 \left(\frac{\widehat{p} + p^*}{2}, p^* \right). \end{aligned} \tag{C.8.12}$$

By combining (C.8.11) and (C.8.12), we conclude the proof of the first inequality.

Second&Third Inequality. We denote by $\overline{p} = (p + p^*)/2$ for any p. We note the following two facts,

$$\begin{aligned}
&\frac{p_1^{1/2} + p_2^{1/2}}{\overline{p}_1^{1/2} + \overline{p}_2^{1/2}} \leq \sqrt{2}, \\
&\left|\overline{p}_1^{1/2} - \overline{p}_2^{1/2}\right| \left(\overline{p}_1^{1/2} + \overline{p}_2^{1/2}\right) = |\overline{p}_1 - \overline{p}_2| = \left|\frac{p_1 - p_2}{2}\right| = \frac{1}{2}\left|p_1^{1/2} - p_2^{1/2}\right| \left(p_1^{1/2} + p_2^{1/2}\right).
\end{aligned}$$

Thus, we have

$$\left|\overline{p}_1^{1/2} - \overline{p}_2^{1/2}\right| = \frac{1}{2} \cdot \frac{p_1^{1/2} + p_2^{1/2}}{\overline{p}_1^{1/2} + \overline{p}_2^{1/2}} \cdot \left|p_1^{1/2} - p_2^{1/2}\right| \leq \frac{\sqrt{2}}{2} \cdot \left|p_1^{1/2} - p_2^{1/2}\right|,$$

which implies the second inequality. The third inequality can be proved in a similar way.

Forth Inequality. We note that

$$\begin{aligned}
\|p_1(\cdot \,|\, x) - p_2(\cdot \,|\, x)\|_1 &= \int |p_1(y \,|\, x) - p_2(y \,|\, x)| \, \mathrm{d}y \\
&= \int \left(p_1(y \,|\, x)^{1/2} - p_2(y \,|\, x)^{1/2}\right) \left(p_1(y \,|\, x)^{1/2} + p_2(y \,|\, x)^{1/2}\right) \mathrm{d}y \\
&= \sqrt{\int (p_1(y \,|\, x)^{1/2} - p_2(y \,|\, x)^{1/2}) \, \mathrm{d}y} \sqrt{\int (p_1(y \,|\, x)^{1/2} + p_2(y \,|\, x)^{1/2}) \, \mathrm{d}y} \\
&\leq 2\sqrt{\int (p_1(y \,|\, x)^{1/2} - p_2(y \,|\, x)^{1/2}) \, \mathrm{d}y} = 2\sqrt{2}h \left(p_1(\cdot \,|\, x), p_2(\cdot \,|\, x)\right),
\end{aligned}$$

which concludes the proof. □

C.8.4. Proof of Theorem C.7.9

Proof. The proof follows from the proof of Theorem 7.4 in Geer et al. (2000). We define the events

$$\mathcal{E} = \left\{\omega \in \Omega \colon H^2(\widehat{p}, p^*) > \delta^2\right\}.$$

Conditioning on $\mathcal{E}$, we have

$$\left(\widehat{\mathbb{E}} - \mathbb{E}\right)[g_{\widehat{p}}] \geq H^2(\overline{\widehat{p}}, p^*) \geq \frac{1}{16} H^2(\widehat{p}, p^*) > \frac{\delta^2}{16}, \tag{C.8.13}$$

where the first two inequalities come from Lemma C.7.8. We further define

$$\mathcal{E}^\dagger = \left\{\omega \in \Omega \colon \sup_{p \in \mathcal{P} \colon H^2(\overline{p}, p^*) > \delta^2/16} \left(\widehat{\mathbb{E}} - \mathbb{E}\right)[g_p] - H^2(\overline{p}, p^*) \geq 0\right\}.$$

By (C.8.13) and the definitions of $\mathcal{E}$ and $\mathcal{E}^\dagger$, we observe that $\mathcal{E} \subseteq \mathcal{E}^\dagger$. Thus, we only need to upper bound $\mathbb{P}(\mathcal{E}^\dagger)$. To do so, we use a peeling argument as follows. Let $L = \min\{\ell \colon 2^{\ell+1}\delta^2/16 > 1\}$. We observe that

$$\mathbb{P}(\mathcal{E}^\dagger) \leq \sum_{\ell=0}^{L} \mathbb{P}(\mathcal{E}_\ell^\dagger), \tag{C.8.14}$$

where

$$\mathcal{E}_\ell^\dagger = \left\{\omega \in \Omega \colon \sup_{p \in \mathcal{P}_\ell} \left(\widehat{\mathbb{E}}[g_p] - \mathbb{E}[g_p]\right) \geq 2^\ell \delta^2/16\right\}.$$

Here $\mathcal{P}_\ell = \{p \in \mathcal{P} \colon H^2(\overline{p}, p^*) \leq 2^{\ell+1}\delta^2/16\}$. To upper bound $\mathbb{P}(\mathcal{E}_\ell^\dagger)$, we introduce the following result.

Theorem C.8.1. Given a Markov chain $\{Z_t\}_{t\geq 0} \subset \mathcal{Z}$ satisfying Assumption C.7.3, and take

$$v \leq C_1\sqrt{NT}R^2/K, \tag{C.8.15}$$

$$v \leq 8\sqrt{NT}R, \tag{C.8.16}$$

$$v \geq C_0 \cdot \max\left\{\int_{v/(2^6\sqrt{NT})}^{R} \left(\mathcal{H}_{B,K}(u, \mathcal{G}, P)\right)^{1/2} \mathrm{d}u, R\right\}, \tag{C.8.17}$$

$$v \geq C_2/(NT)^2, \tag{C.8.18}$$

$$C_0^2 \geq C^2(C_1 + 1), \tag{C.8.19}$$

where $\mathcal{H}_{B,K}(u, \mathcal{G}, P)$ is the generalized entropy with bracketing. Then we have

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \sqrt{NT}\left(\frac{1}{NT}\sum_{i\in[N]}\sum_{t=0}^{T-1} g(Z_t^i) - \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(Z_t)\right]\right) \geq v\right)$$
$$\leq \frac{4C}{\kappa}\exp\left(-\frac{v^2\kappa}{18C^2(C_1+1)R^2\log(NT)}\right) + \frac{2}{N^2T^2},$$

where $\{Z_t^i\}_{t=0}^{T-1}\}$ is generated from the same distribution as $\{Z_t\}_{t\geq 0}$ for any $i \in [N]$.

Proof. See §C.8.5 for a detailed proof. □

To invoke Theorem C.8.1, we take

$$v = \sqrt{NT}\cdot 2^\ell\delta^2/16, \quad K = 1, \quad R = 2^{\ell/2}\delta, \quad C_1 = 15, \quad C = c/64, \quad C_0 = c/16, \quad C_2 = c.$$

It is easy to verify that (C.8.15), (C.8.16), (C.8.18), and (C.8.19) hold. For (C.8.17), since $\sqrt{NT}\delta_{NT}^2 \geq c\Psi(\delta_{NT})$, which implies that

$$\sqrt{NT} \geq c \cdot \frac{\Psi(\delta_{NT})}{\delta_{NT}^2} \geq c \cdot \frac{\Psi(2^{\ell/2}\delta)}{2^\ell \delta^2},$$

where we use the fact that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Thus, we have

$$16a \geq c \cdot \max\left\{ \int_{v/(2^6\sqrt{NT})}^{R} \left(\mathcal{H}_{B,1}\left(u, \{g_p \colon p \in \mathcal{P}_\ell\}, \mu_0\right)\right)^{1/2} \mathrm{d}u, R \right\},$$

which justifies (C.8.17) by noting that $K = 1$. Here, we use the fact that

$$\mathcal{H}_{B,1}(u, \{g_p \colon p \in \mathcal{P}_\ell\}, P) \leq H_B\left(\frac{u}{\sqrt{2}}, \{\overline{p}^{1/2} \colon p \in \mathcal{P}_\ell\}\right).$$

Thus, by using Theorem C.8.1, we have

$$\mathbb{P}(\mathcal{E}_\ell^\dagger) \leq \frac{c}{\kappa} \exp\left(-\frac{NT\kappa 2^\ell \delta^2}{c^2 \log(NT)}\right) + \frac{2}{N^2T^2}.$$

Further, by combining (C.8.14), we have

$$\mathbb{P}(\mathcal{E}^\dagger) \leq \frac{c}{\kappa} \exp\left(-\frac{NT\kappa\delta^2}{c^2}\right) + \frac{c}{N^2T^2} \log\frac{4}{\delta},$$

which concludes the proof of the theorem. □

C.8.5. Proof of Theorem C.8.1

Proof. We take $\tau = \min\{T, 3/\kappa \cdot \log(\mathcal{G}_{\max}NT)\}$, where $\mathcal{G}_{\max} = \max\{\max_{g\in\mathcal{G}} \max_{z\in\mathcal{Z}} g(z), 1\}$, and denote by $\mathbb{E}_{\text{stat}}[\cdot]$ the expectation taken with respect to the stationary distribution of

$\{Z_t\}_{t\geq 0}$. We have the following decomposition,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g\in\mathcal{G}}\sqrt{NT}\left(\frac{1}{NT}\sum_{i\in[N]}\sum_{t=0}^{T-1}g(Z_t^i)-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}g(Z_t)\right]\right)\geq v\right)\\
&\quad=\mathbb{P}\Bigg(\sup_{g\in\mathcal{G}}\frac{\tau}{T}\bigg(\frac{1}{N\tau}\sum_{i\in[N]}\sum_{t=0}^{\tau-1}g(Z_t^i)-\mathbb{E}\bigg[\frac{1}{\tau}\sum_{t=0}^{\tau-1}g(Z_t)\bigg]\bigg)\\
&\qquad\qquad+\frac{T-\tau}{T}\bigg(\frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1}g(Z_t^i)-\mathbb{E}_{\text{stat}}[g(Z)]\bigg)\\
&\qquad\qquad+\frac{T-\tau}{T}\bigg(\mathbb{E}_{\text{stat}}[g(Z)]-\mathbb{E}\bigg[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1}g(Z_t)\bigg]\bigg)\geq\frac{v}{\sqrt{NT}}\Bigg)\\
&\quad\leq\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{\tau}{T}\left(\frac{1}{N\tau}\sum_{i\in[N]}\sum_{t=0}^{\tau-1}g(Z_t^i)-\mathbb{E}\left[\frac{1}{\tau}\sum_{t=0}^{\tau-1}g(Z_t)\right]\right)\geq\frac{v}{3\sqrt{NT}}\right)
\end{aligned}$$

(C.8.20)

$$\qquad+\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{T-\tau}{T}\left(\frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1}g(Z_t^i)-\mathbb{E}_{\text{stat}}\left[g(Z)\right]\right)\geq\frac{v}{3\sqrt{NT}}\right),$$

where the last inequality comes from the fact that

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}\frac{T-\tau}{T}\left(\mathbb{E}_{\text{stat}}[g(Z)]-\mathbb{E}\left[\frac{1}{T-\tau}\sum_{t=\tau}^{T-1}g(Z_t)\right]\right)\\
&\quad\leq\sup_{g\in\mathcal{G}}\frac{T-\tau}{T}\int g(z)\frac{1}{T-\tau}\sum_{t=\tau}^{T-1}\left(p_{\text{stat}}(z)-\int p_t(z\,|\,z_0)\mathrm{d}\zeta(z_0)\right)\mathrm{d}z\\
&\quad\leq\frac{1}{T}\mathcal{G}_{\max}\sum_{t=\tau}^{T-1}c\cdot\exp(-\kappa t)\leq\mathcal{G}_{\max}c\exp(-\kappa\tau)\leq\frac{v}{3\sqrt{NT}},
\end{aligned}$$

where we use the fact that $v \geq C_2/(NT)^2$ for some constant C_2. Thus, we only need to upper bound the two terms on the RHS of (C.8.20). We first introduce the following supporting results.

Lemma C.8.2. Take

$$
\begin{aligned}
&v \leq C_1\sqrt{n}R^2/K,\\
&v \leq 8\sqrt{n}R,\\
&v \geq C_0 \cdot \max\left\{\int_{v/(2^6\sqrt{n})}^{R} \left(\log \mathcal{N}_{B,K}(u, \mathcal{G}, P)\right)^{1/2} \mathrm{d}u, R\right\},\\
&C_0^2 \geq C^2(C_1+1).
\end{aligned}
$$

Then we have

$$
\mathbb{P}\left(\sup_{g\in\mathcal{G}}\left|\sqrt{n}\left(\frac{1}{n}\sum_{i\in[n]} g(Z_i) - \mathbb{E}\left[g(Z)\right]\right)\right| \geq v\right) \leq C\exp\left(-\frac{v^2}{C^2(C_1+1)R^2}\right),
$$

where $\{Z_i\}_{i\in[n]}$ are i.i.d. samples drawn the same distribution as Z.

Proof. See Theorem 5.11 in Geer et al. (2000) for a detailed proof. □

Lemma C.8.3. Given a β-mixing sequence $\{Z_t\}_{t\geq 0} \subset \mathcal{Z}$ with coefficient $\beta(t)$ for any $t \geq 0$. There exists a sequence $\{Z_t^*\}_{t=0}^{T-1} \subset \mathcal{Z}$ and a set $\mathcal{J}$ such that

(1) $\mathcal{J}$ is a partition of $\{0, 1, \ldots, T-1\}$, i.e., $\cup_{J\in\mathcal{J}} J = \{0, 1, \ldots, T-1\}$ and $J_1 \cap J_2 = \varnothing$ for any $J_1, J_2 \in \mathcal{J}$;

(2) for any $0 \leq t \leq T-1$, Z_t^* and Z_t have the same distribution;

(3) for any $J \in \mathcal{J}$, $\{Z_t^*\}_{t\in J}$ is an independent sequence;

(4) it holds for any $u \in \mathbb{R}$ that

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{1}{T} \sum_{t=0}^{T-1} g(Z_t) - \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} g(Z_t)\right] \geq u\right)$$
$$\leq \sum_{J \in \mathcal{J}} \mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{1}{|J|} \sum_{t \in J} g(Z_t^*) - \mathbb{E}\left[\frac{1}{|J|} \sum_{t \in J} g(Z_t)\right] \geq u\right)$$
$$+ \sum_{J \in \mathcal{J}} |J| \cdot \beta\left(\min\{|t_1 - t_2| \colon t_1 \neq t_2 \in J\}\right).$$

Proof. See Theorem 2.11 in Barrera and Gobet (2021) for a detailed proof. □

We upper bound two terms on the RHS of (C.8.20) as follows.

Upper Bounding the First Term on the RHS of (C.8.20)**.** To upper bound the first term, we invoke Lemma C.8.2. Since the sequence

$$\left\{\frac{1}{\tau} \sum_{t=0}^{\tau-1} g(Z_t^i)\right\}_{i \in [n]}$$

is i.i.d., we have

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{\tau}{T}\left(\frac{1}{N\tau} \sum_{i \in [N]} \sum_{t=0}^{\tau-1} g(Z_t^i) - \mathbb{E}\left[\frac{1}{\tau} \sum_{t=0}^{\tau-1} g(Z_t)\right]\right) \geq \frac{v}{3\sqrt{NT}}\right)$$
$$\leq C \exp\left(-\frac{v^2 T}{9\tau^2 C^2 (C_1 + 1) R^2}\right) \leq C \exp\left(-\frac{v^2}{C^2 (C_1 + 1) R^2}\right). \tag{C.8.21}$$

Upper Bounding the Second Term on the RHS of (C.8.20)**.** To upper bound the second term, we note that

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{T-\tau}{T}\left(\frac{1}{N(T-\tau)} \sum_{i \in [N]} \sum_{t=\tau}^{T-1} g(Z_t^i) - \mathbb{E}_{\text{stat}}\left[g(Z)\right]\right) \geq \frac{v}{3\sqrt{NT}}\right) = \text{(i)} + \text{(ii)},$$

where

$$\text{(i)} = \mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{T-\tau}{T}\left(\frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1} g(Z_t^i) - \mathbb{E}_{\text{stat}}\left[g(Z)\right]\right) \geq \frac{v}{3\sqrt{NT}}\right)$$
$$\qquad - \mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{T-\tau}{T}\left(\frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1} g(\widetilde{Z}_t^i) - \mathbb{E}_{\text{stat}}\left[g(Z)\right]\right) \geq \frac{v}{3\sqrt{NT}}\right),$$
$$\text{(ii)} = \mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{T-\tau}{T}\left(\frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1} g(\widetilde{Z}_t^i) - \mathbb{E}_{\text{stat}}\left[g(Z)\right]\right) \geq \frac{v}{3\sqrt{NT}}\right).$$

Here $\{\widetilde{Z}_t^i\}_{t=0}^{T-1}$ are an auxiliary sequence for any $i \in [N]$, where $\widetilde{Z}_0^i$ is sampled from the stationary distribution of the sequence $\{Z_t\}_{t\geq 0}$. To upper bound (i), we note that

$$\text{(i)} \leq \sum_{i\in[N]}\sum_{t=\tau}^{T-1} c\cdot\exp(-\kappa t) \leq NTc\exp(-\kappa\tau) \leq \frac{1}{N^2T^2}. \tag{C.8.22}$$

To upper bound (ii), we invoke Lemma C.8.3 by taking

$$\mathcal{J} = \{J_1, J_2, \ldots, J_s\}, \qquad J_j = \{\tau+j-1, \tau+j+s-1, \ldots, T-s+j\} \text{ for any } j\in[s].$$

Then there exists a sequence $\{\{\widetilde{Z}_t^{i*}\}_{t=\tau}^{T-1}\}_{i\in[N]}$ such that $\widetilde{Z}_t^{i*}$ and $\widetilde{Z}_t^i$ have the same distribution for any (i,t); $\{\{\widetilde{Z}_t^*\}_{t\in J}\}_{i\in[N]}$ are independent; and it holds for any $u\in\mathbb{R}$ that

$$\text{(ii)} \leq \sum_{j=1}^{s}\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{s}{N(T-\tau)}\sum_{i\in[N]}\sum_{t\in J_j} g(\widetilde{Z}_t^{i*}) - \mathbb{E}_{\text{stat}}\left[g(Z)\right] \geq \frac{v}{3\sqrt{NT}}\frac{T}{T-\tau}\right) + (T-\tau)\cdot\beta(s)$$
$$\leq s\cdot C\exp\left(-\frac{v^2}{9sC^2(C_1+1)R^2}\right) + (T-\tau)\beta(s),$$

where we use Lemma C.8.2 in the last inequality. Now, by taking $s = \min\{T, 3/\kappa \cdot \log(NT)\}$, we have

$$\text{(ii)} \le \frac{3C}{\kappa} \exp\left(-\frac{v^2\kappa}{18C^2(C_1+1)R^2\log(NT)}\right) + \frac{1}{N^2T^2}, \tag{C.8.23}$$

where we use Lemma C.7.4 to upper bound $\beta(s)$. Now, by combining (C.8.22) and (C.8.23), we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{T-\tau}{T}\left(\frac{1}{N(T-\tau)}\sum_{i\in[N]}\sum_{t=\tau}^{T-1} g(Z_t^i) - \mathbb{E}_{\text{stat}}\left[g(Z)\right]\right) \ge \frac{v}{3\sqrt{NT}}\right) \\
&\quad \le \frac{3C}{\kappa}\exp\left(-\frac{v^2\kappa}{18C^2(C_1+1)R^2\log(NT)}\right) + \frac{2}{N^2T^2}.
\end{aligned} \tag{C.8.24}$$

Combining Everything. By plugging (C.8.21) and (C.8.24) into (C.8.20), we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g\in\mathcal{G}} \sqrt{NT}\left(\frac{1}{NT}\sum_{i\in[N]}\sum_{t=0}^{T-1} g(Z_t^i) - \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(Z_t)\right]\right) \ge v\right) \\
&\quad \le \frac{4C}{\kappa}\exp\left(-\frac{v^2\kappa}{18C^2(C_1+1)R^2\log(NT)}\right) + \frac{2}{N^2T^2},
\end{aligned}$$

which concludes the proof of the theorem. □