Efficient and Robust Policy Evaluation for Reinforcement Learning

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Shuze Daniel Liu

May 2025

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Shuze Daniel Liu

This Dissertation has been read and approved by the examing committee:

Advisor: Shangtong Zhang

Advisor:

Committee Member: Aidong Zhang

Committee Member: Chen-Yu Wei

Committee Member: Cong Shen

Committee Member: Feng Lu

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

May 2025

To my parents Zhigang and Wenli
and my family Yuxin

Acknowledgements

I would like to express my deepest gratitude to my advisor Shangtong Zhang, whose insightful guidance and continuous encouragement have been instrumental throughout my doctoral study. Shangtong's mentorship has profoundly shaped my thinking, and his thoughtful advice and support beyond academia have greatly enriched my growth as a researcher and individual. I would also like to thank my committee members for their feedback and appraisal of my work: Aidong Zhang, Cong Shen, Lu Feng, and Chen-Yu Wei. Their advice and encouragement have been instrumental in shaping both the technical depth and clarity of this thesis. I am deeply grateful to my coauthors, Haifeng Xu, Shuhang Chen, and Weiran Shen for their invaluable collaboration. Their collective efforts, insightful discussions, and dedicated feedback have significantly strengthened our research and made this journey a rewarding experience. I also thank my colleagues in Sequential Intelligence Lab for fostering an environment of intellectual curiosity and collaboration. Their openness in sharing ideas and resources made my research experience both productive and enjoyable. My heartfelt thanks go to my parents, Zhigang and Wenli, whose unwavering support and encouragement have been invaluable to me throughout the way. I am especially grateful to Yuxin for her deepest love, support, and understanding, which have been a constant source of strength throughout my journey. I look forward to sharing a joyful and fulfilling life together, filled with growth and companionship.

Abstract

Evaluating the quality of a policy (i.e, a decision making rule that an agent adopts to interact with the environment) is central to reinforcement learning (RL). The conventional approach requires repeatedly executing the policy and averaging its outcome. However, due to high evaluation variance, this method demands massive active interactions with the environment to obtain data, which is both expensive and slow. At the same time, the stability of many RL algorithms remains an open question, as existing analyses often rely on assumptions that fail to hold in practice. This thesis addresses both challenges by proposing algorithms that enhance the efficiency and robustness of policy evaluation, and by providing a stability analysis of RL algorithms under realistic conditions.

The first part of the thesis focuses on algorithmic innovations reducing the amount of data needed for accurate evaluation. We begin by introducing an optimal data collecting policy that significantly lowers evaluation variance compared to traditional approaches. In settings where multiple policies must be evaluated at once, we propose a shared data-collecting policy that reduces evaluation variance across all interested policies simultaneously. Taking a step further, we develop a doubly-optimal policy evaluation method that optimizes both data collection and processing stages, achieving state-of-the-art variance reduction. To enhance robustness and safety, we also design methods that account for uncertainty in the environment and enforce safety constraints during data collection. Together, these contributions offer a framework for reliable and sample-efficient policy evaluation.

In the second part of the thesis, we focus on the theoretical analysis of RL algorithms. We study the stability of stochastic approximation methods (i.e., iterative methods that update estimates using noisy observations of a target quantity) under Markovian noise (i.e., when the randomness in updates comes from data generated by a Markov process, so that samples

are correlated rather than independent). Our framework provides almost sure convergence guarantees under practical conditions and applies to a broader class of RL algorithms than previous results.

Collectively, this thesis advances both the algorithmic and theoretical foundations of policy evaluation, offering tools that are efficient, reliable, and applicable to real-world RL systems.

Preface

This proposal is based on several papers I authored. In particular, Chapter 4 is based on Liu and Zhang (2024); Chapter 5 is based on Liu et al. (2025c); Chapter 6 is based on Liu et al. (2025a); Chapter 7 is based on my equally-contributed first-authored paper Chen et al. (2025), and Chapter 9 is based on Liu et al. (2025b). Besides, Chapter 8 focuses on my proposed method.

1. Liu, S. and Zhang, S. (2024). Efficient policy evaluation with offline data informed behavior policy design. In *Proceedings of the International Conference on Machine Learning*.
2. Liu, S., Chen, C., and Zhang, S. (2025a). Doubly optimal policy evaluation for reinforcement learning. In *Proceedings of the International Conference on Learning Representations*.
3. Liu, S., Chen, Y., and Zhang, S. (2025c). Efficient multi-policy evaluation for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Oral Presentation Honor).
4. Chen, C., Liu, S., and Zhang, S. (2025). Efficient policy evaluation with safety constraint for reinforcement learning. In *Proceedings of the International Conference on Learning Representations*.
5. Liu, S., Chen, S., and Zhang, S. (2025b). The ode method for stochastic approximation and reinforcement learning with markovian noise. *Journal of Machine Learning Research*.

Apart from the above papers on reinforcement learning, I also have the following paper during my PhD study.

1. Liu, S., Shen, W., and Xu, H. (2021a). Optimal pricing of information. In *Proceedings of the ACM Conference on Economics and Computation*.

Contents

List of Figures

Chapter 1

Introduction

Reinforcement learning (RL, Sutton and Barto (2018)) has achieved remarkable success in various sequential decision-making problems. For example, RL algorithms have reduced energy consumption for Google data centers' cooling by 40% (Chervonyi et al., 2022), predicted protein structures with competitive accuracy (Jumper et al., 2021), and discovered faster matrix multiplication algorithms (Fawzi et al., 2022). When applying RL algorithms, *policy evaluation* plays a critical role to allow practitioners to estimate the performance of a policy before committing to its full deployment and test different algorithmic choices. A commonly used approach among RL practitioners for policy evaluation is the on-policy Monte Carlo method, where a policy (i.e., the target policy) is evaluated by directly executing itself. However, using the target policy itself as the behavior policy is not optimal (Liu and Zhang, 2024), leading to evaluation with high variance. This suboptimality of on-policy evaluation results in extensive needs for collecting online samples to achieve a desired level of accuracy.

In many scenarios, heavily relying on online data is not preferable, since collecting massive online data through real-world interaction is both expensive and slow (Li, 2019; Zhang, 2023). Even with a well-developed simulator, complex tasks like data center cooling take 10 seconds per step (Chervonyi et al., 2022), making the evaluation of a policy requiring millions of steps prohibitively expensive. To address the expensive nature of online data, *offline RL* is proposed to mitigate the dependency on online data. However, there are often mismatches between the offline data distribution and the data distribution induced by the target policy, leading to uncontrolled and ineliminable bias (Jiang and Li, 2016; Farahmand and Szepesvári, 2011; Marivate, 2015). As a result, a policy with high performance on offline data may actually perform very poorly in real deployment (Levine, 2018). Consequently, both online and offline RL practitioners still heavily rely on online policy evaluation methods (Kalashnikov et al., 2018; Vinyals et al., 2019).

To improve the online sample efficiency for policy evaluation, existing methods utilize the off-policy evaluation (OPE) strategy, designing different data-collecting behavior policies to reduce the evaluation variance (Hanna et al., 2017; Zhong et al., 2022; Liu and Zhang, 2024; Liu et al., 2025a). Under this regime, our prior work (Liu and Zhang, 2024) designs a closed-formed optimal behavior policies that minimize estimation variance by leveraging existing offline data. This approach significantly reduces the variance compared to traditional on-policy methods, as confirmed by empirical studies across multiple benchmarks.

In many reinforcement learning applications such as hyperparameter tuning and model selection, evaluating multiple target policies simultaneously is a common need. Traditional methods typically require separate data collections for each policy evaluation, which leads to high inefficiency. To tackle this problem, we develop a shared variance-reducing behavior policy for multiple target policies (Liu et al., 2025c). Our method gives unbiased estimation and reduce evaluation variance for multiple target policies at the same time. It significantly reduces the total number of online samples needed and empirically outperforms existing multi-policy evaluation techniques.

Building on these insights, we further design a doubly optimal policy evaluation framework that simultaneously optimizes both data collection and data processing phases (Liu et al., 2025a). This doubly optimal estimator is obtained from a bi-level optimization problem, incorporating the optimal data-collecting policy with the optimal data-processing baselines. Our method achieves state-of-the-art performance in reducing variance. Theoretically, we prove the estimator's unbiasedness and variance reduction properties compared to previous best-performing methods. Empirically, Liu et al. (2025a) saves substantial amount of online data across various environments, demonstrating state-of-the-art performance.

In many real-world applications, ensuring safety during policy execution is as crucial as achieving sample efficiency. Considering the safety concern in running the behavior policy, we introduce a safety-constrained off-policy evaluation method (Chen et al., 2025). This method incorporates safety constraints directly into the behavior policy design, ensuring that policy evaluations maintain rigorous safety standards without compromising efficiency. Our theoretical results guarantee variance reduction and safety constraint satisfaction simultaneously, with empirical experiments demonstrating superior performance compared to existing methods.

While these methods effectively reduce variance, a critical limitation persists: existing OPE approaches typically assume fixed transition dynamics between training

and deployment environments. However, transition dynamics often change in real-world scenarios due to factors such as system drift or unforeseen perturbations, leading to unreliable evaluations (Wang et al., 2023). To tackle this problem, we further propose an efficient and robust off-policy evaluation method, which explicitly models uncertainty in transition dynamics through transition gradient optimization. Our approach formulates policy evaluation as a minimax optimization problem to identify and handle worst-case transition scenarios systematically. This ensures robust evaluations even under significant environmental shifts. Analytical derivations and empirical tests highlight the effectiveness and robustness of this method in handling dynamic transition uncertainties.

In addition to improving empirical performance and robustness in policy evaluation, we also contribute theoretical foundations that enhance the understanding of stability in reinforcement learning algorithms. Stochastic approximation forms the backbone of many RL algorithms, which iteratively update estimates based on noisy samples from the environment. This framework underlies widely used methods such as temporal difference learning (Sutton, 1988). Analyzing the convergence of these algorithms typically requires strong assumptions on the noise process, such as i.i.d. or martingale difference noise. However, in RL, the noise is often Markovian and potentially unbounded, particularly when eligibility traces and function approximation are involved. To address this, we develop a new ordinary differential equation (ODE) framework that extends the classical Borkar-Meyn theorem to the Markovian noise setting under verifiable and broadly applicable assumptions (Liu et al., 2025b). This framework guarantees almost sure stability and convergence of stochastic approximation iterates and provides a general-purpose theoretical tool for analyzing the dynamics of RL algorithms under realistic conditions.

This thesis explores a series of advanced strategies for efficient, robust, and safe policy evaluation, addressing key challenges such as variance reduction, multi-policy evaluation, safety constraints, and transition uncertainty. Beyond empirical advancements, we also provide a general theoretical framework that strengthens the stability guarantees for a wide class of reinforcement learning algorithms modeled as stochastic approximation. Together, these contributions offer both practical tools for high-stakes RL deployments and foundational insights for the design and analysis of RL algorithms in complex, dynamic environments.

Chapter 2

Background

2.1 Finite Markov Decision Process

A finite-horizon Markov Decision Process (MDP) is defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r, p_0, T)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, respectively. The transition dynamics are characterized by a probability distribution $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, where $p(s' \mid s, a)$ denotes the probability of transitioning to state s' upon taking action a in state s. The reward function is defined as $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, assigning a scalar reward to each state-action pair. The initial state distribution is denoted by $p_0 : \mathcal{S} \to [0, 1]$, and the horizon length T specifies the finite length of an episode.

In reinforcement learning, the agent interacts with the MDP sequentially over time steps $t \in \{0, 1, \ldots, T-1\}$. At each time step t, the agent observes a state $S_t \in \mathcal{S}$, selects an action $A_t \in \mathcal{A}$ according to a policy $\pi_t(a \mid s)$, receives a reward $R_{t+1} = r(S_t, A_t)$, and transitions to the next state S_{t+1} drawn from $p(\cdot \mid S_t, A_t)$. We denote the generated trajectory in a finite MDP as $\tau = (S_0, A_0, R_1, \ldots, S_T)$. We also use abbreviations $\pi_{i:j} \doteq \{\pi_i, \pi_{i+1}, \ldots, \pi_j\}$ and $\pi \doteq \pi_{0:T-1}$.

The return at time step t is defined as the cumulative sum of future rewards:

$$G_t \doteq \sum_{i=t+1}^{T} R_i.$$

Correspondingly, the state-value function and action-value function of policy π at time step t are defined as

$$v_{\pi,t}(s) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s \right], \quad q_{\pi,t}(s, a) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s, A_t = a \right].$$

The overall performance of a policy π is measured by the expected total return over the initial state distribution:

$$J(\pi) \doteq \sum_{s \in \mathcal{S}} p_0(s) v_{\pi,0}(s).$$

In this thesis, we are interested in the evaluation of policies in this finite-horizon MDP framework, which serves as the foundation for the subsequent chapters.

2.2 Discounted Markov Decision Process

In addition to the finite-horizon setting, reinforcement learning is also studied in the context of an infinite-horizon Markov Decision Process (MDP) with discounted rewards. A discounted MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r, p_0, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, p_0 is the initial state distribution, and $\gamma \in (0, 1)$ is the discount factor.

The agent interacts with the environment over an infinite sequence of time steps. At each time step t, the agent observes a state $S_t \in \mathcal{S}$, selects an action $A_t \in \mathcal{A}$ according to a stationary policy $\pi(a \mid s)$, receives a reward $R_{t+1} = r(S_t, A_t)$, and transitions to a successor state S_{t+1} sampled from $p(\cdot \mid S_t, A_t)$.

The discounted return at time step t is defined as:

$$G_t \doteq \sum_{i=t+1}^{\infty} \gamma^{i-t-1} R_i.$$

The state-value function $v_{\pi,t}(s)$, state-action-value function $q_{\pi,t}(s, a)$, and the expected total return $J(\pi)$ follow the same definitions as in the finite-horizon MDP.

The discount factor γ serves to weigh immediate rewards more heavily than distant future rewards and ensures the convergence of the infinite sum defining the return. The discounted infinite-horizon MDP setting is widely used due to its analytical convenience and relevance to long-term sequential decision-making problems.

2.3 Constrained Markov Decision Process

In addition to the standard MDP framework, we also consider a Constrained Markov Decision Process (CMDP), which incorporates an additional cost signal to model safety-critical scenarios. A (finite) CMDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r, c, p_0, T)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the cost function, p_0 is the initial state distribution, and T is the finite horizon length.

The CMDP process starts at time step 0 by sampling an initial state S_0 from p_0. At each time step $t \in \{0, 1, \ldots, T-1\}$, the agent selects an action $A_t \in \mathcal{A}$ according

to a policy $\pi_t(a \mid s)$, receives a reward $R_{t+1} = r(S_t, A_t)$ and a cost $C_{t+1} = c(S_t, A_t)$, and transitions to the next state S_{t+1} sampled from $p(\cdot \mid S_t, A_t)$.

The return for the reward and the cost at time step t are defined as:

$$G_t \doteq \sum_{i=t+1}^{T} R_i, \quad G_t^c \doteq \sum_{i=t+1}^{T} C_i.$$

The state-value and action-value functions for the reward are defined as:

$$v_{\pi,t}(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s], \quad q_{\pi,t}(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a].$$

Similarly, the state-value and action-value functions for the cost are defined as:

$$v_{\pi,t}^c(s) \doteq \mathbb{E}_\pi[G_t^c \mid S_t = s], \quad q_{\pi,t}^c(s,a) \doteq \mathbb{E}_\pi[G_t^c \mid S_t = s, A_t = a].$$

The overall performance of policy π is measured by the expected total reward:

$$J(\pi) \doteq \sum_{s \in \mathcal{S}} p_0(s) v_{\pi,0}(s).$$

Similarly, the total cost under policy π is defined as:

$$J^c(\pi) \doteq \sum_{s \in \mathcal{S}} p_0(s) v_{\pi,0}^c(s).$$

We denote the generated trajectory in a finite CMDP as

$$\tau = \{S_0, A_0, R_1, C_1, S_1, A_1, R_2, C_2, \ldots, S_{T-1}, A_{T-1}, R_T, C_T\},$$

which is simplified by the shorthand $\tau_{t:T-1}^{\mu_{t:T-1}}$.

2.4 Off-Policy Evaluation

In reinforcement learning, an important problem is to assess the performance of a policy without executing it in the environment. This problem is referred to as *off-policy evaluation* (OPE). The objective of OPE is to estimate the expected total reward $J(\pi)$ of a target policy π, using data generated by a different policy μ, known as the *behavior policy*.

Formally, we are interested in estimating:

$$J(\pi) \doteq \sum_{s \in \mathcal{S}} p_0(s) v_{\pi,0}(s),$$

where the value function $v_{\pi,0}(s)$ is defined under the target policy π, but the available data consists of trajectories generated under the behavior policy μ. Each trajectory collected under μ takes the form:

$$\{S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T\},$$

where $S_0 \sim p_0$ and $A_t \sim \mu_t(\cdot \mid S_t)$ for all t.

Off-policy evaluation is essential in settings where directly deploying the target policy is impractical due to safety concerns, high interaction costs, or low data efficiency. By leveraging data collected from the behavior policy, properly designed OPE enables practitioners to estimate policy performance with existing data or smaller amount of online interaction.

However, OPE faces fundamental challenges due to the *distribution mismatch* between the target and behavior policies. In particular, naively using behavior policy data to estimate $J(\pi)$ may result in biased estimates. Classical solutions address this issue by employing techniques such as *importance sampling* to reweight observed data and account for the difference between π and μ.

While importance sampling-based methods are known to suffer from high variance—especially when the target policy differs significantly from the behavior policy—recent work, including the contributions of this thesis, demonstrates that off-policy evaluation can be made efficient and reliable through proper algorithmic design and data collection strategies. In the next section, we will provide a detailed discussion of importance sampling and its various formulations for off-policy evaluation.

2.5 Importance Sampling for Off-Policy Evaluation

A standard approach to address the distribution mismatch in off-policy evaluation is the technique of *importance sampling*. Importance sampling adjusts the contribution of behavior policy data to reflect the target policy's distribution by reweighting the observed rewards.

For each trajectory collected under the behavior policy μ, we define the *importance sampling ratio* at time step t as:

$$\rho_t \doteq \frac{\pi_t(A_t \mid S_t)}{\mu_t(A_t \mid S_t)}.$$

The product of importance sampling ratios from time step t to $t' \geq t$ is denoted as:

$$\rho_{t:t'} \doteq \prod_{k=t}^{t'} \frac{\pi_k(A_k \mid S_k)}{\mu_k(A_k \mid S_k)}.$$

Several variants of importance sampling estimators have been proposed in the literature (Geweke, 1988; Hesterberg, 1995; Koller and Friedman, 2009; Thomas, 2015). A classical approach is the *trajectory-wise importance sampling estimator* (IS), which reweights the total return of an entire trajectory:

$$\mathrm{IS}(\tau) \doteq \rho_{0:T-1} \sum_{t=0}^{T-1} R_{t+1}.$$

While this estimator is unbiased under the standard policy coverage assumption (Precup et al., 2000a; Maei, 2011; Sutton et al., 2016; Zhang, 2022), it typically suffers from high variance, especially when the importance weights $\rho_{0:T-1}$ deviate significantly from one.

To mitigate this issue, the *per-decision importance sampling estimator* (PDIS) was introduced by Precup et al. (2000a). PDIS reweights rewards at each time step separately, leading to reduced variance compared to trajectory-wise IS. It is defined as:

$$G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \doteq \sum_{k=t}^{T-1} \rho_{t:k} R_{k+1}.$$

The PDIS estimator can also be expressed recursively as:

$$G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) = \begin{cases} \rho_t \left(R_{t+1} + G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \right) & t \in [T-2], \\ \rho_t R_{t+1} & t = T-1. \end{cases}$$

Under the standard policy coverage assumption,

$$\forall t, s, a, \quad \mu_t(a \mid s) = 0 \implies \pi_t(a \mid s) = 0,$$

both IS and PDIS estimators are unbiased. In this thesis, we further loosen this condition and characterize a broader class of behavior policies under which the PDIS estimator remains unbiased.

Traditionally, the variance issue associated with importance sampling has led to the belief that off-policy evaluation using such estimators is inherently inefficient. In this thesis, we revisit this belief and demonstrate that, with properly designed behavior policies and variance reduction techniques, importance sampling can be leveraged not only to correct distribution mismatch but also to serve as a *variance minimization tool*.

2.6 Fitted Q Evaluation

Fitted Q Evaluation (FQE) is a regression-based algorithm for offline policy evaluation. Its objective is to estimate the expected total return of a target policy using a fixed dataset collected by potentially unknown behavior policies.

FQE is motivated by the Bellman equation, which characterizes the action-value function $q_{\pi,t}(s,a)$ of a target policy π at time step t as:

$$q_{\pi,t}(s,a) = \mathbb{E}_\pi \left[R_{t+1} + q_{\pi,t+1}(S_{t+1}, A_{t+1}) \,\middle|\, S_t = s, A_t = a \right],$$

with the terminal condition $q_{\pi,T}(s,a) = 0$. This recursive structure motivates a backward dynamic programming approach for estimating $q_{\pi,t}$ from offline data.

Given an offline dataset $\mathcal{D} = \{(S_t^i, A_t^i, R_{t+1}^i, S_{t+1}^i)\}_{i=1}^n$, FQE iteratively estimates the Q-function at each time step. For each $t = T-1, T-2, \ldots, 0$, the Q-function is updated by solving:

$$q_t^{(\ell+1)} \in \arg\min_{q \in \mathcal{F}} \sum_{i=1}^n \left(q(S_t^i, A_t^i) - \left(R_{t+1}^i + \sum_{a \in \mathcal{A}} \pi_{t+1}(a \mid S_{t+1}^i) q_{t+1}^{(\ell)}(S_{t+1}^i, a) \right) \right)^2,$$

where $\ell \geq 0$ denotes the iteration index, and the initialization is $q_T^{(0)} \equiv 0$.

This backward procedure continues until the Q-function estimates converge. Once the estimates $\{q_t^{(\ell)}\}_{t=0}^{T-1}$ are obtained, the expected total reward of the target policy is estimated by:

$$J(\pi) \approx \sum_{s \in \mathcal{S}} p_0(s) \sum_{a \in \mathcal{A}} \pi_0(a \mid s) q_0^{(\ell)}(s,a).$$

In this thesis, we adopt FQE as a fundamental tool for offline policy evaluation. Specifically, we use FQE to estimate the extended value function of target policies from existing offline data, which in turn enables the learning of closed-form behavior policies for efficient online data collection.

2.7 Policy Gradient

Policy gradient methods are a foundational class of reinforcement learning algorithms that optimize the policy directly in the parameter space. Unlike value-based methods that learn action-value functions and derive the policy indirectly, policy gradient methods parameterize the policy as π_θ, where $\theta \in \mathbb{R}^d$ is a vector of parameters, and perform gradient ascent on the expected return.

Formally, the objective is to maximize the expected total reward of the policy:

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} R_{t+1}\right].$$

The policy gradient theorem (Sutton et al., 1999) provides a tractable expression for the gradient of $J(\theta)$ with respect to the policy parameters:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(A_t \mid S_t) G_t\right],$$

where $G_t = \sum_{i=t+1}^{T} R_i$ is the cumulative return starting from time step t. This gradient estimator is unbiased under the assumption that the trajectories are generated by the current policy π_θ.

Policy gradient algorithms, such as REINFORCE (Williams, 1992) and its variants, perform stochastic gradient ascent by updating the policy parameters as:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta),$$

where $\alpha > 0$ is the learning rate.

One of the key advantages of policy gradient methods is their applicability to high-dimensional or continuous action spaces, where value-based methods may struggle. However, a major challenge associated with policy gradient estimators is their high variance, which can result in unstable or sample-inefficient learning. Several variance reduction techniques have been proposed to address this issue, including the use of baselines (Greensmith et al., 2004), advantage functions (Schulman et al., 2016), and control variates (Wu et al., 2018).

Beyond policy optimization, the policy gradient framework also provides a powerful tool for sensitivity analysis and variance evaluation. In particular, the gradient of the return variance with respect to the policy parameters can be computed and optimized, enabling variance-aware learning and behavior policy design.

In this thesis, we leverage the idea for policy gradient to control the variance of policy evaluation. Specifically, we utilize a variance gradient formulation to control the variance of policy evaluation under adversarial transition dynamics and to guide the learning of behavior policies that are robust to environmental perturbations.

2.8 Linear Function Approximation

In large or continuous state spaces, representing the exact value function is often impractical due to the curse of dimensionality. To overcome this challenge, a common

approach is to approximate the value function using a linear combination of features. Specifically, let $\phi : \mathcal{S} \to \mathbb{R}^K$ denote a feature mapping from states to K-dimensional vectors. The approximate value function is then defined as

$$v_\theta(s) \doteq \phi(s)^\top \theta,$$

where $\theta \in \mathbb{R}^K$ is the parameter vector to be learned.

We further define the feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}| \times K}$, whose s-th row is given by $\phi(s)^\top$. Under this notation, the value function over all states can be written compactly as $v_\theta = \Phi\theta$.

Linear function approximation has been widely adopted in reinforcement learning, particularly in theoretical analyses and algorithm development, due to its simplicity and analytical tractability. Throughout this thesis, we assume access to such a fixed feature representation and adopt linear function approximation in both algorithmic design and theoretical analysis. In particular, the algorithms and stochastic approximation procedures studied in this thesis, including TD learning, GTD, and their stability analysis, are all developed under this setting.

2.9 Off-policy Temporal Difference Learning

Temporal Difference (TD) learning is a widely used method for estimating value functions by bootstrapping from future estimates. In the off-policy setting, data is collected under a behavior policy μ while evaluating a different target policy π. When combined with linear function approximation, the off-policy TD(λ) algorithm updates its value estimate θ_t using the following recursion:

$$\begin{aligned} e_t &= \lambda\gamma\rho_{t-1}e_{t-1} + \phi_t, \\ \delta_t &= R_{t+1} + \gamma\phi_{t+1}^\top\theta_t - \phi_t^\top\theta_t, \\ \theta_{t+1} &= \theta_t + \alpha_t\rho_t\delta_t e_t, \end{aligned}$$

where $\phi_t = \phi(S_t)$, $\rho_t = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$ is the importance sampling ratio, and e_t is the eligibility trace vector.

While off-policy TD methods are widely used, they can exhibit instability even with linear function approximation and regularization. Recent work shows that regularization may fail to prevent divergence and can introduce additional instability (Manek and Kolter, 2022). These issues motivate the development of more stable alternatives such as GTD, which optimizes a well-defined objective and enables theoretical convergence analysis.

2.10 Gradient Temporal Difference Learning

Temporal difference learning with linear function approximation is one of the most widely used approaches in reinforcement learning. However, when applied in the off-policy setting, conventional TD methods suffer from instability and divergence due to the so-called deadly triad. Gradient Temporal Difference (GTD) learning is a family of algorithms specifically designed to address this challenge by performing stochastic gradient descent on a well-defined objective function. Throughout this section, we assume linear function approximation.

The idea of GTD is to minimize the off-policy mean squared projected Bellman error (MSPBE) objective directly and use a weight duplication trick or Fenchel's duality to address the double sampling issue in estimating the gradient (Sutton et al., 2009; Liu et al., 2015). Various GTD variants have been developed in the literature (Sutton et al., 2008b, 2009; Maei, 2011; Yu, 2017; Zhang et al., 2021a; Qian and Zhang, 2025). In this thesis, we focus on the following representative variant, referred to as GTD(λ).

In GTD(λ), an additional weight vector $\nu \in \mathbb{R}^K$ is introduced, and θ and ν are updated recursively as follows:

$$\begin{aligned} e_t &= \lambda\gamma\rho_{t-1}e_{t-1} + \phi_t, \\ \delta_t &= R_{t+1} + \gamma\phi_{t+1}^\top\theta_t - \phi_t^\top\theta_t, \\ \nu_{t+1} &= \nu_t + \alpha_t\left(\rho_t\delta_t e_t - \phi_t\phi_t^\top\nu_t\right), \\ \theta_{t+1} &= \theta_t + \alpha_t\rho_t(\phi_t - \gamma\phi_{t+1})e_t^\top\nu_t. \end{aligned}$$

The introduction of the auxiliary weight vector ν facilitates the avoidance of the double sampling problem.

To analyze the algorithm, it is convenient to rewrite the updates compactly. Let $x_t \doteq \begin{bmatrix}\nu_t \\ \theta_t\end{bmatrix}$ and define an augmented Markov chain $\{Y_t\}$ as $Y_{t+1} \doteq (S_t, A_t, S_{t+1}, e_t)$. We further define:

$$\begin{aligned} A(y) &\doteq \rho(s,a)e(\gamma\phi(s') - \phi(s))^\top, \\ b(y) &\doteq \rho(s,a)r(s,a)e, \\ C(y) &\doteq \phi(s)\phi(s)^\top, \\ H(x,y) &\doteq \begin{bmatrix}-C(y) & A(y) \\ -A(y)^\top & 0\end{bmatrix}x + \begin{bmatrix}b(y) \\ 0\end{bmatrix}. \end{aligned}$$

Then, the GTD(λ) update can be rewritten as:

$$x_{t+1} = x_t + \alpha_t H(x_t, Y_{t+1}).$$

In this thesis, GTD(λ) serves as a fundamental example of stochastic approximation algorithms under Markovian noise, whose stability and convergence we study more generally.

2.11 Emphatic Temporal Difference Learning

Emphatic Temporal Difference (ETD) learning is a family of off-policy policy evaluation methods designed to stabilize learning in the presence of function approximation and distribution mismatch. The core idea is to reweight the standard off-policy linear TD updates with an additional emphasis factor that accounts for the mismatch between the target policy π and the behavior policy μ (Sutton et al., 2016; Yu, 2015). Similar to GTD, there are many variants of ETD. In this section, we focus on the original ETD(λ) algorithm (Yu, 2015; Sutton et al., 2016).

ETD(λ) updates the value parameter θ recursively with the help of a followon trace F_t, an eligibility trace e_t, and an interest function $i : \mathcal{S} \to (0, \infty)$:

$$\begin{aligned}
F_t &= \gamma \rho_{t-1} F_{t-1} + i(S_t), \\
M_t &= \lambda i(S_t) + (1 - \lambda) F_t, \\
e_t &= \lambda \gamma \rho_{t-1} e_{t-1} + M_t \phi_t, \\
\theta_{t+1} &= \theta_t + \alpha_t \rho_t \left(R_{t+1} + \gamma \phi_{t+1}^\top \theta_t - \phi_t^\top \theta_t \right) e_t.
\end{aligned}$$

The interest function $i(\cdot)$ encodes user-defined preferences over states, and is usually constant (e.g., $i(s) \equiv 1$), although non-trivial choices are possible.

Compared with GTD, ETD modifies the eligibility trace update by introducing the scalar M_t, which incorporates both the interest function and the followon trace F_t. This modification, known as the emphasis, enables ETD to better align the distribution of updates with the stationary distribution under the target policy.

ETD(λ) is included in this thesis as a representative instance of stochastic approximation procedures operating under Markovian noise, whose stability properties are studied under our proposed theoretical framework.

Chapter 3

Related Work

3.1 Variance Reduction in Policy Evaluation

Reducing the variance for policy evaluation in reinforcement learning (RL) has been widely studied. One rising approach is variance reduction by designing a proper data-collecting policy, also known as the behavior policy. Noticing that the target policy itself is not the best behavior policy, Hanna et al. (2017) formulate the task of finding a variance-reduction behavior policy as an optimization problem. They use stochastic gradient descent to update a parameterized behavior policy. However, the stochastic method has been known to easily get stuck in highly suboptimal points in just moderately complex environments, where various local optimal points exist (Williams, 1992). Moreover, their method requires highly sensitive hyperparameter tuning to learn the behavior policy effectively. Specifically, the learning rate can vary by up to 10^5 times across different environments, as reported in the experiments of Hanna et al. (2017). This extreme sensitivity requires online tuning, consuming massive online data. Furthermore, Hanna et al. (2017) constrain the online data to be complete trajectories.

Zhong et al. (2022) also aim to reduce the variance of policy evaluation through designing a proper behavior policy. They propose adjusting the behavior policy to focus on under-sampled data segments. Nevertheless, their method necessitates complete offline trajectories generated by known policies and assumes a strong similarity between the behavior and target policies, limiting the generalizability. Moreover, the estimates made by Zhong et al. (2022) lack theoretical guarantees of unbiasedness nor consistency. Another approach by Mukherjee et al. (2022) investigates behavior policies aimed at reducing variance in per-decision importance sampling estimators. However, their results are limited to tree-structured MDPs, a significant limitation since most problems do not adhere to tree structure. Moreover, Mukherjee et al.

(2022) explicitly require the knowledge of transition probability and, therefore, suffer from all canonical challenges in model learning (Sutton, 1990; Sutton et al., 2008a; Deisenroth and Rasmussen, 2011; Chua et al., 2018). The current state-of-the-art method in behavior policy design is proposed by Liu and Zhang (2024), where they find an optimal and offline-learnable behavior policy with the per-decision importance sampling estimator.

Besides behavior policy design, another popular approach for reducing the variance in policy evaluation is using the baseline functions. Jiang and Li (2016) propose a doubly robust estimator by incorporating a baseline function into the plain per-decision importance sampling estimator. However, their method assumes that the behavior policy is fixed and given, but does not discuss how to choose a proper behavior policy. Ignoring the choice of behavior policy loses the opportunity to save online samples manyfold. Thomas and Brunskill (2016) extend the method of Jiang and Li (2016) into the infinite horizon setting, proposing a weighted doubly robust estimator. However, their method introduces bias into the estimator, potentially leading the estimation to systematically deviate from the true return of the target policy.

3.2 Multi-Policy Evaluation

3.2.1 Multiple target policies

In multi-policy evaluation, traditional approaches often evaluate each policy separately using on-policy Monte Carlo methods. However, this ordinary method ignores the potential similarity between target policies and is crude for two reasons. First, the method does not utilize data sampled by other policies, causing the number of required online samples to scale quickly with the number of target policies. Second, even for a single target policy, the on-policy evaluation method is still not the optimal choice. Through a tailored behavior policy (Liu and Zhang, 2024), the variance of the on-policy Monte Carlo evaluation can be reduced while achieving an unbiased estimation.

To address the inefficiency in multi-policy evaluation problem, Dann et al. (2023) present an algorithm to reuse online samples from target policies. However, their algorithm works only when all target policies are deterministic, which is also highly restricted. The key difference is that they consider the plain approach by reusing samples from target policies, while we propose a tailored behavior for multiple target policies, which is designed to generate samples that all similar policies can efficiently share.

3.2.2 Multiple logging policies

Other approaches consider using data from multiple logging policies to perform off-policy evaluation, although only aiming at a single target policy. We call them logging policies because in their works (Agarwal et al., 2017; Lai et al., 2020; Kallus et al., 2021), data are previously logged from certain behavior policies and are fixed. This is different from our setting, in which we design an active data-collecting policy for multiple target policies.

Agarwal et al. (2017) point out that directly combining data from different policies may increases the estimation variance. They then propose two new estimators by reweighting data from different policies. However, their method is restricted to the contextual bandit setting. Lai et al. (2020) extend the method from Agarwal et al. (2017) into multi-step RL. Nevertheless, getting the desired weights for different logging policies requires knowing complicated covariance terms between every pair of logging policies. That is, given K logging policies, their method needs to compute K^2 covariances. Such strong prior knowledge is rarely available and is computationally expensive, making the method impractical. Furthermore, they ignore bias from any off-policy estimator. Kallus et al. (2021) also explore off-policy evaluation with multiple target policies in RL setting. They combine the reweighting strategy with the control variate method, leading to a reduced variance estimation. However, getting the weights proposed by their method requires knowledge of state visitation densities, whose approximation is very challenging in MDPs with large stochasticity and function approximation (cf. model-based RL (Sutton, 1990; Sutton et al., 2008a; Deisenroth and Rasmussen, 2011; Chua et al., 2018)). Due to this impracticability, Kallus et al. (2021) only conduct experiment of their method in the contextual bandit setting, remaining the experiment on the multi-step RL setting untouched.

3.3 Safe Reinforcement Learning

Safety in reinforcement learning, often framed as safe RL (Garcıa and Fernández, 2015), has been an active research topic recently. Many recent works focus on safety in policy exploration and optimization (Brunke et al., 2022). For safe exploration, Moldovan and Abbeel (2012) present a method for ensuring safe exploration by keeping the agent within a predefined set of safe states during its learning process. However, their method is a model-based approach, requiring an explicit approximation of the transition function, which introduces challenges common to model learning, such as compounding errors and the need for accurate model dynamics (Sutton, 1990;

Sutton et al., 2008a; Deisenroth and Rasmussen, 2011; Chua et al., 2018). As for safe optimization, Berkenkamp et al. (2017) propose to ensure safety by keeping the agent within safe regions, which are characterized by a Lyapunov function. However, they assume the environment to be deterministic, which is a significant limitation as most MDPs are stochastic. Their method is also model-based, requiring knowledge of the transition functions.

Safe reinforcement learning is often modeled as a Constrained Markov Decision Process (CMDP) (Gu et al., 2022; Liu et al., 2021b), in which we need to maximize the agent reward while making agents satisfy safety constraints. Achiam et al. (2017) enforce a constant threshold to constrain the expected total cost. However, even though they adopt the trust-region method to control policy updates, the expected total cost of the new policy can still exceed the safety threshold at each update step, leading to uncontrolled violations of the safety constraints over time. Wachi and Sui (2020) propose a method for safe reinforcement learning in constrained Markov decision processes (CMDPs) by using a Gaussian Process to model the safety constraints and guide exploration. Nevertheless, their approach needs to compute the covariance matrix between explored states throughout the execution, which is computationally expensive, especially in environments with large state spaces. In addition, they assume that the state transitions are deterministic, making their method highly restricted.

3.4 Robust Reinforcement Learning

Robustness in reinforcement learning has been extensively studied under the framework of Robust Markov Decision Processes (RMDPs), where uncertainty in transition dynamics is modeled explicitly. Classical RMDP formulations introduce uncertainty sets over transition probabilities and aim to optimize policies for worst-case outcomes within these sets (Iyengar, 2005). More recently, robust policy improvement methods have adopted a game-theoretic perspective, modeling adversarial perturbations to the environment. For instance, Wang et al. (2020) proposed a transition gradient method that computes adversarial transitions to degrade policy performance, thereby learning policies that are robust to such perturbations. These works focus primarily on policy optimization rather than evaluation, and their techniques are not directly applicable to off-policy evaluation (OPE), which requires unbiased and low-variance estimates of a fixed policy's performance.

Despite these advancements, the problem of robust policy evaluation, particularly in the off-policy setting, remains underexplored. Unlike policy optimization, where the

policy can adapt to adversarial perturbations, off-policy evaluation requires estimating the performance of a fixed policy accurately—even under dynamics shift. Existing OPE methods typically assume known and stationary dynamics (Hanna et al., 2017; Zhong et al., 2022; Mukherjee et al., 2022), leaving a gap in reliable evaluation techniques when the deployment environment differs from the one used for data collection. Our work addresses this gap by introducing a transition-gradient-based minimax formulation for robust off-policy evaluation.

3.5 Stability of Reinforcement Learning Algorithms

Stochastic approximation is central to the analysis of many reinforcement learning algorithms, with foundational results establishing almost sure convergence by linking stochastic updates to limiting ordinary differential equations (ODEs) (Borkar, 2009; Kushner and Yin, 2003). A key milestone in this area is Borkar and Meyn (2000), which ensures the stability of iterates under Martingale difference noise when the associated ODE at infinity is globally asymptotically stable. This result has been widely used in the analysis of temporal difference learning and other classical RL methods.

However, the Borkar-Meyn theorem and related results generally assume i.i.d. or Martingale difference noise, which does not hold in many RL applications where data is generated by a Markov process. To address this, previous work attempts to extend stability analysis to Markovian noise settings. Ramaswamy and Bhatnagar (2018) consider differential inclusions and require that the limiting update function be uniformly well-behaved for all noise realizations, which can be difficult to verify and does not hold in many standard RL algorithms. Another influential line of work (Kushner and Yin, 2003) focuses on projected stochastic approximation schemes that maintain stability by enforcing bounded iterates through projection operators. While effective in general settings, this approach introduces reflection terms into the limiting ODE, which complicates analysis and typically requires domain-specific arguments to handle properly.

These limitations have left a gap in the literature: a general and verifiable stability result for stochastic approximation under Markovian noise without resorting to restrictive assumptions, compactness, or projection schemes. This gap has particular relevance in reinforcement learning, where Markovian noise and unbounded traces are common.

Chapter 4

Efficient Policy Evaluation with Offline Data Informed Behavior Policy Design

This chapter is based on my paper Liu and Zhang (2024) published at ICML 2024.

In this chapter, we propose novel methods that improve the data efficiency of online Monte Carlo estimators while maintaining their unbiasedness. We first propose a tailored closed-form behavior policy that provably reduces the variance of an online Monte Carlo estimator. We then design efficient algorithms to learn this closed-form behavior policy from previously collected offline data. Theoretical analysis is provided to characterize how the behavior policy learning error affects the amount of reduced variance. Compared with previous works, our method achieves better empirical performance in a broader set of environments, with fewer requirements for offline data.

4.1 Preliminaries

We study a finite horizon Markov Decision Process (MDP, Puterman (2014)) as defined in Section 2.1. We use Monte Carlo methods, as introduced by Kakutani (1945), for estimating the total rewards $J(\pi)$. The most straightforward Monte Carlo method is to draw samples of $J(\pi)$ through the online execution of the policy π. The empirical average of the sampled returns converges to $J(\pi)$ as the number of samples increases. Since this method estimates a policy by executing itself, it is called on-policy learning (Sutton 1988).

Moving forward, we focus on off-policy evaluation, where the goal is to estimate the total rewards $J(\pi)$ of an interested policy π, which is called the *target policy*. Data for off-policy evaluation are collected by executing a different policy

μ, called the *behavior policy*. In off-policy evaluation, we generate each trajectory $\{S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T\}$ by a behavior policy μ with $A_t \sim \mu_t(\cdot|S_t)$. We use a shorthand $\tau_{t:T-1}^{\mu_{t:T-1}} \doteq \{S_t, A_t, R_{t+1}, \ldots, S_{T-1}, A_{T-1}, R_T\}$ for a trajectory generated by the behavior policy μ from time step t to $T-1$ inclusively. We use the importance sampling ratio to reweight the rewards obtained by the behavior policy μ, in order to give an estimate of $J(\pi)$. We define the importance sampling ratio at time step t as $\rho_t \doteq \frac{\pi_t(A_t|S_t)}{\mu_t(A_t|S_t)}$. Then, the product of importance sampling ratios from time t to $t' \geq t$ is defined as $\rho_{t:t'} \doteq \prod_{k=t}^{t'} \frac{\pi_k(A_k|S_k)}{\mu_k(A_k|S_k)}$. In off-policy learning, there are several ways to use the importance sampling ratios (Geweke, 1988; Hesterberg, 1995; Koller and Friedman, 2009; Thomas, 2015).

In this chapter, we adopt the per-decision importance sampling estimator (PDIS, Precup et al. (2000a)). We define the PDIS Monte Carlo estimator as $G^{\text{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \doteq \sum_{k=t}^{T-1} \rho_{t:k} R_{k+1}$, which can also be expressed recursively as

$$G^{\text{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) = \begin{cases} \rho_t \left(R_{t+1} + G^{\text{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}})\right) & t \in [T-2], \\ \rho_t R_{t+1} & t = T-1. \end{cases} \tag{4.1}$$

Under the classic policy coverage assumption (Precup et al., 2000a; Maei, 2011; Sutton et al., 2016; Zhang, 2022) $\forall t, s, a, \mu_t(a|s) = 0 \implies \pi_t(a|s) = 0$, this off-policy estimator $G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})$ provides an *unbiased* estimation for $J(\pi)$, i.e., $\mathbb{E}\left[G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right] = J(\pi)$.

4.2 Variance Reduction in Statistics

In this section, we provide the mathematical foundation for variance reduction with importance sampling ratios. The notations here are independent of the rest of this paper. We use similar notations only for easy interpretation in later sections. Consider a discrete random variable A taking values from a finite space $\mathcal{A}$ according to a probability mass function $\pi : \mathcal{A} \to [0, 1]$ and a function $q : \mathcal{A} \to \mathbb{R}$ mapping a value in $\mathcal{A}$ to a real number. We are interested in estimating $\mathbb{E}_{A\sim\pi}[q(A)]$. The ordinary Monte Carlo methods then sample $\{A_1, \ldots, A_N\}$ from π and use the empirical average $\frac{1}{N}\sum_{i=1}^{N} q(A_i)$ as the estimate. In statistics, importance sampling is introduced as a variance reduction technique for Monte Carlo methods (Rubinstein 1981). The main idea is to sample $\{A_i, \ldots, A_N\}$ from a different distribution μ and use $\frac{1}{N}\sum_{i=1}^{N} \rho(A_i)q(A_i)$ as the estimate, where $\rho(A) \doteq \frac{\pi(A)}{\mu(A)}$ is the importance sampling ratio. Assuming μ covers π, i.e.,

$$\forall a, \mu(a) = 0 \implies \pi(a) = 0, \tag{4.2}$$

the importance sampling ratio weighted empirical average is then unbiased, i.e.,

$$\mathbb{E}_{A\sim\pi}[q(A)] = \mathbb{E}_{A\sim\mu}[\rho(A)q(A)].$$

If the sampling distribution μ is carefully designed, the variance can also be reduced. To adapt this idea for RL, we relax the condition (4.2) in this section. We formulate this problem of searching a variance-reducing sampling distribution as an optimization problem:

$$\min_{\mu\in\Lambda_+} \quad \mathbb{V}_{A\sim\mu}(\rho(A)q(A)). \tag{4.3}$$

Here Λ_+ denotes the set of all the policies that give unbiased estimations, i.e.,

$$\Lambda_+ \doteq \{\mu \in \Delta(\mathcal{A}) \mid \mathbb{E}_{A\sim\mu}[\rho(A)q(A)] = \mathbb{E}_{A\sim\pi}[q(A)]\},$$

where $\Delta(\mathcal{X})$ denotes the set of all probability distributions on the set $\mathcal{X}$. Solving (4.3) is actually very challenging. To see this, consider a concrete example where $\mathcal{A} = \{a_1, a_2, a_3\}$ and

$$\begin{cases} q(a_1) = -10 \\ q(a_2) = 2 \\ q(a_3) = 2 \end{cases}, \quad \begin{cases} \pi(a_1) = 0.1 \\ \pi(a_2) = 0.5 \\ \pi(a_3) = 0.4 \end{cases}, \quad \begin{cases} \mu(a_1) = 0 \\ \mu(a_2) = 0 \\ \mu(a_3) = 1 \end{cases}. \tag{4.4}$$

It can be computed that $\mathbb{E}_{A\sim\pi}[q(A)] = 0.8$ and $\mathbb{E}_{A\sim\mu}[\rho(A)q(A)] = 0.8$. In other words, we could sample A from μ and use $\rho(A)q(A)$ as an estimator. This estimator is unbiased. But apparently, this μ does not cover π. Moreover, since μ is deterministic, the variance of this estimator is 0. Then μ is an optimal sampling distribution. However, μ is hand-crafted based on the knowledge that $q(a_1)\pi(a_1) + q(a_2)\pi(a_2) = 0$. Without such knowledge, we argue that there is little hope to find this μ. This example suggests that searching over the entire Λ_+ might be too ambitious. One natural choice presented by Rubinstein (1981) is to restrict the search to

$$\Lambda_- \doteq \{\mu \in \Delta(\mathcal{A}) \mid \forall a, \mu(a) = 0 \implies \pi(a) = 0\}. \tag{4.5}$$

In other words, we aim to find a variance-minimizing sampling distribution among all distributions that cover π.r we have $\Lambda_- \subseteq \Lambda_+$. In this work, we enlarge Λ_- to Λ defined as

$$\Lambda \doteq \{\mu \in \Delta(\mathcal{A}) \mid \forall a, \mu(a) = 0 \implies \pi(a)q(a) = 0\}. \tag{4.6}$$

following Owen (2013). The space Λ weakens the assumption in (4.5). Owen (2013) proves that any distribution μ in Λ gives unbiased estimation, though μ may not cover π.

Lemma 1. $\forall \mu \in \Lambda, \mathbb{E}_{A\sim\mu}\left[\rho(A)q(A)\right] = \mathbb{E}_{A\sim\pi}\left[q(A)\right].$

For completeness, its proof is in Appendix A.1.1. We now consider the variance minimization problem on Λ, i.e.,

$$\min_{\mu\in\Lambda} \quad \mathbb{V}_{A\sim\mu}(\rho(A)q(A)). \tag{4.7}$$

The following lemma from Owen (2013) gives an optimal solution μ^* to the optimization problem (4.7).

Lemma 2. *Define $\mu^*(a) \propto \pi(a)|q(a)|$. Then μ^* is an optimal solution to (4.7).*

For completeness, its proof is detailed in Appendix A.1. Here by

$$\mu(a) \propto \pi(a)w(a)$$

with some non-negative $w(a)$, we mean

$$\mu(a) \doteq \pi(a)w(a)/\textstyle\sum_b \pi(b)w(b).$$

The reader may notice that if $\pi(a)w(a) = 0$ for all a, the above "reweighted" distribution is not well defined. We then use the convention to interpret $\mu(a)$ as a uniform distribution, i.e., $\mu(a) = 1/|\mathcal{A}|$. *We adopt this convention in using $\propto$ in the rest of the paper to simplify the presentation.* The following lemma gives intuition on the optimality of μ^*, whose proof is in Appendix A.1.3.

Lemma 3. *If $\forall a \in \mathcal{A}, q(a) \geq 0$ or $\forall a \in \mathcal{A}, q(a) \leq 0$, then $\Lambda = \Lambda_+$, and the μ^* defined in Lemma 2 gives a zero variance, i.e., $\mathbb{V}_{A\sim\mu^*}(\rho(A)q(A)) = 0$.*

An optimal sampling distribution proportional to $\pi(a)|q(a)|$ dates back to Kahn and Marshall (1953); Rubinstein (1981); Benjamin Melamed (1998) and is commonly used in RL (Carpentier et al., 2015; Mukherjee et al., 2022). We, however, make two remarks. **First**, we show such a sampling distribution can be suboptimal in Λ_+. For (4.4), such a sampling distribution incurs strictly positive variance, but μ in (4.4) has a zero variance and is also unbiased. **Second**, different from existing literature in RL (Carpentier et al., 2015; Sutton and Barto, 2018; Mukherjee et al., 2022), our μ^* defined in Lemma 2 does not need to cover π. Nevertheless, we note that Lemma 1 still ensures that μ^* gives unbiased estimation (Owen, 2013) and extend unbiasedness to RL settings in Theorem 1.

4.3 Variance Reduction in Reinforcement Learning

We now apply the techniques in Section 4.2 in RL. In particular, we seek to reduce the variance $\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right)$ by designing a proper behavior policy μ. Of course, we need to ensure that the PDIS estimator with this behavior policy is unbiased. In other words, ideally we should search over

$$\Lambda_+ \doteq \left\{\mu \in \Delta(\mathcal{A})^T \mid \mathbb{E}\left[G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right] = J(\pi)\right\}.$$

As discussed in Section 4.2, this is too ambitious without domain-specific knowledge. Instead, we can search over all policies that cover π, i.e.,

$$\Lambda_- \doteq \{\mu \mid \forall t, s, a, \mu_t(a|s) = 0 \implies \pi_t(a|s) = 0\}.$$

The set Λ_- contains all policies that satisfy the policy coverage constraint in off-policy learning (Sutton and Barto 2018). Similar to (4.6), we can also enlarge Λ_- to

$$\Lambda \doteq \{\mu \mid \forall t, s, a, \mu_t(a|s) = 0 \implies \pi_t(a|s) q_{\pi,t}(s,a) = 0\}.$$

The following theorem ensures the desired unbiasedness, which is proved in Appendix A.1.4.

Theorem 1 (Unbiasedness)**.** $\forall \mu \in \Lambda$, $\forall t$, $\forall s$,
$\mathbb{E}\left[G^{PDIS}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right] = v_{\pi,t}(s)$.

One immediate consequence of Theorem 1 is that $\forall \mu \in \Lambda, \mathbb{E}\left[G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right] = J(\pi)$. In this chapter, we consider a set Λ^* such that $\Lambda_- \subseteq \Lambda^* \subseteq \Lambda$. This Λ^* inherits the unbiasedness property of Λ and is less restrictive than Λ_-, the classical search space of behavior policies. This Λ^* will be defined shortly. We now formulate our problem as

$$\min_{\mu \in \Lambda^*} \quad \mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right). \tag{4.8}$$

By the law of total variance, for any $\mu \in \Lambda^*$, we decompose the variance of the PDIS estimator as

$$\begin{aligned}
&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right) \\
=&\mathbb{E}_{S_0}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}}) \mid S_0\right)\right] \\
&+ \mathbb{V}_{S_0}\left(\mathbb{E}\left[G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}}) \mid S_0\right]\right) \\
=&\mathbb{E}_{S_0}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}}) \mid S_0\right)\right] + \mathbb{V}_{S_0}\left(v_{\pi,0}(S_0)\right). \qquad \text{(by Theorem 1)}
\end{aligned}$$

The second term $\mathbb{V}_{S_0}\left(v_{\pi,0}(S_0)\right)$ is a constant given a target policy π and is unrelated to the choice of μ. In the first term, the expectation is taken over S_0 that is determined by the initial probability distribution p_0. Consequently, to solve the problem (4.8), it is sufficient to solve for each s,

$$\min_{\mu\in\Lambda^*} \quad \mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}}) \mid S_0 = s\right). \tag{4.9}$$

Denote the variance of the state value for the next state given the current state-action pair (s,a) as $\nu_{\pi,t}(s,a)$. We have $\nu_{\pi,t}(s,a) = 0$ for $t = T-1$ and otherwise

$$\nu_{\pi,t}(s,a) \doteq \mathbb{V}_{S_{t+1}}\left(v_{\pi,t+1}(S_{t+1}) \mid S_t = s, A_t = a\right). \tag{4.10}$$

We now construct a behavior policy μ^* as

$$\mu_t^*(a|s) \propto \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)}, \tag{4.11}$$

where $u_{\pi,t}(s,a) \doteq q_{\pi,t}^2(s,a)$ for $t = T-1$ and otherwise

$$\begin{aligned} u_{\pi,t}(s,a) &= q_{\pi,t}^2(s,a) + \nu_{\pi,t}(s,a) \\ &+ \textstyle\sum_{s'} p(s'|s,a)\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^*}) \mid S_{t+1} = s'\right). \end{aligned} \tag{4.12}$$

Notably, μ_t^* and $u_{\pi,t}$ are defined backwards and alternatively, i.e., they are defined in the order of $u_{\pi,T-1}, \mu_{T-1}^*, u_{\pi,T-2}, \mu_{T-2}^*, \ldots, u_{\pi,0}, \mu_0^*$. We prove μ^* is optimal in the following sense.

Theorem 2 (Optimal Behavior Policy)**.** *For any t and s, the behavior policy $\mu_t^*(a|s)$ defined above is an optimal solution to the following problem*

$$\min_{\mu_t\in\Lambda_t,\ldots,\mu_{T-1}\in\Lambda_{T-1}} \quad \mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right),$$

where $\Lambda_t \doteq \{\mu_t \in \Delta(\mathcal{A}) \mid \forall s, a, \mu_t(a|s) = 0 \implies \pi_t(a|s)u_{\pi,t}(s,a) = 0\}$.

Its proof is in Appendix A.1.5. We are now ready to define $\Lambda^* \doteq \Lambda_0 \times \cdots \times \Lambda_{T-1}$. Theorem 2 indicates that μ^* achieves optimality for the optimization problem (4.9). Since $u_{\pi,t}(s,a) = 0 \implies q_{\pi,t}(s,a) = 0$ by the non-negativity of the summands in (4.12), we have $\Lambda^* \subseteq \Lambda$. If $\mu_t(a|s) = 0 \implies \pi_t(a|s) = 0$, it follows immediately that $\mu_t(a|s) = 0 \implies \pi_t(a|s)u_{\pi,t}(s,a) = 0$. This indicates $\Lambda_- \subseteq \Lambda^*$. This means that the set of policies Λ^* considered in Theorem 2 are unbiased and includes at least all the policies that cover the target policy, which is the classical behavior policy search space Λ_- (Precup et al., 2000a; Maei, 2011; Sutton et al., 2016; Zhang, 2022).

Unfortunately, empirically implementing μ_t^* requires knowledge of $u_{\pi,t}$ (4.12) that contains the transition function p. Approximating the transition function is very challenging in MDPs with large stochasticity and function approximation (cf. model-based RL (Sutton, 1990; Sutton et al., 2008a; Deisenroth and Rasmussen, 2011; Chua et al., 2018)). Thus, we seek to build another policy $\hat{\mu}$ that can be easily implemented without direct knowledge of the transition function p (cf. model-free RL (Sutton, 1988; Watkins, 1989)).

We achieve this by aiming at one-step optimality instead of global optimality. We try to find the best μ_t assuming in the future we follow $\pi_{t+1}, \dots, \pi_{T-1}$, instead of $\mu_{t+1}^*, \dots, \mu_{T-1}^*$. We refer to this one-step optimal behavior policy as $\hat{\mu}_t$. Similarly, to define optimality, we first need to specify the set of policies we are concerned about. To this end, we define

$$\hat{q}_{\pi,t}(s,a) \doteq q^2_{\pi,t}(s,a) \tag{4.13}$$

for $t = T-1$ and otherwise

$$\begin{aligned} &\hat{q}_{\pi,t}(s,a) \doteq q^2_{\pi,t}(s,a) + \nu_{\pi,t}(s,a) \\ &+ \textstyle\sum_{s'} p(s'|s,a) \mathbb{V}\left(G^{\text{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1} = s'\right). \end{aligned} \tag{4.14}$$

Notably, $\hat{q}_{\pi,t}(s,a)$ is always non-negative since all the summands are non-negative. Accordingly, we define for $t \in [T-1]$, $\hat{\Lambda}_t \doteq \{\mu_t \in \Delta(\mathcal{A}) \mid \forall s, a, \mu_t(a|s) = 0 \implies \pi_t(a|s)\hat{q}_{\pi,t}(s,a) = 0\}$. Comparing (4.12) and (4.14), the optimality of μ^* implies that $\forall s, a, t$, we have $\hat{q}_{\pi,t}(s,a) \geq u_{\pi,t}(s,a) \geq 0$. As a result, if $\mu_t \in \hat{\Lambda}_t$, we have

$$\begin{aligned} \mu_t(a|s) = 0 &\implies \pi_t(a|s)\hat{q}_{\pi,t}(a|s) = 0 \\ &\implies \pi_t(a|s)u_{\pi,t}(a|s) = 0, \end{aligned}$$

indicating $\mu_t \in \Lambda_t$. In other words, we have $\hat{\Lambda}_t \subseteq \Lambda_t$. To search for $\hat{\mu}_{0:T-1}$, we work on $\hat{\Lambda} \doteq \hat{\Lambda}_0 \times \cdots \times \hat{\Lambda}_{T-1}$. To summarize, we have $\Lambda_- \subseteq \hat{\Lambda} \subseteq \Lambda^* \subseteq \Lambda \subseteq \Lambda_+$. Recall that Λ_+ is the set of all behavior policies such that the corresponding PDIS estimator is unbiased. Λ is a sufficient but not necessary condition to ensure such unbiasedness (Theorem 1). Λ^* is a restriction of Λ such that we are able to find an optimal solution. We restrict Λ^* to $\hat{\Lambda}$, aiming for a sub-optimal but implementable policy. $\hat{\Lambda}$ is still larger than Λ_-, which is the space with the coverage assumption (4.2) that previous works (Precup et al., 2000a; Maei, 2011; Sutton et al., 2016; Sutton and Barto, 2018; Zhang, 2022) consider.

After confirming the space of behavior policies, we formulate the optimization problem for designing an efficient behavior policy to achieve one-step optimality as

$$\min_{\mu_t \in \hat{\Lambda}_t} \quad \mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\{\mu_t, \pi_{t+1}, \ldots, \pi_{T-1}\}}) \mid S_t = s\right). \tag{4.15}$$

According to the recursive expression of the variance in Lemma 34 in Appendix A.1.5, we rewrite (4.15) as

$$\begin{aligned}\min_{\mu_t \in \hat{\Lambda}_t} \mathbb{E}_{A_t \sim \mu_t} \Big[\rho_t^2 \Big(\mathbb{E}_{S_{t+1}} \Big[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) \\ \mid S_t, A_t\Big] + \nu_{\pi,t}(S_t, A_t) + q_{\pi,t}^2(S_t, A_t)\Big) \mid S_t\Big],\end{aligned} \tag{4.16}$$

where the objective can be further simplified as

$$\begin{aligned}&\mathbb{E}_{A_t \sim \mu_t} \Big[\rho_t^2 \Big(\mathbb{E}_{S_{t+1}} \Big[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) \\ &\quad \mid S_t, A_t\Big] + \nu_{\pi,t}(S_t, A_t) + q_{\pi,t}^2(S_t, A_t)\Big) \mid S_t\Big] \\ =&\mathbb{E}_{A_t \sim \mu_t} \left[\rho_t^2 \hat{q}_{\pi,t}(S_t, A_t) \mid S_t\right] && \text{(By (4.14))} \\ =&\mathbb{V}_{A_t \sim \mu_t}\left(\rho_t \sqrt{\hat{q}_{\pi,t}(S_t, A_t)} \mid S_t\right) \\ &- \mathbb{E}_{A_t \sim \pi_t}^2 \left[\sqrt{\hat{q}_{\pi,t}(S_t, A_t)} \mid S_t\right]. && \text{(Lemma 1 and } \mu_t \in \hat{\Lambda}_t)\end{aligned}$$

Since the second term is unrelated to μ_t, it is equivalent to solving

$$\min_{\mu_t \in \hat{\Lambda}_t} \quad \mathbb{V}_{A_t \sim \mu_t}\left(\rho_t \sqrt{\hat{q}_{\pi,t}(S_t, A_t)} \mid S_t\right).$$

According to Lemma 2,

$$\hat{\mu}_t(a|s) \propto \pi_t(a|s) \sqrt{\hat{q}_{\pi,t}(s,a)}. \tag{4.17}$$

is an optimal solution to (4.16). We now present our main result that $\hat{\mu}$ *provably* reduces variance.

Theorem 3 (Variance Reduction)**.** *For any t and s,*

$$\begin{aligned}&\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}) \mid S_t = s\right) \\ \leq&\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) - \epsilon_t(s).\end{aligned}$$

To define $\epsilon_t(s)$, first define $c_t(s) =$

$$\textstyle\sum_a \pi_t(a|s) \hat{q}_{\pi,t}(s,a) - \left(\sum_a \pi_t(a|s) \sqrt{\hat{q}_{\pi,t}(s,a)}\right)^2.$$

Then we define $\epsilon_t(s) \doteq c_t(s)$ for $t = T-1$ and otherwise

$$\epsilon_t(s) \doteq c_t(s) + \mathbb{E}_{A_t \sim \hat{\mu}_t}\left[\rho_t^2 \mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) | s, A_t\right]\right]. \tag{4.18}$$

Its proof is in Appendix A.1.6. Notably, this c_t is always non-negative by Jensen's inequality, ensuring the non-negativity of ϵ_t and thus the variance reduction property. Moreover, $c_t(s) = 0$ occurs only when all actions have the same $\hat{q}_{\pi,t}$ on the state s. It is reasonable to conjecture that this is rare in practice. So, $c_t(s)$ is likely to be strictly positive. This shows the variance of the PDIS estimator with $\hat{\mu}$ at a state s is *provably* smaller than or equal to that with π, the straightforward on-policy Monte Carlo estimator, by at least $\epsilon_t(s)$. The magnitude of $\epsilon_t(s)$ depends on a specific target policy and the environment. We empirically show the variance reduction is significant in commonly used benchmarks in Section 4.5.

4.4 Learning Closed-Form Behavior Policies

We now present efficient algorithms to learn the closed-form behavior policy $\hat{\mu}$. Despite that $\hat{q}_{\pi,t}$ in (4.14) has a complicated definition, we prove that it has a concise representation. It is exactly the action value function of the policy π with the same transition function p but a different reward function $\hat{r}$.

Theorem 4. *Define*

$$\hat{r}_{\pi,t}(s,a) \doteq 2r(s,a)q_{\pi,t}(s,a) - r^2(s,a). \tag{4.19}$$

Then $\hat{q}_{\pi,t}(s,a) = \hat{r}_{\pi,t}(s,a)$ *for* $t = T-1$ *and otherwise*

$$\begin{aligned} &\hat{q}_{\pi,t}(s,a) \\ =&\hat{r}_{\pi,t}(s,a) + \textstyle\sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')\hat{q}_{\pi,t+1}(s',a'). \end{aligned} \tag{4.20}$$

Its proof is in Appendix A.1.7. *This observation makes it possible to apply any off-the-shelf offline policy evaluation methods to learn* $\hat{q}$, after which the behavior policy $\hat{\mu}$ can be computed easily with (4.17). For generality, we consider the behavior policy agnostic offline learning setting (Nachum et al., 2019), where the offline data in the form of $\{(t_i, s_i, a_i, r_i, s'_i)\}_{i=1}^m$ consists of m previously logged data tuples. In the i-th data tuple, t_i is the time step, s_i is the state at time step t_i, a_i is the action executed on state s_i, r_i is the sampled reward, and s'_i is the successor state. Those tuples can be generated by one or more, known or unknown behavior policies. Those tuples do not need to form a complete trajectory.

In this chapter, we choose Fitted Q-Evaluation (FQE, Le et al. (2019)) as a demonstration, but our framework is ready to incorporate any state-of-the-art offline policy evaluation methods to approximate $\hat{q}$. To learn $\hat{r}$, it is sufficient to learn r and

Algorithm 1: Offline Data Informed (ODI) algorithm

Input: Estimators $r(s,a)$, $q_{\pi,t}(s,a)$, $\hat{q}_{\pi,t}(s,a)$,
a target policy π,
an offline dataset $\mathcal{D} = \{(t_i, s_i, a_i, r_i, s_i)\}_{i=1}^m$
Output: a behavior policy $\hat{\mu}$
Approximate r from $\mathcal{D}$ using supervised learning
Approximate $q_{\pi,t}$ from $\mathcal{D}$ using any offline RL method (e.g. Fitted Q-Evaluation)
Compute $\hat{r}_i$ by (4.19) for each data pair in $\mathcal{D}$
Construct $\mathcal{D}_{\hat{r}} \doteq \{(t_i, s_i, a_i, \hat{r}_i, s_i)\}_{i=1}^m$ by plugging $\hat{r}_i$ into $\mathcal{D}$
Approximate $\hat{q}_{\pi,t}$ from $\mathcal{D}_{\hat{r}}$ by (4.20) using any offline RL method (e.g. Fitted Q-Evaluation)
Return: $\hat{\mu}_t(a|s) \propto \pi_t(a|s)\sqrt{\hat{q}_{\pi,t}(s,a)}$

q. FQE can be used to learn q, and learning r is a simple regression problem. FQE is then invoked again w.r.t. the learned $\hat{r}$ to learn an approximation of $\hat{q}$. We refer the reader to Algorithm 1 for a detailed exposition of our algorithm. We split the offline data into training sets and test sets to tune all the hyperparameters offline in Algorithm 1, based on the supervised learning loss or the FQE loss on the test set. We remark that FQE loss on the test set is known to be an inaccurate signal (Fujimoto et al., 2022) so our $\hat{q}$ estimation would be poorly tuned in this sense. We, however, notice that even with such a poorly tuned estimation, the variance reduction in the tested environments is still significant. This suggests that $\epsilon_t(s)$ in Theorem 3 is likely to be large and demonstrates the robustness of our approach. Since $\hat{q}_{\pi,t}(s,a)$ is proved to be always non-negative (cf. (4.14)), we use positive function class for FQE in approximating $\hat{q}$, e.g., a neural network with softplus as the last activation function.

In the following, we theoretically analyze how the error in approximating $\hat{q}$ affects the amount of reduced variance in Theorem 3. We assume $\hat{q}_{\pi,t}(s,a)$ is not only non-negative but also positive. Given its non-negative summands in (4.14), we argue that this positivity assumption is not restrictive at all. We use $q^+_{\pi,t}(s,a) > 0$ to denote our approximation to $\hat{q}_{\pi,t}(s,a)$. The approximation error can then be captured by

$$\eta_{\pi,t}(s,a) \doteq \hat{q}^+_{\pi,t}(s,a)/\hat{q}_{\pi,t}(s,a) > 0. \tag{4.21}$$

If $\eta_{\pi,t}(s,a)$ is 1, there is no approximation error for (s,a,t). The actual learned behavior policy is then denoted by

$$\hat{\mu}^+_t(a|s) \propto \pi_t(a|s)\sqrt{\hat{q}^+_{\pi,t}(s,a)}. \tag{4.22}$$

Then, we generalize Theorem 3 to the following theorem.

Theorem 5. *For any t and s,*

$$\begin{aligned}&\mathbb{V}(G^{PDIS}(\tau_{t:T-1}^{\hat{\mu}^+_{t:T-1}}) \mid S_t = s)\\ \leq&\mathbb{V}(G^{PDIS}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s) - \epsilon_t^+(s).\end{aligned}$$

To define $\epsilon_t^+(s)$, first define

$$\begin{aligned}c_t^+(s) \doteq& \textstyle\sum_a \pi_t(a|s)\hat{q}_{\pi,t}(s,a) -\\ &\left(\textstyle\sum_a \pi_t(a|S_t)\sqrt{\eta_{\pi,t}(S_t,a)}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)\\ &\times\left(\textstyle\sum_a \pi_t(a|S_t)\frac{1}{\sqrt{\eta_{\pi,t}(S_t,a)}}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right).\end{aligned}$$

Then we define $\epsilon_t^+(s) \doteq c_t^+(s)$ for $t = T-1$ and otherwise

$$\begin{aligned}&\epsilon_t^+(s)\\ \doteq& c_t^+(s) + \mathbb{E}_{A_t\sim\hat{\mu}_t^+}[\rho_t^2\mathbb{E}_{S_{t+1}}[\epsilon_{t+1}^+(S_{t+1})|s,A_t]].\end{aligned} \tag{4.23}$$

Its proof is in Appendix A.1.8. When there is no estimation error, i.e., $\eta_{\pi,t}(s,a) = 1$, c_t^+ and ϵ_t^+ reduce to c_t and ϵ_t in Theorem 3, which is non-negative by Jensen's inequality. As discussed earlier, it is reasonable to conjecture that $c_t(s)$ is likely to be strictly positive. This leaves room to tolerate estimation errors such that $c_t^+(s)$ can still be positive even if $\eta_t(s,a) \neq 1$. Because the sign of c_t^+ only depends on the current $\eta_{\pi,t}$, the estimation error in the future step does not affect current c_t. Notably, even if some $\epsilon_{t+1}^+(S_{t+1}) < 0$, $\epsilon_t^+(S_t)$ can still be positive. This is because $\epsilon_t^+(s)$ depends on the expectation of the $\epsilon_{t+1}^+(S_{t+1})$, not a single value, and c_t^+ can still be positive. This makes our approach robust to the approximation error. *It is important to note that the PDIS estimator with $\hat{\mu}_t(a|s)$ is always unbiased, regardless of the approximation error η.*

Theorem 5 makes it straightforward to analyze how the offline data affects the amount of the reduced variance. For example, if FQE is used, one can resort to Munos (2003); Antos et al. (2008); Munos and Szepesvári (2008); Chen and Jiang (2019) to connect offline data and the approximation error η. Theorem 5 then directly relays the approximation error to the amount of reduced variance. We, however, omit such analysis since it deviates from our main contribution.

4.5 Empirical Results

In this section, we present empirical results comparing our methods against three baselines: **(1)** the canonical on-policy Monte Carlo estimator, **(2)** off-policy Monte

On-policy MC	Ours with 2.3% data coverage	Ours with 4.6% data coverage	Ours with 18.4% data coverage	BPG	ROS
300	150	90	60	300	300
600	330	180	120	540	540
1200	540	420	270	990	990

Table 4.1: The above table is an extension of Figure 4.1 by adding experiments with 4.6%/18.4% offline data coverage for our algorithm in Gridworld with size $= 27,000$. Each number is the number of steps needed to achieve the same estimation accuracy that the naive Monte Carlo achieves with 300/600/1200 steps. All numbers are averaged from 900 different runs over a wide range of policies. Standard errors are visualized in Figure 4.1 of our paper and are invisible for some algorithm curves because they are too small.

Carlo estimator with behavior policy search (BPS, Hanna et al. (2017)), and **(3)** robust on-policy sampling (ROS, Zhong et al. (2022)). We do not implement ReVar (Mukherjee et al., 2022) because it will incur infinite loops if the MDP is not tree-structured. Our method first learns a behavior policy with given offline data using Algorithm 1, then the PDIS Monte Carlo estimator (4.1) is used to estimate the performance of the target policy, where the learned behavior policy is used to interact with the environment. We call our method Offline Data Informed (ODI) algorithm. Our method is superior in data requirements and applicability as summarized in Table 4.2.

Gridworld: We first conduct experiments with linear function approximation in Gridworld with n^3 states, i.e., it is an $n \times n$ grid with the time horizon also being n. Specifically, we use Gridworld with $n^3 = 1,000$ and $n^3 = 27,000$. We use randomly generated reward functions with 30 randomly generated target policies. The offline data is generated by selecting random actions on uniformly random state distribution. We report the *normalized estimation error* of the four methods against the number of environment interactions (steps). Intuitively, this normalized estimation error is the estimation error of an estimator normalized by that of the on-policy Monte Carlo estimator. Precisely speaking, define the *estimation error* at step t as the absolute difference between an estimator and the ground truth divided by the ground truth. The *normalized estimation error* is then the estimation error divided by the average estimation error of the on-policy Monte Carlo estimator after the first episode. Thus, the normalized estimation error of the on-policy Monte Carlo estimator starts from 1.

	On-policy MC	Ours	BPG	ROS	Improvement in Saved Episodes
Ant	100	**81**	91	103	(100-81)/(100-91)≈211.1%
Hopper	100	**54**	89	100	(100-54)/(100-89)≈418.2%
I. Pendulum	100	**72**	103	99	(100-72)/(100-99)=2800%
I. D. Pendulum	100	**35**	95	90	(100-35)/(100-90)=650%
Walker	100	**70**	92	91	(100-70)/(100-91)≈333.3%

Table 4.2: Episodes needed to achieve the same of estimation accuracy that on-policy Monte Carlo achieves with 100 episodes.

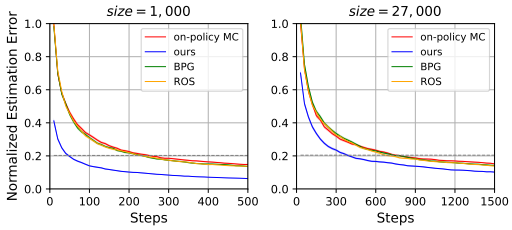

Figure 4.1: Results on Gridworld. The curves are averaged over 900 trials (30 target policies, each having 30 independent runs). The shaded regions denote standard errors and are invisible for some curves because they are too small.

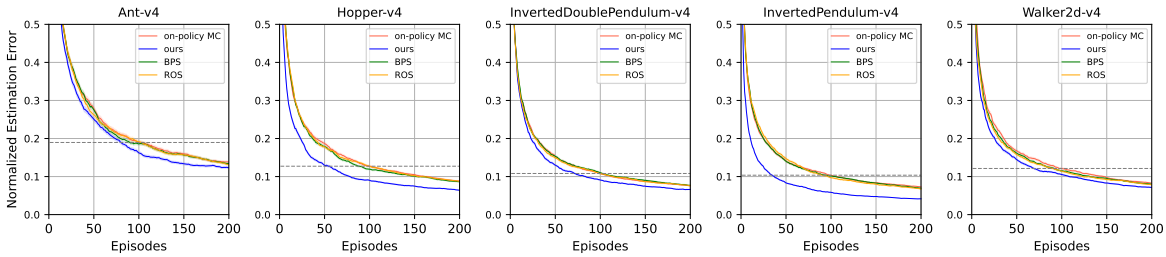

Figure 4.2: Results on Mujoco environments. Each curve is averaged over 900 trials (30 target policies, each having 30 independent runs). The shaded regions denote standard errors and are invisible for some curves because they are too small.

As shown in Figure 4.1, our method outperforms baselines by a large margin. In particular, as shown by the dotted line, in Gridworld with size $1,000$, to achieve the same estimation error that the on-policy Monte Carlo estimator achieves with 250 steps, our methods only need around 50 steps. In Gridworld with size $27,000$, to achieve the same estimation error that on-policy Monte Carlo estimator achieves with 750 steps, our methods only need around 400 steps, saving more than 40% of online iteractions. The improvement in environments with size $= 27,000$ is smaller than environments with size $= 1,000$ because the amount of offline data is the same for

both environments, i.e., the offline data coverage is worse for the Gridworld with size $= 27,000$. In fact, the offline data coverage for the Gridworld with size $= 1,000$ and size $= 27,000$ are 62.5% and 2.3%, respectively. More experiment details are in Appendix A.2.1.

MuJoCo: We then conduct experiments with neural network function approximation in MuJoCo (Todorov et al., 2012) robot simulation tasks. Since our methods are designed for discrete action space, we discretize the MuJoCo action space. Details about action space discretization, target policy generation, and offline data generation are provided in Appendix A.2.2. We report the normalized estimator error in Figure 4.2, where our methods are consistently better than baselines. In particular, as shown by the dotted line in Figure 4.2 and Table 4.2, our methods need much fewer episodes (save up to 65% episodes) to achieve the estimation error that the on-policy Monte Carlo estimator achieves with 100 episodes. Recognizing episodes may have different lengths in MuJoCo, we also provide in Appendix A.2.2 a version of Figure 4.2 with the x-axis being steps, where our methods are still consistently better.

It is worth mentioning that all hyperparameters of our methods required to learn $\hat{\mu}$ are tuned offline and are the same across all MuJoCo and Gridworld experiments.

4.6 Discussion

Monte Carlo methods are the most dominant approach for evaluating a policy. The development and deployment of almost all RL algorithms, including offline RL algorithms, implicitly or explicitly depend on Monte Carlo methods more or less. For example, when an RL researcher wants to plot a curve of the agent performance against training steps, Monte Carlo methods are usually the first choice. Our method improves the online data efficiency of Monte Carlo evaluation while maintaining its unbiasedness by learning a tailored behavior policy from offline data. The two main contributions are the provably better closed-form behavior policy (Theorem 3) and its alternative representation (Theorem 4). Extending them to temporal difference learning (Sutton, 1988) is a possible future work.

Chapter 5

Efficient Multi-Policy Evaluation for Reinforcement Learning

This chapter is based on my paper Liu et al. (2025c) published at AAAI 2025 with an oral presentation honor.

To unbiasedly evaluate multiple target policies, the dominant approach among RL practitioners is to run and evaluate each target policy separately. However, this evaluation method is far from efficient because samples are not shared across policies, and running target policies to evaluate themselves is actually not optimal. In this chapter, we address these two weaknesses by designing a tailored behavior policy to reduce the variance of estimators across all target policies. Theoretically, we prove that executing this behavior policy with manyfold fewer samples outperforms on-policy evaluation on every target policy under characterized conditions. Empirically, we show our estimator has a substantially lower variance compared with previous best methods and achieves state-of-the-art performance in a broad range of environments.

5.1 Preliminaries

We consider the task of multi-policy evaluation in the context of a finite Markov Decision Process (MDP), as defined in Section 2.1. Specifically, we aim to evaluate a set of K target policies. In this chapter, any index with parenthesis around it (e.g. $\pi^{(k)}$) is related to the *policy index*. We define abbreviations $\pi^{(k)}_{i:j} \doteq \left\{\pi^{(k)}_i, \pi^{(k)}_{i+1}, \ldots, \pi^{(k)}_j\right\}$ and $\pi^{(k)} \doteq \pi^{(k)}_{0:T-1}$, where $\pi^{(k)}_t : \mathcal{A} \times \mathcal{S} \to [0, 1]$ defines the probability of selecting action A_t given the state S_t at time $t \in [T-1]$. An initial state S_0 is sampled from p_0 at time step 0. At each time step t, after the execution of an action, a finite reward $R_{t+1} = r(S_t, A_t)$ is obtained and a successor state S_{t+1} is sampled from $p(\cdot \mid S_t, A_t)$.

We define the return at time step t as $G_t \doteq \sum_{i=t+1}^{T} R_i$. The state- and action-value function is defined as

$$v_{\pi^{(k)},t}(s) \doteq \mathbb{E}_{\pi^{(k)}}\left[G_t \mid S_t = s\right]$$

and

$$q_{\pi^{(k)},t}(s,a) \doteq \mathbb{E}_{\pi^{(k)}}\left[G_t \mid S_t = s, A_t = a\right].$$

The performance of the policy π is defined as $J(\pi^{(k)}) \doteq \sum_s p_0(s) v_{\pi^{(k)},0}(s)$. We adopt the total rewards performance metric, introduced by Puterman (2014), as a measurement of the performance. In this work, we focus on the Monte Carlo methods, which have been widely adopted since their introduction by Kakutani (1945). We draw samples of $J(\pi^{(k)})$ by executing the policy $\pi^{(k)}$ online. The empirical average of the sampled returns converges to $J(\pi^{(k)})$ as the number of samples increases. Since this method estimates the performance of a policy $\pi^{(k)}$ by running itself, it is called on-policy learning (Sutton 1988).

Henceforth, we study off-policy learning, in which we need to estimate the total rewards $J(\pi^{(k)})$ of a policy $\pi^{(k)}$, called the target policy, by running a different policy μ, known as the behavior policy. Each trajectory $\{S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T\}$ is generated by a behavior policy μ with $S_0 \sim p_0$, $A_t \sim \mu_t(\cdot|S_t)$, $t \in [T-1]$. We use

$$\tau_{t:T-1}^{\mu_{t:T-1}} \doteq \{S_t, A_t, R_{t+1}, \ldots, S_{T-1}, A_{T-1}, R_T\}$$

to denote a segment of a random trajectory generated by the behavior policy μ from the time step t to the time step $T-1$ inclusively. The key tool for off-policy learning is importance sampling (IS) (Rubinstein, 1981), which is used to reweight rewards collected by μ to give an unbiased estimate of $J(\pi^{(k)})$. For each policy $\pi^{(k)}$, the importance sampling ratio at time step t is defined as $\rho_t^{\pi^{(k)},\mu} \doteq \frac{\pi_t^{(k)}(A_t|S_t)}{\mu_t(A_t|S_t)}$. Then, the product of importance sampling ratios from time t to $t' \geq t$ is defined as $\rho_{t:t'}^{\pi^{(k)},\mu} \doteq \prod_{i=t}^{t'} \frac{\pi_i^{(k)}(A_i|S_i)}{\mu_i(A_i|S_i)}$. Among the various ways to use the importance sampling ratios in off-policy learning (Geweke, 1988; Hesterberg, 1995; Koller and Friedman, 2009; Thomas, 2015), we use the per-decision importance sampling estimator (PDIS, Precup et al. (2000a)) In this chapter and leave the study of others for future work. For $\pi^{(k)}$, the PDIS Monte Carlo estimator is defined as $G_k^{\text{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \doteq \sum_{i=t}^{T-1} \rho_{t:i}^{\pi^{(k)},\mu} R_{i+1}$, which is unbiased for any behavior policy μ that covers target policy $\pi^{(k)}$ (Precup et al., 2000a). That is, when $\forall s$, $\forall a$, $\mu_t(a|s) = 0 \implies \pi_t^{(k)}(a|s) = 0$, we have $\forall t$, $\forall s$, $\mathbb{E}[G_k^{\text{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t = s] = v_{\pi^{(k)},t}(s)$. We also leverage the recursive form of the

PDIS estimator:

$$
\begin{aligned}
&G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \\
&= \begin{cases} \rho_t^{\pi^{(k)},\mu}\left(R_{t+1} + G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right)\right) & t \in [T-2], \\ \rho_t^{\pi^{(k)},\mu} R_{t+1} & t = T-1. \end{cases}
\end{aligned} \tag{5.1}
$$

Because the PDIS estimator is unbiased, reducing its variance is sufficient for the improvement of its sample efficiency. We achieve this variance reduction goal for multiple policies by designing a tailored behavior policy.

5.2 Variance Reduction in Statistics

In this section, we propose the mathematical framework for variance reduction using importance sampling ratios. Let A be a discrete random variable with a finite set of possible values $\mathcal{A}$, and assume it follows a probability mass function $\pi^{(k)} : \mathcal{A} \to [0, 1]$, called target policy. Additionally, let $q : \mathcal{A} \to \mathbb{R}$ be a function that maps elements of $\mathcal{A}$ to real numbers. Our objective is to estimate $\mathbb{E}_{A\sim\pi^{(k)}}[q(A)]$ for each $\pi^{(k)}$, where k is an index within a finite set $[K]$. Since In this chapter, data can be generated from multiple distributions, we specify their source clearly. We reserve the superscript with brackets $[\cdot, \cdot]$ to denote the source and the index of samples. For example, $A^{[\pi^{(k)},i]}$ is the ith sample generated by running $\pi^{(k)}$. We use n_k to denote the total number of samples sampled by policy $\pi^{(k)}$. The plain Monte Carlo methods then samples $\left\{A^{[\pi^{(k)},1]}, \ldots, A^{[\pi^{(k)},n_k]}\right\}$ from each $\pi^{(k)}$ and use the empirical average $\frac{1}{n_k}\sum_{i=1}^{n_k} q(A^{[\pi^{(k)},i]})$ as the estimate for each $\mathbb{E}_{A\sim\pi^{(k)}}[q(A)]$.

The importance sampling is introduced as a variance reduction technique in statistics, where the main idea is to sample $\left\{q(A^{[\mu,i]})\right\}_{i=1}^{N}$ following a distribution μ and use $\frac{1}{N}\sum_{i=1}^{N} \rho^{\pi^{(k)},\mu}(A^{[\mu,i]})q(A^{[\mu,i]})$ as the estimate, where $\rho^{\pi^{(k)},\mu}(A) \doteq \frac{\pi^{(k)}(A)}{\mu(A)}$ is the importance sampling ratio. In this statistics section, we propose the optimal behavior policy μ that evaluates all target policies $\pi^{(k)}$ simultaneously by sharing samples. We also define the similarity of policies and prove when target policies satisfy the similarity condition, samples needed to estimate all of them do not scale with the number of policies K. These ideas are later extended into the multi-step reinforcement learning (RL) setting in the following section.

Assuming that $\forall i, \mu$ covers $\pi^{(k)}$, i.e.,

$$
\forall a, \mu(a) = 0 \implies \pi^{(k)}(a) = 0. \tag{5.2}
$$

Then, the importance sampling ratio weighted empirical average is unbiased, i.e., $\forall k$, $\mathbb{E}_{A\sim\pi^{(k)}}[q(A)] = \mathbb{E}_{A\sim\mu}[\rho^{\pi^{(k)},\mu}(A)q(A)]$. If we carefully design the sampling distribution μ, the variance can be reduced. We formulate this problem of searching a variance-reducing sampling distribution for K policies as an optimization problem

$$\min_{\mu\in\Lambda_-} \quad \sum_{k\in[K]} \mathbb{V}_{A\sim\mu}(\rho^{\pi^{(k)},\mu}(A)q(A)),$$

where Λ_- is the classical search space (Rubinstein, 1981; Zhang, 2022; Liu et al., 2025b; Qian et al., 2024) defined as

$$\Lambda_- \doteq \left\{\mu \in \Delta(\mathcal{A}) \mid \forall a, \forall k, \mu(a) = 0 \Rightarrow \pi^{(k)}(a) = 0\right\}.$$

Here, $\Delta(\mathcal{A})$ denotes the set of all probability distributions on the set $\mathcal{A}$. In other words, Λ_- includes all distributions that cover $\left\{\pi^{(k)}\right\}_{k=1}^{K}$. In this work, we enlarge Λ_- to Λ, which is defined as

$$\Lambda \doteq \left\{\mu \in \Delta(\mathcal{A}) \mid \forall a, \forall k, \mu(a) = 0 \Rightarrow \pi^{(k)}(a)q(a) = 0\right\}. \tag{5.3}$$

The space Λ weakens the assumption in (5.2). We prove that any distribution μ in Λ still gives unbiased estimation, though $\Lambda_- \subseteq \Lambda$.

Lemma 4. $\forall \mu \in \Lambda, \forall k,$

$$\mathbb{E}_{A\sim\mu}\left[\rho^{\pi^{(k)},\mu}(A)q(A)\right] = \mathbb{E}_{A\sim\pi^{(k)}}\left[q(A)\right].$$

Its proof is in the appendix. We now consider the variance minimization problem on Λ, i.e.,

$$\min_{\mu\in\Lambda} \quad \sum_{k\in[K]} \mathbb{V}_{A\sim\mu}(\rho^{\pi^{(k)},\mu}(A)q(A)). \tag{5.4}$$

The following lemma gives an optimal solution μ^* to the optimization problem (5.4).

Lemma 5. *Define* $\mu^*(a) \propto \sqrt{\sum_{k\in[K]} \pi^{(k)}(a)^2 q(a)^2}$. *Then* μ^* *is an optimal solution to* (5.4).

Its proof is in the appendix. Here, $\mu(a) \propto f(a)$ with some non-negative $f(a)$ means

$$\mu(a) \doteq f(a)/\textstyle\sum_b f(b).$$

If $f(a) = 0$ for all a, the above "reweighted" distribution is not well defined. We then use the convention to interpret $\mu(a)$ as a uniform distribution, i.e., $\mu(a) = 1/|\mathcal{A}|$. *This convention in using* $\propto$ *is adopted in the rest of the paper for simplicity.*

When estimating $\mathbb{E}_{A\sim\pi^{(k)}}[q(A)]$, $\pi^{(k)}(a)q(a)$ shows how much an action contributes to the expectation and is heavily used (Owen, 2013; Liu and Zhang, 2024). Denote

$$w^{(k)}(a) \doteq \left(\pi^{(k)}(a)q(a)\right)^2, \tag{5.5}$$

$$\bar{w}(a) \doteq \textstyle\sum_{j\in[K]} w^{(j)}(a)/K. \tag{5.6}$$

We use $\eta^{(k)}(a)$ to denote the similarity between $\pi^{(k)}$ and the average $\bar{w}(a)$,

$$\eta^{(k)}(a) \doteq w^{(k)}(a)/\bar{w}(a). \tag{5.7}$$

Naturally, $\eta^{(k)}(a) = 1$ when all policies are the same on a. Define $\underline{\eta} \doteq \min_{k,a} \eta^{(k)}(a)$ and $\overline{\eta} \doteq \max_{k,a} \eta^{(k)}(a)$, we have $\forall k, a$,

$$\underline{\eta} \leq \eta^{(k)}(a) \leq \overline{\eta}. \tag{5.8}$$

In the following theorem, we compare the variance of estimation methods. For off-policy evaluation, our designed μ^* generates n samples. For on-policy evaluation, when evaluating multiple policies, it is common for different policies to generate different numbers of samples. Thus, to achieve a fair and general enough comparison, each target policy $\pi^{(k)}$ generates n_k samples. There is no constraint on n_k, as long as $\sum_{k=1}^K n_k = n$. Using $A^{[\pi^{(k)},i]}$ to denote the ith sample generated following $\left\{\pi^{(k)}\right\}$, we define the empirical average for all $\pi^{(k)}$ as

$$E^{\text{on},\pi^{(k)}} \doteq \frac{\sum_{i=1}^{n_k} q(A^{[\pi^{(k)},i]})}{n_k}. \tag{5.9}$$

Similarly, using $A^{[\mu^*,i]}$ to denote the ith sample generated by μ^*, We define the empirical average for all $\pi^{(k)}$ as

$$E^{\text{off},\pi^{(k)}} \doteq \frac{\sum_{i=1}^{n} \rho^{\pi^{(k)},\mu^*}(A^{[\mu^*,i]})q(A^{[\mu^*,i]})}{n}. \tag{5.10}$$

Then, we characterize sufficient conditions on policy similarity such that with the same total samples, off-policy evaluation with our tailored behavior policy μ^* achieves a lower variance than on-policy Monte Carlo on each $\pi^{(k)}$.

Lemma 6. *$\forall k \in [K]$,*

$$\mathbb{V}_{A\sim\mu^*}\left(E^{off,\pi^{(k)}}\right) \leq \mathbb{V}_{A\sim\pi^{(k)}}\left(E^{on,\pi^{(k)}}\right),$$

if the similarity $\eta(\cdot)$ has $\forall k$,

$$\begin{aligned}&\sqrt{\frac{\overline{\eta}}{\underline{\eta}}}\left(\textstyle\sum_a \pi^{(k)}(a)q(a)\right)^2 - \left(\frac{n}{n_k} - 1\right)\Delta^{(k)}\\ \leq &\textstyle\sum_a \pi^{(k)}(a)q(a)^2,\end{aligned} \tag{5.11}$$

where

$$\Delta^{(k)} \doteq \left[\textstyle\sum_a \pi^{(k)}(a)q(a)^2 - \left(\sum_a \pi^{(k)}(a)q(a)\right)^2\right].$$

Its proof is in the appendix. In Lemma 6, we show under characterized conditions, using only the same total samples n generated by μ^*, the off-policy estimator already achieves a lower variance than on-policy estimator for each target policy $\pi^{(k)}$. Now, we present a stronger lemma by allowing each target policy to also generate n samples, resulting in a total of nK samples, which is K times larger than n. Using the empirical average for on-policy estimator as defined in (5.9), we now have, for all $\pi^{(k)}$,

$$E^{\text{on},\pi^{(k)}} = \sum_{i=1}^{n} q(A^{[\pi^{(k)},i]})/n. \tag{5.12}$$

Then, we simplify the variance of the on-policy estimator for $\pi^{(k)}$ as

$$\begin{aligned}
&\mathbb{V}_{A\sim\pi^{(k)}}(E^{\text{on},\pi^{(k)}}) \\
=&\mathbb{V}_{A\sim\pi^{(k)}}(\tfrac{\sum_{i=1}^{n} q(A^{[\pi^{(k)},i]})}{n}) && \text{(By (5.12))} \\
=&\frac{1}{n}\mathbb{V}_{A\sim\pi^{(k)}}(\textstyle\sum_{i=1}^{n} q(A^{[\pi^{(k)},i]})) \\
=&\mathbb{V}_{A\sim\pi^{(k)}}(q(A)).
\end{aligned}$$

In the last step, we leverage the independence of samples. Similarly, using the definition of empirical average for off-policy estimator as defined in (5.10), we have

$$\mathbb{V}_{A\sim\pi^{(k)}}(E^{\text{off},\pi^{(k)}}) = \mathbb{V}_{A\sim\mu^*}\Big(\rho^{\pi^{(k)},\mu^*}(A)q(A)\Big).$$

Then, we formalize the superiority for the "n-to-Kn" comparison in the following theorem.

Lemma 7. *$\forall k \in [K]$,*

$$\mathbb{V}_{A\sim\mu^*}\Big(\rho^{\pi^{(k)},\mu^*}(A)q(A)\Big) \leq \mathbb{V}_{A\sim\pi^{(k)}}(q(A)),$$

if the similarity $\eta(\cdot)$ has $\forall k$,

$$\sqrt{\tfrac{\overline{\eta}}{\underline{\eta}}}\big(\textstyle\sum_{a\in\mathcal{A}} \pi^{(k)}(a)q(a)\big)^2 \leq \sum_{a\in\mathcal{A}} \pi^{(k)}(a)q(a)^2. \tag{5.13}$$

Its proof is in the appendix. The superiority of using our designed behavior policy μ^* comes from two sources. First, μ^* generates samples that all similar policies can efficiently share. Second, it is designed to generate low-variance and unbiased samples compared with the on-policy evaluation.

5.3 Variance Reduction in Reinforcement Learning

We extend the techniques discussed in the statistics section into multi-step reinforcement learning (RL). In this section, Theorem 6 is the RL version of Lemma 4 for unbiasedness. Theorem 7 is the RL version of Lemma 5 for behavior policy design. Theorem 8 and 9 are the RL version of Lemma 6 and 7, respectively, for variance reduction.

As discussed in the related work section, the major caveat in multi-policy evaluation problems is data sharing. Without efficient data sharing, the total number of samples required for evaluating all policies increases rapidly with the number of target policies. Previous works try to reuse collected data across multiple target policies. However, their method rely on either (1) **restrictive assumptions**, namely, deterministic policies and flexible environment starting at any desired state (Dann et al., 2023), or (2) **impractical knowledge**, namely, complicated covariances (Lai et al., 2020) and state visitation densities at very step (Kallus et al., 2021). Thus, none of the existing methods (Dann et al., 2023; Lai et al., 2020; Kallus et al., 2021; Agarwal et al., 2017) is implementable in the multi-step RL setting.

In this work, we tackle this notorious problem of efficient multi-policy evaluation in RL without any impracticability. We seek to reduce the variance $\sum_{k\in[K]} \mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\mu_{0:T-1}}\right)\right)$ by designing a proper behavior policy μ. Certainly, we need to ensure that the PDIS estimator with this behavior policy is unbiased.

In the off-policy evaluation problem, classic reinforcement learning (Sutton and Barto 2018) requires coverage assumption to ensure unbiased estimation. This means they only consider a set of policies that cover $\left\{\pi^{(k)}\right\}_{k=1}^{K}$, i.e.,

$$\Lambda_{-} \doteq \{\mu \mid \forall k, t, s, a, \mu_t(a|s) = 0 \implies \pi_t^{(k)}(a|s) = 0\}.$$

Similar to (5.3), we enlarge Λ_{-} to

$$\begin{aligned}\Lambda \doteq &\{\mu \mid \forall k, t, s, a, \mu_t(a|s) = 0 \\ &\implies \pi_t^{(k)}(a|s) q_{\pi^{(k)},t}(s,a) = 0\}.\end{aligned}$$

We prove every policy $\mu \in \Lambda$ still achieves unbiased estimation in the following theorem.

Theorem 6 (Unbiasedness)**.** $\forall \mu \in \Lambda$, $\forall k$, $\forall t$, $\forall s$,

$$\mathbb{E}\left[G_k^{PDIS}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t = s\right] = v_{\pi^{(k)},t}(s).$$

Its proof is in the appendix. One immediate consequence of Theorem 6 is that $\forall \mu \in \Lambda, \forall k, \mathbb{E}\left[G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\mu_{0:T-1}}\right)\right] = J(\pi^{(k)})$. In this chapter, we consider a set $\hat{\Lambda}$ such that $\Lambda_- \subseteq \hat{\Lambda} \subseteq \Lambda$. $\hat{\Lambda}$ inherits the unbiasedness property of Λ and is less restrictive than Λ_-, the classical search space of behavior policies. This $\hat{\Lambda}$ will be defined shortly. We now formulate our problem as

$$\min_{\mu \in \hat{\Lambda}} \quad \sum_{k \in [K]} \mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\mu_{0:T-1}}\right)\right). \tag{5.14}$$

By the law of total variance, for any $\mu \in \hat{\Lambda}$, we decompose the variance of the PDIS estimator as

$$\begin{aligned} & \sum_{k \in [K]} \mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\mu_{0:T-1}}\right)\right) \\ =& \sum_{k \in [K]} \mathbb{E}_{S_0}\left[\mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\mu_{0:T-1}}\right) \mid S_0\right)\right] \\ & + \mathbb{V}_{S_0}\left(\mathbb{E}\left[G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\mu_{0:T-1}}\right) \mid S_0\right]\right) \\ =& \sum_{k \in [K]} \mathbb{E}_{S_0}\left[\mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\mu_{0:T-1}}\right) \mid S_0\right)\right] \\ & + \mathbb{V}_{S_0}\left(v_{\pi^{(k)},0}(S_0)\right). \qquad \text{(by Theorem 6)} \end{aligned} \tag{5.15}$$

The second term in (5.15) is a constant given a target policy $\pi^{(k)}$ and is unrelated to the choice of μ. In the first term, the expectation is taken over S_0 that is determined by the initial probability distribution p_0. Consequently, to solve the problem (5.14), it is sufficient to solve for each s,

$$\min_{\mu \in \hat{\Lambda}} \quad \sum_{k \in [K]} \mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\mu_{0:T-1}}\right) \mid S_0 = s\right).$$

Denote the variance of the state value for the next state given the current state-action pair (s, a) as $\nu_{\pi^{(k)},t}(s, a)$. We have $\nu_{\pi^{(k)},t}(s, a) = 0$ for $t = T - 1$ and otherwise

$$\nu_{\pi^{(k)},t}(s, a) \doteq \mathbb{V}_{S_{t+1}}\left(v_{\pi^{(k)},t+1}(S_{t+1}) \mid S_t = s, A_t = a\right). \tag{5.16}$$

To achieve variance reduction compared with on-policy evaluation, we aim to design $\hat{\mu}_t$ as an optimal solution to the following problem

$$\min_{\mu_t \in \hat{\Lambda}} \quad \sum_k \mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{t:T-1}^{\left\{\mu_t, \pi_{t+1}^{(k)}:\pi_{T-1}^{(k)}\right\}}\right) \mid S_t = s\right), \tag{5.17}$$

The high-level intuition is that we aim to find the optimal behavior policy μ_t for the current step, assuming that in the future we perform the on-policy evaluation. To define optimality, we first specify the set of policies we are concerned about. To this end, we define that $\forall k$, $\hat{q}_{\pi^{(k)},t}(s, a) \doteq q_{\pi^{(k)},t}(s, a)^2$ for $t = T - 1$ and otherwise

$$\begin{aligned} & \hat{q}_{\pi^{(k)},t}(s, a) \doteq q_{\pi^{(k)},t}(s, a)^2 + \nu_{\pi^{(k)},t}(s, a) \\ & + \sum_{s'} p(s'|s, a) \mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{t+1:T-1}^{\pi_{t+1:T-1}^{(k)}}\right) \mid S_{t+1} = s'\right). \end{aligned} \tag{5.18}$$

Notably, $\hat{q}_{\pi^{(k)},t}(s,a)$ is always *non-negative* since all the summands are non-negative. Accordingly, we define $\hat{\Lambda} \doteq \{\mu \mid \forall k,t,s,a, \mu_t(a|s)=0 \Rightarrow \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)=0\}$. From (5.18), we observe for any k, t, s, a, $\hat{q}_{\pi^{(k)},t}(s,a) \geq q_{\pi^{(k)},t}(s,a) \geq 0$. As a result, if $\mu_t \in \hat{\Lambda}$, we have $\mu_t(a|s)=0 \Rightarrow \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)=0 \Rightarrow \pi_t^{(k)}(a|s)q_{\pi^{(k)},t}(s,a)=0$. Thus, $\hat{\Lambda} \subseteq \Lambda$. To summarize, we have $\Lambda_- \subseteq \hat{\Lambda} \subseteq \Lambda$. $\hat{\Lambda}$ inherits the unbiased property of Λ (Theorem 6) and is larger than the classic space Λ_- considered in previous works (Precup et al., 2000a; Maei, 2011; Sutton et al., 2016; Sutton and Barto, 2018).

Now, we define the optimal behavior policy as

$$\hat{\mu}_t(a|s) \propto \sqrt{\textstyle\sum_{k=1}^{K} \pi_t^{(k)}(a|s)^2 \hat{q}_{\pi^{(k)},t}(s,a)}. \tag{5.19}$$

$\hat{q}$ defined in (5.18) is different from q, and is always non-negative. We confirm the optimality of $\hat{\mu}_t$ in the following theorem.

Theorem 7 (Behavior Policy Design)**.** *For any k, t and s, the behavior policy $\hat{\mu}_t(a|s)$ defined in (5.19) is an optimal solution to the following problem*

$$\min_{\mu_t \in \hat{\Lambda}} \quad \sum_k \mathbb{V}\left(G_k^{PDIS}\left(\tau_{t:T-1}^{\left\{\mu_t, \pi_{t+1}^{(k)}:\pi_{T-1}^{(k)}\right\}} \right) \mid S_t = s \right).$$

Its proof is in the appendix. Next, we formalize the similarity between target policies. Similar to (5.5), (5.6) in the statistics setting, $\forall k$, $\forall t$, $\forall s$, we denote

$$w_t^{(k)}(s,a) \doteq \pi_t^{(k)}(a|s)^2 \hat{q}_{\pi^{(k)},t}(s,a), \tag{5.20}$$

$$\bar{w}_t(s,a) \doteq \left(\textstyle\sum_{j\in[K]} w_t^{(j)}(s,a)\right)/K. \tag{5.21}$$

Then, adopting the notation from (5.7) and (5.8), we denote the similarity between $\pi_t^{(k)}$ and the average $\bar{w}_t$ as

$$\eta_t^{(k)}(s,a) \doteq w_t^{(k)}(s,a)/\bar{w}_t(s,a). \tag{5.22}$$

When policies are the same, $\forall k,t,s$, $\eta_t^{(k)}(s,a)=1$. Define $\underline{\eta}_t \doteq \min_{k,s,a} \eta_t^{(k)}(s,a)$ and $\overline{\eta} \doteq \max_{k,a} \eta_t^{(k)}(s,a)$, we have $\forall t,k,s,a$,

$$\underline{\eta}_t \leq \eta_t^{(k)}(s,a) \leq \overline{\eta}_t. \tag{5.23}$$

Next, to extend the variance reduction property from statistics (Lemma 6) into reinforcement learning, we also allow each target policy to generate n_k samples. With a similar notation, we have the empirical average for all $\pi^{(k)}$ as

$$E_{t:T-1}^{\text{on},\pi^{(k)}} \doteq \frac{\sum_{i=1}^{n_k} G_k^{\text{PDIS}}\left(\tau_{t:T-1}^{[\pi_{t:T-1}^{(k)},i]}\right)}{n_k}, \tag{5.24}$$

where $\tau^{[\pi^{(k)},i]}$ is the ith trajectory obtained by running $\pi^{(k)}$. To achieve a fair comparison, when doing off-policy estimation by following $\hat{\mu}$, we generate $n = \sum_{k=1}^{K} n_k$ samples. Likewise, define

$$E_{t:T-1}^{\text{off},\pi^{(k)}} \doteq \frac{\sum_{i=1}^{n} G_k^{\text{PDIS}}\left(\tau_{t:T-1}^{[\hat{\mu}_{t:T-1},i]}\right)}{n}. \tag{5.25}$$

We have the following theorem.

Theorem 8 (Variance Reduction with Same Sample Sizes)**.** *$\forall k$, $\forall t$, $\forall s$,*

$$\mathbb{V}\left(E_{t:T-1}^{off,\pi^{(k)}} \mid S_t = s\right) \leq \mathbb{V}\left(E_{t:T-1}^{on,\pi^{(k)}} \mid S_t = s\right).$$

if the similarity η has $\forall k, \forall t, \forall s$,

$$\begin{aligned} &\sqrt{\tfrac{\overline{\eta_t}}{\underline{\eta_t}}}\left(\textstyle\sum_a \pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(a|s)}\right)^2 - \left(1 - \tfrac{n_k}{n}\right)\Delta_t^{(k)}(s) \\ \leq &\textstyle\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a), \end{aligned} \tag{5.26}$$

where

$$\begin{aligned} \Delta_t^{(k)}(s) \doteq &\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2 \nu_{\pi^{(k)},t}(S_t,A_t) \mid S_t = s\right] \\ &+\mathbb{V}_{A_t\sim\hat{\mu}_t}\left(\rho^{\pi^{(k)},\hat{\mu}} q_{\pi^{(k)},t}(S_t,A_t) \mid S_t = s\right). \end{aligned}$$

Its proof is in the appendix. We then compare the datasets when the behavior policy $\hat{\mu}$ and each target policy $\pi^{(k)}$ both generate n samples, resulting in a "n-to-nK" comparison, similar to Lemma 7.

Theorem 9 (Variance Reduction)**.** *$\forall k$, $\forall t$, $\forall s$,*

$$\begin{aligned} &\mathbb{V}\left(G_k^{PDIS}\left(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}\right) \mid S_t = s\right) \\ \leq &\mathbb{V}\left(G_k^{PDIS}\left(\tau_{t:T-1}^{\pi_{t:T-1}^{(k)}}\right) \mid S_t = s\right), \end{aligned}$$

if the similarity η has $\forall k, \forall t, \forall s$,

$$\begin{aligned} &\sqrt{\tfrac{\overline{\eta_t}}{\eta_t}}\left(\textstyle\sum_a \pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(s,a)}\right)^2 \\ \leq &\textstyle\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a). \end{aligned} \tag{5.27}$$

Its proof is in the appendix. This theorem implies that in the multi-step RL setting, running our tailored behavior policy $\hat{\mu}$ also ensures that the number of required samples does not scale with the number of target policies under similarity

Algorithm 2: Multi-Policy Evaluation (MPE) algorithm

Input: K target policies $\pi^{(k)}$,
an offline dataset $\mathcal{D} = \{(t_i, s_i, a_i, r_i, s_i')\}_{i=1}^m$
Output: a behavior policy $\hat{\mu}$
Approximate $q_{\pi^{(k)},t}$ from $\mathcal{D}$ using any offline RL method (e.g. Fitted Q-Evaluation)
Compute $\hat{r}_{\pi^{(k)},i}$ for data pairs in $\mathcal{D}$ by (4.19)
Construct $\mathcal{D}^{(k)} \doteq \left\{(t_i, s_i, a_i, \hat{r}_{\pi^{(k)},i}, s_i')\right\}_{i=1}^m$
Approximate $\hat{q}_{\pi^{(k)},t}$ from $\mathcal{D}^{(k)}$ by (C.22) using any offline method (e.g. Fitted Q-Evaluation)
Return: $\hat{\mu}_t(a|s) \propto \sqrt{\sum_{k=1}^K \pi_t^{(k)}(a|s)^2 \hat{q}_{\pi^{(k)},t}(s,a)}$

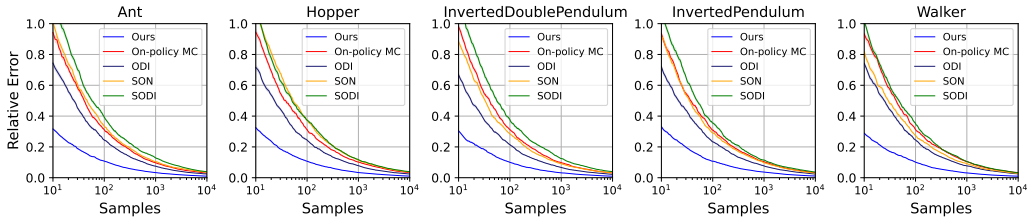

Figure 5.2: Results on MuJoCo. Each curve is averaged over 900 runs (30 groups of target policies, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves because they are too small.

conditions. The reduced variance of our method depends on the similarity between target policies, which can be easily checked through learning $\hat{q}$ with offline data. Thus, if RL practitioners are not confident in the similarity between target policies, they can verify it before actual deployment without consuming any online data.

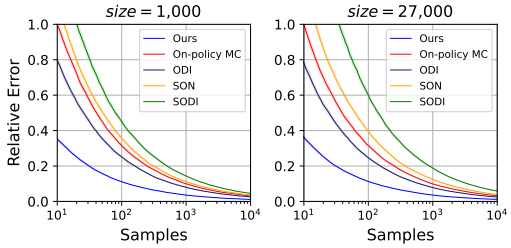

Figure 5.1: Results on Gridworld. Each curve is averaged over 900 runs (30 groups of policies, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves because they are too small.

5.4 Empirical Results

We evaluate $K = 10$ target policies simultaneously by executing the tailored behavior policy $\hat{\mu}$ with n total samples. We name our method multiple policy evaluation (MPE) estimator. We present our empirical comparisons with the following baselines: **(1)** The canonical on-policy Monte Carlo estimator with n_k samples for each target policy $\pi^{(k)}$, summing to a total of $n = \sum_{k=1}^{K} n_k$ samples. **(2)** The offline data informed estimator (ODI, Liu and Zhang (2024)) that runs each behavior policy (designed for each target policy $\pi^{(k)}$) for n_k samples, summing to a total of $n = \sum_{k=1}^{K} n_k$ samples. **(3)** The shared-sample on-policy Monte Carlo estimator (SON), where we evaluate each target policy with shared data collected by canonical on-policy Monte Carlo estimators of all K policies, resulting in $n = \sum_{k=1}^{K} n_k$ samples used to evaluate every target policy. **(4)** The shared-sample ODI estimator (SODI), where we evaluate each target policy with shared data collected by ODI estimators of all K policies. Since each single behavior policy from the ODI estimator collects n_k samples, each target policy in SODI leverages $n = \sum_{k=1}^{K} n_k$ samples.

As a demonstration of concept, we set $K = 10$ and $n_k = \frac{n}{K}$ for each of the 10 target policies. Target policies are drawn from the training process of proximal policy optimization (PPO) algorithm (Schulman et al., 2017). We learn our behavior policy $\hat{\mu}$ using Algorithm 2. Hyperparameters are the same across all MuJoCo and Gridworld experiments. Experimental details are in the appendix.

Gridworld: We use Gridworld with $m^3 = 1,000$ and $m^3 = 27,000$ states, where each Gridworld has a width m and height m with a time horizon $T = m$.

Env Size	Ours	On-policy MC	ODI	SON	SODI
1,000	**0.125**	1.000	0.637	1.289	2.073
27,000	**0.129**	1.000	0.601	1.561	3.532

Table 5.1: Relative variance of estimators on Gridworld. The relative variance is defined as the variance of each estimator divided by the variance of the on-policy Monte Carlo estimator. Numbers are averaged over 900 independent runs (30 groups of target policies, each having 30 independent runs).

Env Size	Ours	On-policy MC	ODI	SON	SODI
1,000	**126**	1000	632	1264	2046
27,000	**131**	1000	629	1568	3501

Table 5.2: Episodes needed to achieve the same of estimation accuracy that on-policy Monte Carlo achieves with 1000 episodes. Numbers are averaged over 900 independent runs (30 groups of target policies, each having 30 independent runs) and their standard errors are shown in Figure 5.1.

Figure 5.1 shows our method outperforms all baselines by a large margin. The *relative error* is defined as the estimation error divided by the estimation error of the on-policy MC at the beginning of x-axis. The *samples* on the x-axis represents the total online episodes for multi-policy evaluation. The blue line in the graph is below other lines, indicating that our method requires fewer samples to achieve the same accuracy. To quantify the variance reduction, Table 5.1 shows our method reduces variance to about 12.5% compared with the on-policy Monte Carlo estimator. Table 5.2 shows that to achieve the same estimation error that the on-policy Monte Carlo estimator achieves with 1000 samples, our estimator only needs about 130 samples saving about 87% of online interactions, achieving state-of-the-art performance.

MuJoCo: Next, we conduct experiments in MuJoCo robot simulation tasks (Todorov et al., 2012). MuJoCo is a physics engine containing various stochastic environments, where the goal is to control a robot to achieve different behaviors such as walking, jumping, and balancing. Figure 5.2 shows our method is consistently better than all baselines. The tables in the appendix show similar patterns as in the Gridworld experiment. In particular, our estimator reduces the variance to about 10% compared with the on-policy Monte Carlo estimator and saves about 90% of online interactions.

An interesting observation to demonstrate the discrepancy among target policies is that SODI and SON generally perform worse than On-policy MC and ODI. This result suggests that when target policies lack sufficient similarity, reusing data without a carefully designed joint behavior policy leads to high-variance estimation. Additionally, while ODI outperforms On-policy MC, SODI performs worse than SON. This may be because each behavior policy in SODI is specially tailored for its own target policy, making it vulnerable to target policy change. *These observations confirm the notorious difficulty of data sharing across multiple policies, highlighting the need for a tailored and shared behavior policy to efficiently facilitate data sharing.*

5.5 Discussion

In this chapter, we introduce a novel approach for multi-policy evaluation by designing a tailored behavior policy that efficiently and unbiasedly evaluates multiple target policies.

Theoretically, our method eliminates the need for restrictive assumptions or infeasible knowledge required by previous methods. Our method achieves lower variance compared to on-policy evaluation for each target policy under similarity conditions (Theorem 8, Theorem 9) and ensures the number of required samples does not scale with the number of target policies when similarity conditions hold.

Empirically, our method outperforms previously best-performing methods, achieving state-of-the-art performance across various environments. One promising future direction is to extend our variance reduction method to policy improvement and achieve efficient policy learning.

Chapter 6

Doubly Optimal Policy Evaluation

This chapter is based on my paper Liu et al. (2025a) published at ICLR 2025.

Policy evaluation estimates the performance of a policy by (1) collecting data from the environment and (2) processing raw data into a meaningful estimate. Due to the sequential nature of reinforcement learning, any improper data-collecting policy or data-processing method substantially deteriorates the variance of evaluation results over long time steps. Thus, policy evaluation often suffers from large variance and requires massive data to achieve the desired accuracy. In this work, we design an optimal combination of data-collecting policy and data-processing baseline. Theoretically, we prove our doubly optimal policy evaluation method is unbiased and guaranteed to have lower variance than previously best-performing methods. Empirically, compared with previous works, we show our method reduces variance substantially and

6.1 Preliminaries

In this chapter, we study the task of off-policy evaluation in the context of a finite Markov Decision Process (MDP), as defined in Section 2.1. In off-policy evaluation, a notorious curse is that the importance sampling ratios can be extremely large, resulting in infinite variance (Sutton and Barto, 2018). Even with the PDIS method, this fundamental issue still remains if the behavior policy significantly differs from the target policy, particularly when the behavior policy assigns very low probabilities to actions favored by the target policy. Moreover, such degeneration of important sampling ratios typically grows with the dimensions of state and action spaces as well as the time horizon (Levine et al., 2020). One way to control for the violation in important sampling ratios is to subtract a baseline from samples (Williams, 1992; Greensmith et al., 2004; Jiang and Li, 2016; Thomas and Brunskill, 2017). Using b to

denote an arbitrary baseline function, the PDIS estimator with baseline is defined as

$$G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) = \begin{cases} \rho_t \left(R_{t+1} + G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) - b_t(S_t, A_t)\right) + \bar{b}_t(S_t) & t \in [T-2], \\ \rho_t(R_{t+1} - b_t(S_t, A_t)) + \bar{b}_t(S_t) & t = T-1, \end{cases} \tag{6.1}$$

where

$$\bar{b}_t(S_t) \doteq \mathbb{E}_{A_t \sim \pi}[b_t(S_t, A_t)]. \tag{6.2}$$

The variance of (6.1) highly depends on the importance sampling ratio $\rho_t = \frac{\pi_t(A_t|S_t)}{\mu_t(A_t|S_t)}$ and the choice of baseline function b.

6.2 Variance Reduction in Reinforcement Learning

We seek to reduce the variance $\mathbb{V}(G^b(\tau_{0:T-1}^{\mu_{0:T-1}}))$ by designing an optimal behavior policy and an optimal baseline function at the same time. We solve the bi-level optimization problem

$$\begin{aligned} \min_b \min_\mu \quad & \mathbb{V}(G^b(\tau_{0:T-1}^{\mu_{0:T-1}})) \\ \text{s.t.} \quad & \mathbb{E}\left[G^b(\tau_{0:T-1}^{\mu_{0:T-1}})\right] = J(\pi), \end{aligned} \tag{6.3}$$

where the optimal behavior policy μ^* and the optimal baseline function b^* are carefully tailored to each other to guarantee both unbiasedness and substantial variance reduction.

Our paper proceeds as follows. In Section 6.2, we solve this bi-level optimization problem in closed-form. In Section 6.3, we mathematically quantify the superiority in variance reduction of our designed optimal behavior policy and baseline function, in comparison with cutting-edge methods (Jiang and Li, 2016; Liu and Zhang, 2024). In Section 6.5, we empirically show that such doubly optimal design reduces the variance substantially compared with the on-policy Monte Carlo estimator and previously best methods (Jiang and Li, 2016; Liu and Zhang, 2024) in a broad set of environments.

To ensure that the off-policy estimator $G^b(\tau_{0:T-1}^{\mu_{0:T-1}})$ is unbiased, the classic reinforcement learning wisdom (Precup et al., 2000a; Maei, 2011; Sutton et al., 2016; Zhang, 2022) requires that the behavior policy μ covers the target policy π. This means that they constraint μ to be in a set

$$\begin{aligned} \Lambda^- \doteq & \{\mu \mid \forall t, s, a, \pi_t(a|s) \neq 0 \implies \mu_t(a|s) \neq 0\} \\ = & \{\mu \mid \forall t, s, a, \mu_t(a|s) = 0 \implies \pi_t(a|s) = 0\}, \end{aligned}$$

which contains all policies that satisfy the policy coverage constraint in off-policy learning (Sutton and Barto 2018). By specifying the policy coverage constraint, the optimization problem (6.3) is reformulated as

$$\begin{aligned} \min_{b} \min_{\mu} \quad & \mathbb{V}(G^b(\tau_{0:T-1}^{\mu_{0:T-1}})) \\ \text{s.t.} \quad & \mu \in \Lambda^-. \end{aligned} \tag{6.4}$$

In this chapter, compared with the classic reinforcement learning literature, we enlarge the search space of μ from this set Λ^- to a set Λ. To achieve a superior and reliable optimization solution, we require Λ to have two properties.

1. (Broadness) Λ must be broad enough such that it includes all policies satisfying the classic policy coverage constraint (Precup et al., 2000a; Sutton and Barto, 2018). Formally,

$$\Lambda^- \subseteq \Lambda. \tag{6.5}$$

2. (Unbiasedness) Every behavior policy in Λ must be well-behaved such that the data collected by it can be used by the off-policy estimator to achieve unbiased estimation for all state s and time step t. Formally, $\forall \mu \in \Lambda$,

$$\forall t, \forall s, \quad \mathbb{E}\left[G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right] = v_{\pi,t}(s). \tag{6.6}$$

The space Λ that satisfies those two properties will be defined shortly. We now reformulate our bi-level optimization problem as

$$\begin{aligned} \min_{b} \min_{\mu} \quad & \mathbb{V}(G^b(\tau_{0:T-1}^{\mu_{0:T-1}})) \\ \text{s.t.} \quad & \mu \in \Lambda. \end{aligned} \tag{6.7}$$

Compared with the classic approach (6.4), our bi-level optimization problem (6.7) searches for μ in a broader space Λ such that $\Lambda^- \subseteq \Lambda$. Thus, the optimal solution of our optimization problem must be *superior* to the optimal solution of the optimization problem with the classic policy coverage constraint. To solve our bi-level optimization problem (6.7), we first give a closed-form solution of the inner optimization problem

$$\begin{aligned} \min_{\mu} \quad & \mathbb{V}(G^b(\tau_{0:T-1}^{\mu_{0:T-1}})) \\ \text{s.t.} \quad & \mu \in \Lambda \end{aligned} \tag{6.8}$$

for any baseline function b. Notably, this baseline function b does not need to be any kind of oracle. We design the optimal solution of (6.8) for this baseline function b

without requiring any property on b. Now, we decompose the variance of our off-policy estimator $G^b(\tau_{0:T-1}^{\mu_{0:T-1}})$. By the law of total variance, $\forall b, \forall \mu \in \Lambda$,

$$\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{0:T-1}^{\mu_{0:T-1}})\right) \\
=&\mathbb{E}_{S_0}\left[\mathbb{V}\left(G^b(\tau_{0:T-1}^{\mu_{0:T-1}}) \mid S_0\right)\right] + \mathbb{V}_{S_0}\left(\mathbb{E}\left[G^b(\tau_{0:T-1}^{\mu_{0:T-1}}) \mid S_0\right]\right) \\
=&\mathbb{E}_{S_0}\left[\mathbb{V}\left(G^b(\tau_{0:T-1}^{\mu_{0:T-1}}) \mid S_0\right)\right] + \mathbb{V}_{S_0}\left(v_{\pi,0}(S_0)\right). \qquad \text{(by (6.6))} \quad (6.9)
\end{aligned}$$

The second term in (6.9) is a constant given a target policy π and is unrelated to the choice of μ. In the first term, the expectation is taken over S_0 that is determined by the initial probability distribution p_0. Consequently, given any baseline function b, to solve the problem (6.8), it is sufficient to solve

$$\begin{aligned}
\min_{\mu} \quad & \mathbb{V}(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s) \\
\text{s.t.} \quad & \mu \in \Lambda
\end{aligned} \tag{6.10}$$

for all s and t. If we can find one optimal behavior policy μ^* that simultaneously solves the optimization problem (6.10) on all states and time steps, μ^* is also the optimal solution for the optimization problem (6.8). Denote the variance of the state value function for the next state given the current state-action pair as $\nu_{\pi,t}(s,a)$. Recall the notation $[T-2]$ is a shorthand for the set $\{0,1,\ldots,T-2\}$. We have $\nu_{\pi,t}(s,a) \doteq 0$ for $t = T-1$, and $\forall t \in [T-2]$,

$$\nu_{\pi,t}(s,a) \doteq \mathbb{V}_{S_{t+1}}\left(v_{\pi,t+1}(S_{t+1}) \mid S_t = s, A_t = a\right). \tag{6.11}$$

Given any baseline function b, we construct a behavior policy μ^* as

$$\mu_t^*(a|s) \propto \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)} \tag{6.12}$$

where $u_{\pi,t}(s,a) \doteq [q_{\pi,t}(s,a) - b_t(s,a)]^2$ for $t = T-1$, and $\forall t \in [T-2]$,

$$u_{\pi,t}(s,a) \doteq \left(q_{\pi,t}(s,a) - b_t(s,a)\right)^2 + \nu_{\pi,t}(s,a) + \textstyle\sum_{s'} p(s'|s,a)\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^*}) \mid S_{t+1} = s'\right). \tag{6.13}$$

Notably, $u_{\pi,t}$ and μ_t^* are defined backwards and alternatively, i.e., they are defined in the order of $u_{\pi,T-1}, \mu_{T-1}^*, u_{\pi,T-2}, \mu_{T-2}^*, \ldots, u_{\pi,0}, \mu_0^*$. We now break down each term in $u_{\pi,t}(s,a)$.

1. $\left(q_{\pi,t}(s,a) - b_t(s,a)\right)^2$ is the squared difference between the state-value function $q_{\pi,t}$ and the baseline function b. This term is always non-negative because of the square operation. Its magnitude is mainly controlled by the baseline function b.

2. $\nu_{\pi,t}(s,a)$ defined in (6.11) is the variance of the value for the next state. This term is always non-negative by the definition of variance. Its magnitude is mainly controlled by the stochasticity of the environment (i.e. transition function p).

3. $\sum_{s'} p(s'|s,a)\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}=s'\right)$ is the expected future variance given the current state s and action a. This term is always non-negative by the definition of variance. Its magnitude is jointly controlled by the choice of behavior policy μ^*, the baseline function b, and the transition function p.

$u_{\pi,t}(s,a)$ is non-negative because it is the sum of three non-negative terms. Therefore, $\sqrt{u_{\pi,t}(s,a)}$ is always well-defined. In (6.12), $\mu_t^*(a|s) \propto \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)}$ means $\mu_t^*(a|s) \doteq \frac{\pi_t(a|s)\sqrt{u_{\pi,t}(s,a)}}{\sum_b \pi_t(b|s)\sqrt{u_{\pi,t}(s,b)}}$. If $\forall a, \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)} = 0$, the denominator is zero. In this case, we use the convention to interpret $\mu_t^*(a|s)$ as a uniform distribution, i.e., $\forall a, \mu_t^*(a|s) = 1/|\mathcal{A}|$. We adopt this convention for $\propto$ in the rest of the paper to simplify the presentation. Now, we define the enlarged space Λ as

$$\Lambda \doteq \{\mu \mid \forall t,s,a, \mu_t(a|s)=0 \implies \pi_t(a|s)u_{\pi,t}(s,a)=0\}. \tag{6.14}$$

We prove that this policy space Λ defined above satisfies the broadness (6.5) and the unbiasedness (6.6) by the following lemmas.

Lemma 8 (Broadness)**.** *$\forall b$, $\Lambda^- \subseteq \Lambda$.*

Its proof is in Appendix B.1.1.

Lemma 9 (Unbiasedness)**.** *$\forall b, \forall \mu \in \Lambda$, $\forall t, \forall s$, $\mathbb{E}\left[G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right] = v_{\pi,t}(s)$.*

Its proof is in Appendix B.1.2. After confirming the broadness and unbiasedness of the space Λ, we now prove that the behavior policy μ^* is the optimal solution for the inner optimization problem.

Theorem 10. *For a baseline function b, the behavior policy μ^* defined in (6.12) is an optimal solution to the optimization problems $\forall t, s$,*

$$\begin{aligned} \min_{\mu} \quad & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\ s.t. \quad & \mu \in \Lambda. \end{aligned}$$

Its proof is in Appendix B.1.3. Theorem 10 proves that $\forall b$, the behavior policy μ^* (6.12) is the closed-form optimal solution for all t and s. With Theorem 10, for any t and s, we reduce the bi-level optimization problem

$$\begin{aligned} \min_{b} \min_{\mu} \quad & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\ \text{s.t.} \quad & \mu \in \Lambda \end{aligned}$$

to a single-level unconstrained optimization problem

$$\min_{b} \quad \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right).$$

In this unconstrained optimization problem, we design a function b that influences both the data processing estimator G^b (6.1) and the optimal behavior policy μ^* (6.12). Notably, the optimal behavior policy μ^* depends on the baseline b because it is tailored to a baseline function b in Theorem 10. Unless otherwise noted, we omit explicitly writing this dependency in the notation of μ^* for simplicity. We show that although both G^b and μ^* depend on b, through the mathematical proof in the appendix, the optimal baseline function b^* has a concise format. Define $\forall t, s, a$,

$$b_t^*(s,a) \doteq q_{\pi,t}(s,a). \tag{6.15}$$

Theorem 11. *b^* is the optimal solution to the optimization problems $\forall t, s$,*

$$\min_{b} \quad \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right). \tag{6.16}$$

Its proof is in Appendix B.1.4. By solving each level of the optimization problem, we show (μ^*, b^*) is the optimal solution for the bi-level optimization problem by utilizing Theorem 10 and Theorem 11.

Theorem 12. *(μ^*, b^*) is the optimal solution to the bi-level optimization problems $\forall t$, s,*

$$\begin{aligned} \min_{b} \min_{\mu} \quad & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\ s.t. \quad & \mu \in \Lambda. \end{aligned}$$

Proof. $\forall b, \forall \mu \in \Lambda$, we have $\forall t, \forall s$

$$\begin{aligned} & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\ \geq & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) && \text{(Theorem 10)} \\ \geq & \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right). && \text{(Theorem 11)} \end{aligned}$$

Thus, (μ^*, b^*) achieves the minimum value of $\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right)$ for all t and s. □

6.3 Variance Comparison

Theorem 12 shows (μ^*, b^*) is the optimal behavior policy and baseline function. This means (μ^*, b^*) is superior to any other choice of (μ, b). In this section, we further quantify its superiority. We quantify the variance reduction in reinforcement learning. We show that the variance reduction compounds over each step, bringing substantial advantages. Specifically, we provide a theoretical comparison of our method—the doubly optimal estimator—with the following baselines: (1) the on-policy Monte Carlo estimator, (2) the offline data informed estimator (Liu and Zhang, 2024), and (3) the doubly robust estimator (Jiang and Li, 2016). We use $u_t^{b^*}$ to denote u_t (6.13) using b^* as the baseline function. First, we compare our off-policy estimator with the on-policy Monte Carlo estimator (ON).

Theorem 13. $\forall t, s,$

$$\begin{aligned}
&\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) \\
=&\underbrace{\mathbb{V}_{A_t\sim\pi_t}\left(\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t = s\right)}_{(4.1)} + \underbrace{\mathbb{V}_{A_t\sim\pi_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t = s\right)}_{(4.2)} + \underbrace{\delta_t^{ON,\ ours}(s)}_{(4.3)},
\end{aligned}$$

where $\delta_t^{ON,\ ours}(s) \doteq 0$ *for* $t = T-1$ *and* $\forall t \in [T-2]$, $\delta_t^{ON,\ ours}(s) \doteq$

$$\mathbb{E}_{A_t\sim\pi_t, S_{t+1}}\left[\mathbb{V}\left(G^{PDIS}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t = s\right].$$

Moreover, we prove $\forall t, s, \delta_t^{ON,\ ours}(s) \geq 0$ *meaning the variance reduction in future steps is compounded into the current step.*

Its proof is in Appendix B.1.5. In Theorem 13, we show that the variance reduction of our method includes three sources. First, a part of the future variance (4.1) is eliminated by choosing an optimal behavior policy μ^*. Second, the variance of the q function (4.2) is eliminated by the optimal baseline function b^*. Third, the variance reduction in the future step (4.3) is compounded into the current step.

Next, the following theorem quantifies the variance reduction of our method compared with the offline data informed (ODI) method in Liu and Zhang (2024). Because the behavior policy μ^* is tailored for the baseline function b, we use $\mu^{*,b}$ to denote μ^* with a baseline function b and $\mu^{*,\text{PDIS}}$ to denote μ^* with no baseline function (i.e., the plain PDIS estimator considered in offline data informed (ODI) method (Liu and Zhang, 2024)).

Theorem 14. $\forall t, s,$

$$\begin{aligned}
&\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,PDIS}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b^*}}) \mid S_t = s\right) \\
\geq &\underbrace{\mathbb{V}_{A_t \sim \mu_t^{*,PDIS}}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right)}_{(5.1)} + \underbrace{\delta_t^{ODI,\ ours}(s)}_{(5.2)},
\end{aligned}$$

where $\delta_t^{ODI,\ ours}(s) \doteq 0$ *for* $t = T-1$ *and* $\forall t \in [T-2]$, $\delta_t^{ODI,\ ours}(s) \doteq$

$$\mathbb{E}_{A_t \sim \mu_t^{*,PDIS}, S_{t+1}}\left[\rho_t^2\left[\mathbb{V}\left(G^{PDIS}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,PDIS}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{PDIS}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right)\right] \mid S_t\right].$$

Moreover, we prove $\forall t, s, \delta_t^{ODI,\ ours}(s) \geq 0$ *meaning the variance reduction in future steps is compounded into the current step.*

Its proof is in Appendix B.1.6. The variance reduction of our estimator includes two sources. First, the variance of the q function (5.1) is eliminated. Second, the variance reduction in the future step (5.2) is compounded to the current step.

We also quantify the variance reduction of our estimator with the doubly robust (DR) estimator defined in Jiang and Li (2016). Since Jiang and Li (2016) does not specify any candidate behavior policy, we leverage the conventional wisdom, supposing they use the canonical target policy π as the data-collecting policy.

Theorem 15. $\forall t, s,$

$$\begin{aligned}
&\mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^*}) \mid S_t = s\right) \\
= &\underbrace{\mathbb{V}_{A_t \sim \pi_t}\left(\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t = s\right)}_{(6.1)} + \underbrace{\delta_t^{DR,\ ours}(s)}_{(6.2)},
\end{aligned}$$

where $\delta_t^{DR,\ ours}(s) \doteq 0$ *for* $t = T-1$ *and* $\forall t \in [T-2]$, $\delta^{DR,\ ours_t}(s) \doteq$

$$\mathbb{E}_{A_t \sim \pi_t, S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^*}) \mid S_{t+1}\right) \mid S_t\right].$$

Moreover, we prove $\forall t, s, \delta_t^{DR,\ ours}(s) \geq 0$ *meaning the variance reduction in future steps is compounded into the current step.*

Its proof is in Appendix B.1.7. Similarly, there are two sources of the variance reduction for our method. First, with an optimal behavior policy μ^*, we eliminate a part of the future variance (6.1). Second, the variance reduction in the future steps (6.2) is compounded to the current step.

Algorithm 3: Doubly Optimal (DOpt) Policy Evaluation

Input: a target policy π,
an offline dataset $\mathcal{D} = \{(t_i, s_i, a_i, r_i, s_i')\}_{i=1}^m$
Output: a behavior policy μ^*,
a baseline function b^*
Approximate $q_{\pi,t}$ from $\mathcal{D}$ using offline RL methods (e.g. Fitted Q-Evaluation)
Construct $\nu_{\pi,t}$ from $\mathcal{D}$ by (B.22)
Construct $\mathcal{D}_\nu \doteq \{(t_i, s_i, a_i, \nu_i, s_i')\}_{i=1}^m$
Approximate $u_{\pi,t}$ from $\mathcal{D}_\nu$ by Lemma 10
Return: $\mu_t^*(a|s) \propto \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)}$, $b_t^*(s,a) = q_{\pi,t}(s,a)$

6.4 Learning Closed-Form Behavior Policies

In this section, we present an efficient Algorithm 3 to learn our doubly optimal method including the optimal behavior policy μ^* and the optimal baseline function b^*. Specifically, we learn (μ^*, b^*) from offline data pairs. By definition (6.15), we can apply any off-the-shelf offline policy evaluation methods to learn $b_t^*(s,a) \doteq q_{\pi,t}(s,a)$ (e.g., Fitted Q-Evaluation (Le et al., 2019)). By (6.12), $\mu_t^*(a|s) \propto \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)}$, where u is defined in (6.13) as

$$u_{\pi,t}(s,a) \doteq (q_{\pi,t}(s,a) - b_t(s,a))^2 + \nu_{\pi,t}(s,a) + \textstyle\sum_{s'} p(s'|s,a)\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1} = s'\right).$$

Learning u from this perspective is very inefficient because it requires the approximation of the complicated variance term $\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1} = s'\right)$ regarding future trajectories. To solve this problem, we propose the following recursive form of u.

Lemma 10 (Recursive form of u)**.** *With $b = b^*$, when $t = T-1$, $\forall s, a$, $u_{\pi,t}(s,a) = 0$, when $t \in [T-2]$, $\forall s, a$,*

$$u_{\pi,t}(s,a) = \nu_{\pi,t}(s,a) + \textstyle\sum_{s',a'} \rho_{t+1} p(s'|s,a)\pi_{t+1}(a'|s')u_{\pi,t+1}(s',a').$$

Its proof is in Appendix B.1.8. This lemma allows us to learn u recursively without approximating the complicated trajectory variance. Subsequently, the desired optimal behavior policy μ^* can be easily calculated using (6.12). To ensure broad applicability, we utilize the behavior policy-agnostic offline learning setting (Nachum et al., 2019), in which the offline data consists of $\{(t_i, s_i, a_i, r_i, s_i')\}_{i=1}^m$, with m previously logged data tuples. Those tuples can be generated by various unknown behavior policies, and they are not required to form a complete trajectory. In the i-th data tuple, t_i represents the time step, s_i is the state at time step t_i, a_i is the action executed, r_i is

the sampled reward, and s_i' is the successor state. In this chapter, we learn (μ^*, b^*) from cheaply available offline data using Fitted Q -Evaluation (FQE, (Le et al., 2019)), but our framework is ready to integrate any state-of-the-art offline policy evaluation technique. As for constructing ν, we use the learned q function and r_i, s_i' from the data tuples, according to the derivation (B.22) in Appendix B.2.

6.5 Empirical Results

In this section, we show the empirical comparison between our methods and three baselines: **(1)** the on-policy Monte Carlo estimator, **(2)** the offline data informed estimator (ODI, Liu and Zhang (2024)), and **(3)** the doubly robust estimator (DR, Jiang and Li (2016)). In the doubly robust estimator, because they do not design a tailored behavior policy, we leverage the conventional wisdom to use the target policy π as the behavior policy. Given previously logged offline data, we learn our optimal behavior policy and the optimal baseline tuple (μ^*, b^*) using Algorithm 3. All baseline methods learn their required quantities from the same offline dataset to ensure fair comparisons. We use the behavior policy μ^* for data collection and the baseline b^* for data processing. Since our method reduces variance in both the data-collecting and the data-processing phases, we name our method doubly optimal (DOpt) policy evaluation. More experiment details are in Appendix B.2.

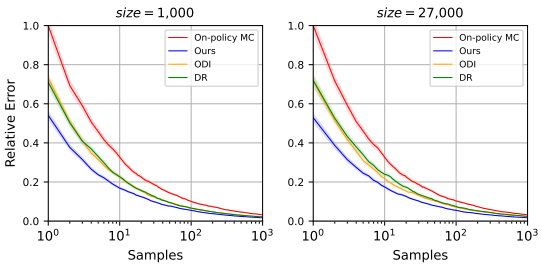

Figure 6.1: Results on Gridworld. Each curve is averaged over 900 runs (30 target policies, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves because they are too small.

Env Size	On-policy MC	Ours	ODI	DR
1,000	1.000	**0.274**	0.467	0.450
27,000	1.000	**0.283**	0.481	0.541

Table 6.1: Relative variance of estimators on Gridworld. The relative variance is defined as the variance of each estimator divided by the variance of the on-policy Monte Carlo estimator. Numbers are averaged over 900 independent runs (30 target policies, each having 30 independent runs).

Gridworld: We begin by conducting experiments in Gridworld with n^3 states, i.e., an $n \times n$ grid with n as the time horizon. The number of states in this Gridworld environment scales cubically with n, offering a suitable tool to test algorithm scalability.

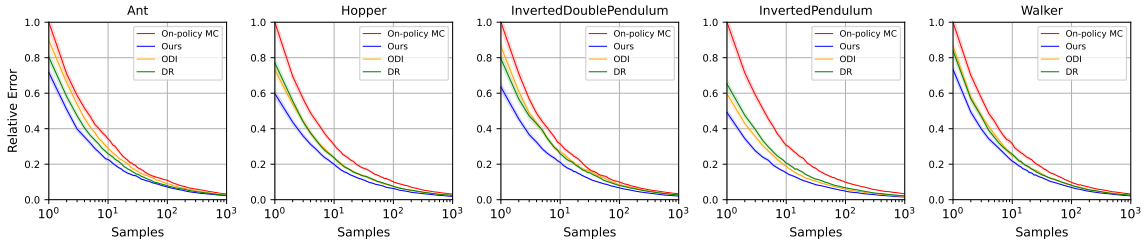

Figure 6.2: Results on MuJoCo. Each curve is averaged over 900 independent runs (30 target policies, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves because they are too small.

We choose Gridworld with $n^3 = 1,000$ and $n^3 = 27,000$, which are the largest Gridworld environment tested among related works (Jiang and Li, 2016; Hanna et al., 2017; Liu and Zhang, 2024). We use randomly generated reward functions with 30 randomly generated target policies. The offline data is generated by various unknown policies to simulate cheaply available but segmented offline data. Because MC methods use each episode as one empirical return sample, we view each episode as one online sample. We report the *relative error* of the four methods against the number of online samples. This relative error is the estimation error normalized by the estimation error of the on-policy Monte Carlo estimator after the first episode. Thus, the relative error of the on-policy Monte Carlo estimator starts from 1.

Figure 6.1 shows our method outperforms all baselines by a large margin. The blue line in the graph is below all other lines, indicating that our method requires fewer samples to achieve the same accuracy. This is because our designed (μ^*, b^*) substantially reduces estimation variance. In Table 6.1, we quantify such variance reduction, showing our method reduces variance by around 75% in both Gridworld with size $1,000$ and $27,000$.

One observation is that DR performs slightly better than ODI in smaller Gridworld but is slightly worse in larger Gridworld, which shows that there might be no dominating relationship between those two methods. Meanwhile, our method is superior to both approaches because the variance reduction of our method comes from both data-collecting and data-processing.

MuJoCo: We also conduct experiments in MuJoCo robot simulation tasks (Todorov et al., 2012). MuJoCo is a physics engine containing various stochastic environments, where the goal is to control a robot to achieve different behaviors such as walking, jumping, and balancing. Figure 6.2 shows our method is consistently better than all baselines in various MuJoCo robot environments. Table 6.2 shows our method requires substantially fewer samples to achieve the same estimation accuracy

	On-policy MC	**Ours**	ODI	DR	Saved Episodes Percentage
Ant	1000	**492**	810	636	(1000 - 492)/1000 = **50.8%**
Hopper	1000	**372**	544	582	(1000 - 372)/1000 = **62.8%**
I. D. Pendulum	1000	**426**	727	651	(1000 - 426)/1000 = **57.4%**
I. Pendulum	1000	**225**	356	439	(1000 - 225)/1000 = **77.5%**
Walker	1000	**475**	705	658	(1000 - 475)/1000 = **52.5%**

Table 6.2: Episodes needed to achieve the same of estimation accuracy that on-policy Monte Carlo achieves with 1000 episodes. Standard errors are plotted in Figure 6.2. Each number is averaged over 900 independet runs.

compared with the on-policy Monte Carlo method. Specifically, our method saves 50.8% to 77.5% of online interactions in different tasks, achieving state-of-the-art performance in policy evaluation.

It is worth mentioning that our method is robust to hyperparameter choices—all hyperparameters required to learn (μ^*, b^*) in our method are tuned offline and stay the same across all environments.

6.6 Discussion

Due to the sequential nature of reinforcement learning, policy evaluation often suffers from large variance and requires massive data to achieve the desired level of accuracy. In this work, we design an optimal combination of data-collecting policy μ^* and data-processing baseline b^*.

Theoretically, we prove our method considers larger policy space (Lemma 8), and is unbiased (Lemma 9) and optimal (Theorem 12). Further, we mathematically quantify the superiority of our method in variance reduction compared with existing methods (Theorem 13, 14, 15).

Empirically, compared with previous best-performing methods, we show our method reduces variance substantially in a broad range of environments, achieving state-of-the-art performance in policy evaluation.

One limitation is, as there is no free lunch, if the offline data size is too small—perhaps consisting of just a single data tuple—the behavior policy and baseline approximated by our method may be inaccurate. In this case, we recommend on-policy evaluation. The future work of our paper is to extend the variance reduction technique to temporal difference learning.

Chapter 7

Efficient Off-Policy Evaluation with Safety Constraint for Reinforcement Learning

This chapter is based on my paper Chen et al. (2025), published at ICLR 2025, in which I am a co–first author. I contributed to the initial idea, participated in the theoretical development, and was responsible for the experimental implementation.

In reinforcement learning, classic on-policy evaluation methods often suffer from high variance and require massive online data to attain the desired accuracy. Previous studies attempt to reduce evaluation variance by searching for or designing proper behavior policies to collect data. However, these approaches ignore the safety of such behavior policies—the designed behavior policies have no safety guarantee and may lead to severe damage during online executions. In this chapter, to address the challenge of reducing variance while ensuring safety simultaneously, we propose an optimal variance-minimizing behavior policy under safety constraints. Theoretically, while ensuring safety constraints, our evaluation method is unbiased and has lower variance than on-policy evaluation. Empirically, our method is the only existing method to achieve both substantial variance reduction and safety constraint satisfaction. Furthermore, we show our method is even superior to previous methods in both variance reduction and execution safety.

7.1 Preliminaries

In this chapter, we study off-policy evaluation under a constrained Markov decision process (CMDP), as defined in Section 2.3. For any set $\mathcal{X}$, we use $|\mathcal{X}|$ to denote its

cardinality. We use $\Delta^{|\mathcal{X}|-1}$ to denote the $(|\mathcal{X}|-1)$-dimensional probability simplex, representing the set of all probability distributions over the set $\mathcal{X}$.

In this work, we focus on off-policy evaluation. We restate the key ideas for off-policy evaluation here to facilitate reading. The goal is to estimate the total rewards $J(\pi)$ of an interested policy π, called the *target policy* by executing a different policy μ, called the *behavior policy*. We generate each trajectory $\{S_0, A_0, R_1, C_1, S_1, A_1, R_2, C_2, \ldots, S_{T-1}, A_{T-1}, R_T, C_T\}$ by a behavior policy μ with $S_0 \sim p_0, A_t \sim \mu_t(\cdot|S_t)$. For simplicity, we use a shorthand $\tau_{t:T-1}^{\mu_{t:T-1}}$ for a trajectory generated by the behavior policy μ from the time step t to the time step $T-1$ inclusively. It is defined as $\tau_{t:T-1}^{\mu_{t:T-1}} \doteq \{S_t, A_t, R_{t+1}, C_{t+1} \ldots, S_{T-1}, A_{T-1}, R_T, C_T\}$. In off-policy evaluation, to give an estimate of $J(\pi)$, we adopt the importance sampling ratio to reweigh rewards collected by the behavior policy μ. We define the importance sampling ratio at time t as $\rho_t \doteq \frac{\pi_t(A_t|S_t)}{\mu_t(A_t|S_t)}$. We also define the product of importance sampling ratios from time t to $t' \geq t$ as $\rho_{t:t'} \doteq \prod_{k=t}^{t'} \frac{\pi_k(A_k|S_k)}{\mu_k(A_k|S_k)}$. Various methods utilize importance sampling ratios within off-policy learning frameworks (Geweke, 1988; Hesterberg, 1995; Koller and Friedman, 2009; Thomas, 2015). In this chapter, We study the per-decision importance sampling estimator (PDIS, Precup et al. (2000a)). The PDIS Monte Carlo estimator is defined as $G^{\text{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \doteq \sum_{k=t}^{T-1} \rho_{t:k} R_{k+1}$. We also use the recursive expression of the PDIS estimator as

$$G^{\text{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) = \begin{cases} \rho_t \left(R_{t+1} + G^{\text{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}})\right) & t \in [T-2], \\ \rho_t R_{t+1} & t = T-1. \end{cases} \tag{7.1}$$

With the classic policy coverage assumption (Precup et al., 2000a; Maei, 2011; Sutton et al., 2016; Zhang, 2022; Liu et al., 2025b) $\forall t, s, a, \quad \mu_t(a|s) = 0 \implies \pi_t(a|s) = 0$, G^{PDIS} provides an *unbiased* estimation for $J(\pi)$, i.e., $\mathbb{E}\left[G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right] = J(\pi)$. Since the PDIS estimator is unbiased, reducing its variance is sufficient for improving its sample efficiency. We achieve this variance reduction by designing and learning proper behavior policies.

7.2 Constrained Variance Minimization for Contextual Bandits

In this section, we focus on variance minimization in policy evaluation under safety constraints in contextual bandits. These discussions provide the foundation for the more complicated optimization problems in sequential reinforcement learning settings, which we explore in Section 7.3. Notations defined in this section are independent of the rest of the paper.

We consider contextual bandits as one-step CMDPs, where the trajectories are in the form of (s, a, r, c). To estimate the performance of the target policy π, $\mathbb{E}_{a\sim\pi}[r(s,a)]$, with data collected by a behavior policy μ, we adopt the importance sampling ratio (Rubinstein, 1981) to reweigh the reward collected by μ. That is, we use $\mathbb{E}_{a\sim\mu}[\rho(a|s)r(s,a)]$ as an estimator, where $\rho(a|s) = \frac{\pi(a|s)}{\mu(a|s)}$. Recall $\Delta^{|\mathcal{A}|-1}$ is the probability simplex representing all probability distributions over the set $\mathcal{A}$. To ensure that this off-policy evaluation is unbiased, a classic choice by Rubinstein (1981) searches for μ in

$$\Lambda_{-} \doteq \left\{\mu \mid \forall s, a, \mu(a|s) = 0 \Rightarrow \pi(a|s) = 0 \wedge \forall s, \mu(\cdot|s) \in \Delta^{|\mathcal{A}|-1}\right\}.$$

In this work, we search in an enlarged space Λ (Owen, 2013; Liu and Zhang, 2024), where

$$\Lambda \doteq \left\{\mu \mid \forall s, a, \mu(a|s) = 0 \Rightarrow \pi(a|s)r(s,a) = 0 \wedge \forall s, \mu(\cdot|s) \in \Delta^{|\mathcal{A}|-1}\right\}. \tag{7.2}$$

Although a behavior policy μ in Λ may not cover the target policy π, μ still gives unbiased estimation in statistics. In the following lemma, we show that searching for μ in this enlarged space Λ guarantees unbiasedness in the contextual bandits setting.

Lemma 11. $\forall \mu \in \Lambda$, $\forall s$,

$$\mathbb{E}_{a\sim\mu}[\rho(a|s)r(s,a)] = \mathbb{E}_{a\sim\pi}[r(s,a)].$$

Its proof is in Appendix D.1.1. Our goal is to search for a variance-minimizing behavior policy μ. Except for the unbiasedness guaranteed by the search space Λ, we also require μ to satisfy safety constraints which will be defined later. We formulate the variance minimization objective as, $\forall s$,

$$\min_{\mu\in\Lambda} \quad \mathbb{V}_{a\sim\mu}(\rho(a|s)r(s,a)). \tag{7.3}$$

Then, with the unbiasedness in Lemma 11, we can further decompose the objective in (7.3) as

$$\begin{aligned}
\mathbb{V}_{a\sim\mu}(\rho(a|s)r(s,a)) =& \mathbb{E}_{a\sim\mu}[(\rho(a|s)r(s,a))^2] - \mathbb{E}_{a\sim\mu}[\rho(a|s)r(s,a)]^2 && (7.4)\\
=& \mathbb{E}_{a\sim\mu}[\rho(a|s)^2 r(s,a)^2] - \mathbb{E}_{a\sim\pi}[r(s,a)]^2. && \text{(By Lemma 11)}
\end{aligned}$$

Since the second term is a constant and is unrelated to μ, it suffices to solve

$$\min_{\mu\in\Lambda} \quad \mathbb{E}_{a\sim\mu}[\rho(a|s)^2 r(s,a)^2]. \tag{7.5}$$

Next, to ensure the safety of executing the behavior policy μ, we incorporate a safety constraint into the variance minimization problem. Since measuring safety by the expected cost is a common approach in the safety RL community (Berkenkamp et al., 2017; Achiam et al., 2017; Chow et al., 2018), we require that the expected cost of μ remains within a threshold related to the expected cost of π. Given a safety parameter $\epsilon \in [0, \infty)$, define a cost threshold

$$\delta_\epsilon(s) \doteq (1+\epsilon)\mathbb{E}_{a\sim\pi}[c(s,a)].$$

We impose the following constraint to the optimization problem (7.5)

$$\mathbb{E}_{a\sim\mu}[c(s,a)] \leq \delta_\epsilon(s), \quad \forall s. \tag{7.6}$$

This constraint requires that the expected cost of the designed behavior policy μ should be smaller than the multiple of the expected cost of the target policy π. By satisfying this constraint, we maintain a desired level of safety during the execution of the behavior policy μ. This safety is defined with respect to the target policy π, which is executed in the classic on-policy evaluation method. By setting $\epsilon = 0$, behavior policies satisfying this constraint are guaranteed to be safer than the target policy.

Notably, another line of research focused on policy safety chooses a constant threshold for the expected cost. We can simply modify (7.6) into a constant-threshold constraint by replacing the threshold function $\delta_\epsilon(s)$ with a constant δ. However, such absolute thresholds may make optimization problems infeasible. Strong assumptions on environments and policies have to be made to guarantee the existence of feasible solutions under absolute threshold (Achiam et al., 2017). In this chapter, we impose the safety constraint with respect to the target policy π, because our goal is to design a safe behavior policy to address the high variance associated with classic on-policy evaluation methods. The parameter ϵ in our threshold allows RL practitioners to adjust safety tolerance based on the specific requirements of the problem, as safety constraints are often highly problem-dependent (Achiam et al., 2017). In Section 7.5, we demonstrate our method in sequential reinforcement learning with a harsh threshold, $\epsilon = 0$, achieving both variance and cost reduction compared to the on-policy method.

We formally define our optimization problem and prove its convexity and feasibility in the following theorem.

Lemma 12. *For all ϵ and s, the following optimization problem is convex and feasible.*

$$\min_{\mu\in\Lambda} \quad \mathbb{E}_{a\sim\mu}[\rho(a|s)^2 r(s,a)^2], \tag{7.7}$$

$$s.t. \quad \mathbb{E}_{a\sim\mu}[c(s,a)] \leq \delta_\epsilon(s). \tag{7.8}$$

Its proof is in Appendix D.1.2. Use μ^* to denote the optimal solution of the above optimization problem. We have the following lemma.

Lemma 13. *For all ϵ and s, let μ^* be the optimal solution of optimization problem (7.7), we have*

$$\mathbb{V}_{a\sim\mu^*}(\rho(a|s)r(s,a)) \leq \mathbb{V}_{a\sim\pi}(r(s,a)).$$

Proof. We first show that the target policy π is always in the feasible set of the optimization problem (7.7). We define the set of feasible policies as

$$\mathcal{F} \doteq \{\mu \in \Lambda \mid \forall \epsilon, s, \mathbb{E}_{a\sim\mu}[c(s,a)] \leq \delta_\epsilon(s)\}. \tag{7.9}$$

Because $\epsilon \in [0, \infty)$, for the safety constraint, we have

$$\mathbb{E}_{a\sim\pi}[c(s,a)] \leq (1+\epsilon)\mathbb{E}_{a\sim\pi}[c(s,a)] = \delta_\epsilon(s).$$

By the definition of Λ (7.2), $\pi \in \Lambda$. Thus, $\pi \in \mathcal{F}$. Because

$$\mu^* \doteq \operatorname*{argmin}_{\mu\in\mathcal{F}} \mathbb{E}_{a\sim\mu}[\rho(a|s)^2 r(s,a)^2] \tag{7.10}$$

is the optimal solution, we have

$$\begin{aligned}
&\mathbb{V}_{a\sim\mu^*}(\rho(a|s)r(s,a)) \\
=&\mathbb{E}_{a\sim\mu^*}[\rho(a|s)^2 r(s,a)^2] - \mathbb{E}_{a\sim\pi}[r(s,a)]^2 && \text{(by (7.4))} \\
\leq&\mathbb{E}_{a\sim\pi}[\rho(a|s)^2 r(s,a)^2] - \mathbb{E}_{a\sim\pi}[r(s,a)]^2 && \text{(by (7.10))} \\
=&\mathbb{E}_{a\sim\pi}[r(s,a)^2] - \mathbb{E}_{a\sim\pi}[r(s,a)]^2 \\
=&\mathbb{V}_{a\sim\pi}(r(s,a)).
\end{aligned}$$

□

In Section 7.3, we expand Lemma 12 and Lemma 13 from contextual bandits to sequential reinforcement learning in Theorem 16 and Theorem 17. We show that with a recursive expression of the estimation variance, we can reduce the sequential problem into bandits in each time step t, and thereafter obtain the optimal behavior policy μ^* that minimizes variance under safety constraints.

7.3 Constrained Variance Minimization for Sequential Reinforcement Learning

In this section, we extend the techniques from contextual bandits to the sequential reinforcement learning setting. We seek to find an optimal behavior policy μ that reduces the variance $\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right)$ under safety constraints. Before defining the optimization problem, we first define the policy space we search for the behavior policy to ensure the unbiasedness of the PDIS estimator. Conventional methods search μ in the set of all policies that cover the target policy π (Sutton and Barto, 2018), i.e.,

$$\Lambda_{-} \doteq \{\mu \mid \forall t, s, a, \mu_t(a|s) = 0 \Rightarrow \pi_t(a|s) = 0 \wedge \forall t, s, \mu_t(\cdot|s) \in \Delta^{|\mathcal{A}|-1}\}.$$

In this chapter, similar to the bandits setting (7.2), we search in an enlarged set Λ, which is defined as

$$\Lambda \doteq \{\mu \mid \forall t, s, a, \mu_t(a|s) = 0 \Rightarrow \pi_t(a|s) q_{\pi,t}(s,a) = 0 \wedge \forall t, s, \mu_t(\cdot|s) \in \Delta^{|\mathcal{A}|-1}\} \quad (7.11)$$

The following lemma from Liu and Zhang (2024) ensures the unbiasedness of the off-policy estimator with the behavior policy $\mu \in \Lambda$.

Lemma 14. *$\forall \mu \in \Lambda$, $\forall t$, $\forall s$,*

$$\mathbb{E}\left[G^{PDIS}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right] = v_{\pi,t}(s).$$

Its proof is in Appendix D.1.3. A natural idea to do variance minimization under safety constraints with a safety parameter $\epsilon \in [0, \infty)$ is to solve the following optimization problem

$$\begin{aligned} \min_{\mu \in \Lambda} \quad & \mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu_{0:T-1}})\right), \\ \text{s.t.} \quad & J^c(\mu) \leq (1+\epsilon) J^c(\pi), \end{aligned} \quad (7.12)$$

where $J^c(\mu) \doteq \sum_s p_0(s) v_{\mu,0}^c(s)$ is the expected cost of the behavior policy μ. Solving this problem directly is very challenging. When designing a policy at a time step t, we need to consider not only the immediate reward generated by this action but also the future consequences. Hanna et al. (2017) try to solve this problem without safety constraints by directly optimizing the behavior policy μ with gradient descent. However, this approach requires online data to optimize μ and struggles in even moderately complicated environments as shown in Zhong et al. (2022) and Liu and Zhang (2024).

In this chapter, we therefore propose to solve this problem in a backward way while ensuring safety constraints. Given an ϵ, use

$$\delta_{\epsilon,t}(s) \doteq (1+\epsilon) v^c_{\pi,t}(s) \tag{7.13}$$

to denote the safety threshold. We define an extended reward function $\tilde{r}_t(s,a)$ and a behavior policy μ^*. They are defined in the order of $\left\{\tilde{r}_{T-1}, \mu^*_{T-1}, \tilde{r}_{T-2}, \mu^*_{T-2}, \cdots, \tilde{r}_0, \mu^*_0\right\}$. Denote the variance of the state value for the next state given the current state-action pair (s,a) as $\nu_{\pi,t}(s,a)$. We have

$$\nu_{\pi,t}(s,a) \doteq \begin{cases} 0 & t = T-1, \\ \mathbb{V}_{S_{t+1}}\left(v_{\pi,t+1}(S_{t+1}) \mid S_t = s, A_t = a\right) & t \in [T-2]. \end{cases}$$

Then, the extended reward function is defined as

$$\tilde{r}_t(s,a) \doteq \begin{cases} r_{\pi,t}(s,a)^2 & t = T-1, \\ \nu_{\pi,t}(s,a) + q_{\pi,t}(s,a)^2 + \mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau^{\mu^*_{t+1:T-1}}_{t+1:T-1}) \mid S_{t+1}\right) \mid s,a\right] & t \in [T-2]. \end{cases} \tag{7.14}$$

The behavior policy μ^*_t is defined as the optimal solution to the following problem. $\forall t, s$,

$$\begin{aligned} \min_{\mu_t \in \Lambda} \quad & \mathbb{E}_{a\sim\mu_t}[\rho_t^2 \tilde{r}_t(s,a)], \\ \text{s.t.} \quad & \mathbb{E}_{a\sim\mu_t}[q^c_{\mu,t}(s,a)] \leq \delta_{\epsilon,t}(s). \end{aligned}$$

We have the following theorem showing the convexity and feasibility of (7.15), thus ensuring the existence of the behavior policy μ^*.

Theorem 16. *For all $\epsilon \geq 0$, t, and s, the following optimization problem is convex and feasible.*

$$\min_{\mu_t \in \Lambda} \quad \mathbb{E}_{a\sim\mu_t}[\rho_t^2 \tilde{r}_t(s,a)], \tag{7.15}$$

$$s.t. \quad \mathbb{E}_{a\sim\mu_t}[q^c_{\mu,t}(s,a)] \leq \delta_{\epsilon,t}(s). \tag{7.16}$$

Its proof is in Appendix D.1.4. We notice that the constrained optimization problem (7.15) is similar to (7.7), which is the optimization problem introduced in Section 7.2. In the contextual bandit setting (7.7), we optimize the objective with respect to the reward function r, ensuring variance reduction (Lemma 13). In sequential reinforcement learning (7.15), we optimize with respect to the extended reward function $\tilde{r}$, achieving variance reduction (Theorem 17 and (7.17)), while

simultaneously guaranteeing safety (7.18). This observation provides a key insight: the step-wise optimization problem in *sequential reinforcement learning* can be viewed as a reduced optimization problem in one-step *contextual bandits*, where the reward is $\tilde{r}$. In Section 7.4, we further propose an efficient algorithm to learn $\tilde{r}$ without directly addressing the complicated trajectory variance $\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right)$, making long-horizon RL problems more tractable.

Theorem 17. *The behavior policy μ^* reduces variance compared with the on-policy evaluation method.*

$$\forall t, s, \mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\mu_{t:T-1}^*}) \mid S_t = s\right) \leq \mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right).$$

Its proof is in Appendix D.1.5. We also present the following theorem to demonstrate variance reduction and safety guarantee with respect to the original constrained optimization problem (7.12).

Theorem 18. *For all $\epsilon \geq 0$, the corresponding behavior policy μ^* has the following property*

$$1. \quad \mathbb{V}\left(G^{PDIS}(\tau_{0:T-1}^{\mu_{0:T-1}^*})\right) \leq \mathbb{V}\left(G^{PDIS}(\tau_{0:T-1}^{\pi_{0:T-1}})\right) \tag{7.17}$$

$$2. \quad J^c(\mu^*) \leq (1+\epsilon) J^c(\pi) \tag{7.18}$$

Its proof is in Appendix D.1.6. Notably, (7.18) shows that our step-wise safety-constraint (7.16) is stricter than the original constraint (7.12).

Algorithm 4: Safety-Constrained Optimal Policy Evaluation (SCOPE)

Input: a target policy π,
an offline dataset $\mathcal{D} = \{(t_i, s_i, a_i, r_i, c_i, s_i')\}_{i=1}^m$
Output: a behavior policy μ^*
Approximate $q_{\pi,t}, q_{\pi,t}^c$ from $\mathcal{D}$
for $t = T-1$ to 0 **do**
  Approximate $\tilde{r}_t$ from $\mathcal{D}$ by Lemma 15
  Approximate $\mu_t^*(a|s)$ following (7.15)
end for
Return: the approximated behavior policy μ^*

7.4 Learning the Optimal Behavior Policy

In this section, we propose an efficient algorithm to learn $\tilde{r}$ with previously logged offline data, and subsequently derive the optimal behavior policy μ^* under safety

constraints. We notice that learning $\tilde{r}$ by (7.14) is inefficient since we need to approximate the complicated variance $\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t\right)$, which involves the entire future trajectory. To tackle this challenge, we present a recursive expression of $\tilde{r}$ in the following lemma.

Lemma 15. *$\forall s, a$, when $t = T-1$, $\tilde{r}_t(s,a) = r_{\pi,t}(s,a)^2$. When $t \in [T-2]$,*

$$\tilde{r}_t(s,a) = 2q_{\pi,t}(s,a)r(s,a) - r(s,a)^2 + \mathbb{E}_{s'\sim p, a'\sim\mu^*}\left[\frac{\pi_{t+1}(a'|s')}{\mu^*_{t+1}(a'|s')}\tilde{r}_{\pi,t+1}(s',a')\right].$$

Its proof is in Appendix D.1.7. With this lemma, we can learn $\tilde{r}$ recursively without approximating the complicated trajectory variance. Then, by (D.14) in the appendix, we can also decompose the widely interested variance target in a succinct form

$$\underbrace{\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right)}_{(a)} = \underbrace{\mathbb{E}_{a\sim\mu^*}[\rho_t^2\tilde{r}_t(s,a)]}_{(b)} - \underbrace{v_{\pi,t}(s)^2}_{(c)}, \quad \forall s,t. \tag{7.19}$$

This succinct form offers a way to approximate the complicated trajectory variance term (a) from (b) and (c), which do not contain any variance term themselves. This is a surprising result because previously the best simplification of the variance for off-policy estimator (a) still depends on state-value variance terms (Jiang and Li, 2016; Liu and Zhang, 2024). With (7.19), we can approximate the variance of the off-policy estimator in a model-free way with only segmented offline data.

For broad applicability, we adopt the behavior policy-agnostic offline learning setting (Nachum et al., 2019), where the offline data has m previously logged data tuples in the form of $\{(t_i, s_i, a_i, r_i, c_i, s_i')\}_{i=1}^m$. These data tuples can be generated by one or more possibly unknown behavior policies, and they are not required to form a complete trajectory. In the i-th data tuple, t_i is the time step, s_i is the state at time step t_i, a_i is the action taken, r_i is the observed reward, c_i is the observed cost, and s_i' is the successor state. In this chapter, we learn $\tilde{r}$ from previously logged offline data. Previously logged offline data are cheap and readily available compared with online data. This makes them a great engine for improving policy evaluation in the online phase. Compared with gradient-based methods (Hanna et al., 2017; Zhong et al., 2022) which need complete online trajectories, our method does not require a long online warm-up time to find a good behavior policy because we are able to utilize offline data. Subsequently, the optimal variance-reducing behavior policy μ^* under safety constraints is approximated through standard convex optimization solvers (Nocedal and Wright, 1999; Agrawal et al., 2018).

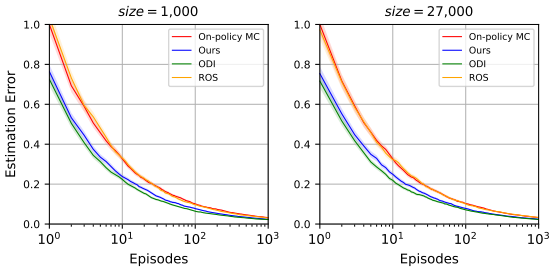

Figure 7.1: Results on Gridworld with *episodes* as x-axis. Each curve is averaged over 900 runs (30 target policies, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves as they are too small.

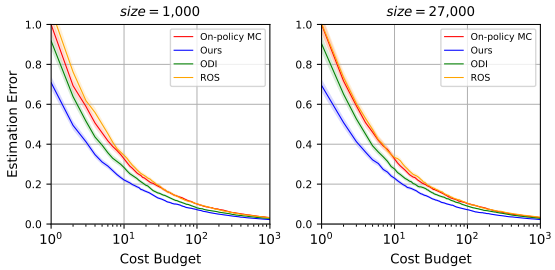

Figure 7.2: Results on Gridworld with *cost budget* as x-axis. *Cost budget* is the total cost of execution. Each curve is averaged over 900 runs (30 target policies, each having 30 independent runs). Shaded regions denote standard errors.

Env Size	On-policy MC	Ours	ODI	ROS
1,000	1.000	**0.861**	1.602	1.083
27,000	1.000	**0.849**	1.590	1.067

Table 7.1: Average trajectory cost on Gridworld. Numbers are normalized by the cost of the on-policy estimator. ODI and ROS have much larger costs because they both ignore safety constraints. Our method is the only method achieving both variance reduction and constraint satisfaction.

7.5 Empirical Results

In this section, we demonstrate the empirical results comparing our methods against three baselines: **(1)** the on-policy Monte Carlo estimator, **(2)** the robust on-policy sampling estimator (ROS, Zhong et al. (2022)), and **(3)** the offline data informed estimator (ODI, Liu and Zhang (2024)). To ensure our method attains lower cost and is thus even safer than the on-policy estimator, we choose $\epsilon = 0$ in the threshold $\delta_{\epsilon,t}$ (7.13). All methods learn their required parameters from the same offline dataset to ensure fair comparisons. Given previously logged offline data, our method learns the optimal behavior policy under safety constraints using Algorithm 4.

We name our algorithm Safety-Constrained Optimal Policy Evaluation (SCOPE) to emphasize that safety constraints are inherently considered in the design of the variance-minimizing behavior policy, unlike previous methods that overlook safety concerns. A metaphor for SCOPE is that it builds a bridge focused on efficient transportation (evaluation efficiency) while simultaneously ensuring traffic safety (satisfying safety constraints).

Gridworld: We first conduct experiments in Gridworld with n^3 states. Each Gridworld is an $n \times n$ grid with the time horizon also being n. Gridworld environments offer a great tool to test algorithm scalability, because the number of states scales cubically with n. Gridworld in our experiments have $n^3 = 1,000$ and $n^3 = 27,000$ number of states, which are the largest Gridworld environments tested among related works (Zhong et al., 2022; Liu and Zhang, 2024). We test all methods on target policies with various performances. The offline data is generated by many different policies to simulate previously logged offline data. In Figure 7.1, we report the estimation error against episodes. The estimation error for any line is the absolute error normalized by the absolute error of the on-policy estimator after the first episode. Thus, the estimation error of the on-policy estimator starts from 1. In Figure 7.2, we report the estimation error against the total cost of execution.

If considering *solely* variance reduction, Figure 7.1 shows our method outperforms the on-policy estimator and ROS by a large margin. Admittedly, ODI (Liu and Zhang, 2024) is slightly better than our method in terms of variance reduction. However, this slight advantage comes with a huge *trade-off* of safety. As shown in Table 7.1, ODI has a much larger cost than on-policy evaluation method (more than 1.5 times) and our method (almost twice as much). **This addresses the underestimated fact—solely reducing variance without safety constraints leads to high-cost (unsafe) methods.**

To further demonstrate the superiority of balancing variance reduction and safety cost of our method, we provide Figure 7.2 to compare the variance reduction each method achieves with the same cost budget. Since our method SCOPE is optimal for safety-constrained variance minimization, it consistently outperforms all baselines in Figure 7.2, as shown by the lowest blue line. This means that compared with existing best-performing methods, SCOPE needs less cost to achieve the same level of accuracy. From Figure 7.2, we compute that to achieve the same accuracy that the on-policy estimator achieves with 1000 costs (each on-policy episode has expected cost 1 by normalization), ODI costs 880 and SCOPE costs only 425. Following this computation, our method saves 57.5% of costs compared to the on-policy method, and 50% compared to ODI. **This reinforces the underestimated fact from the opposite direction—ensuring safety constraints along with the variance minimization leads to a low-cost method.** Also, notably, our estimator outperforms the on-policy and ROS estimators in *reducing both variance and cost*.

MuJoCo: Next, we conduct experiments in MuJoCo robot simulation tasks (Todorov et al., 2012). MuJoCo is a physics engine with a variety of stochastic

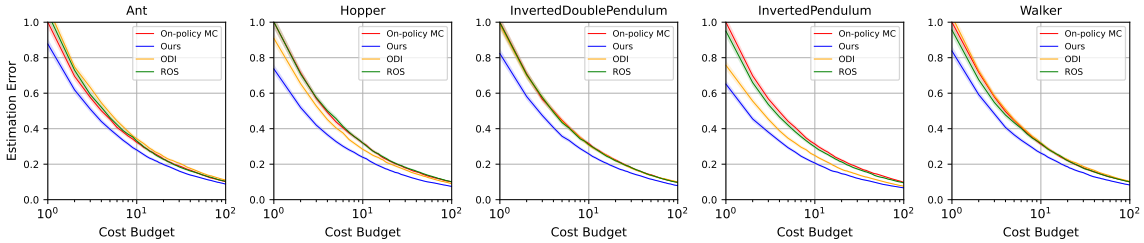

Figure 7.3: Results on MuJoCo. *Cost budget* on the x-axis is the total cost of execution. Each curve is averaged over 900 runs (30 of target policies, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves because they are too small. Results with a larger x-axis range are in the appendix.

	On-policy MC	**Ours**	ODI	ROS	Saved Cost Percentage
Ant	1000	**746**	1136	1063	(1000 - 746)/1000 = **25.4%**
Hopper	1000	**552**	824	1026	(1000 - 552)/1000 = **44.8%**
I. D. Pendulum	1000	**681**	1014	1003	(1000 - 681)/1000 = **31.9%**
I. Pendulum	1000	**425**	615	890	(1000 - 425)/1000 = **57.5%**
Walker	1000	**694**	1031	960	(1000 - 694)/1000 = **30.6%**

Table 7.2: Cost needed to achieve the same estimation accuracy that on-policy Monte Carlo achieves with 1000 episodes on MuJoCo. Each curve is averaged over 900 runs. Standard errors are plotted in Figure 7.3.

environments The goal is to control a robot to achieve different behaviors such as walking, jumping, and balancing.

As confirmed in Table D.3 and Table D.4 in the appendix, our method is the only method consistently achieving both variance reduction and safety constraint satisfaction. Figure 5.2 again indicates that our method consistently outperforms all baselines on reducing variance under the same cost budget. This advantage is observed across all five environments, demonstrating the stableness of our method in balancing variance reduction and cost management. Numerically, in Table 7.2, we show that our method, SCOPE, saves up to 57.5% cost to achieve the desired evaluation accuracy. More experiment details are in Appendix D.2. It is worth mentioning that our method is robust to hyperparameter choices—all hyperparameters in our method are tuned offline and stay the same across all environments.

7.6 Discussion

In reinforcement learning, due to the sequential nature, policy evaluation often suffers from large variance and thus requires massive data to achieve the desired level of accuracy. In addition, safety is a critical concern for policy execution, since unsafe actions can lead to significant risks and irreversible damage. In this chapter, we address these two challenges simultaneously: we propose an optimal variance-minimizing behavior policy under safety constraints.

Theoretically, we show that our estimate is unbiased. Moreover, while simultaneously satisfying safety constraints, our behavior policy is proven to achieve lower variance than the classic on-policy evaluation method (Theorem 17, Theorem 18). We solve the constrained optimization problem without approximating the complicated trajectory variance (Lemma 15), pointing out a promising direction for addressing long-horizon sequential reinforcement learning challenges.

Empirically, compared with existing best-performing methods, we show our method is the only one that achieves both substantial variance reduction and constraint satisfaction for policy evaluation in sequential reinforcement learning. Moreover, it is even superior to previous methods in both variance reduction and execution safety.

Admittedly, as there is no free lunch, if the offline data size is too small—perhaps containing merely a single data tuple—the learned behavior policy in our method may be inaccurate. In this case, for a safe backup, we recommend the on-policy evaluation method. The future work of our paper is to extend the constrained variance minimization technique to temporal difference learning.

Chapter 8

Efficient and Robust Policy Evaluation for Reinforcement Learning through Transition Gradient

In this chapter, we consider a setting where access to a simulator allows us to collect data and obtain heuristics about policy performance prior to a final evaluation stage in the real-world environment. Our goal is to reduce the amount of data required during this final online evaluation by minimizing the variance of off-policy estimates. To this end, we learn a variance-reducing behavior policy using data collected from the simulator. However, many reinforcement learning applications suffer from distributional shifts between the simulator and the real world, making standard off-policy evaluation methods unreliable. To address this, we propose a robust evaluation framework that optimizes against adversarial perturbations of the transition dynamics. By formulating the problem as a double-loop optimization, we improve evaluation robustness in potentially mismatched environments.

8.1 Preliminaries

In this work, we study a finite horizon Markov Decision Process (MDP, Puterman (2014)), which is introduced in Section 2.1. We redefine the concepts here to facilitate reading. The finite MDP contains a finite action space $\mathcal{A}$, a finite action space $\mathcal{A}$, a transition probability function $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, a reward function $r : \mathcal{S} \times \mathcal{A} \to [r_{\min}, r_{\max}]$, an initial state distribution $p_0 : \mathcal{S} \to [0, 1]$, and a constant horizon length T. For simplifying notations, we consider the undiscounted setting without loss of

generality. Our method naturally applies to the discounted setting as long as the horizon is fixed and finite (Puterman, 2014).

We define $\Delta(\mathcal{X})$ for a finite set $\mathcal{X}$ as the probability simplex over $\mathcal{X}$, i.e., $\Delta(\mathcal{X}) \doteq \{p : \mathcal{X} \to [0,1] \mid \sum_{x\in\mathcal{X}} p(x) = 1\}$. Then, the policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is the function mapping states to probability distribution over the action space $\mathcal{A}$. For gradient based methods, we consider the parameterized policies π_θ obtained via a neural network with a softmax output. The parameters $\theta \in \Theta$ is a vector, where $\Theta \in \mathbb{R}^n$ for some constant n. Likewise, we parameterize the transition function $p_\omega : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ by a parameters $\omega \in \Omega$ with $\Omega \in \mathbb{R}^m$, implemented using a neural network with a softmax layer.

We define a shorthand $[n] \doteq \{0, 1, \ldots, n\}$ for any integer n. The MDP process begins at time step 0, where an initial state S_0 is sampled from p_0. At each time step $t \in [T-1]$, an action A_t is sampled based on $\pi(\cdot \mid S_t)$. Then, a finite reward $R_{t+1} \doteq r(S_t, A_t)$ is given by the environment and a successor state S_{t+1} is obtained based on $p(\cdot \mid S_t, A_t)$. After T steps, the agent's interaction with the environment terminates. If the agent reaches any terminal state before time step T, it stays there and receives zero reward.

In this chapter, we use $h \doteq \{S_0, A_0, R_1, S_1, A_1, ...S_{T-1}, A_{T-1}, R_T\}$ to denote the trajectory of this MDP. We then define the *return* of h as $g(h) \doteq \sum_{t=0}^{T-1} R_{t+1}$. For any policy, we have a distribution over the trajectory as $\Pr(H = h|\pi)$, where H is a random variable used to denote the trajectory. Lastly, we define the *value* of a policy as $v(\pi) \doteq \mathbb{E}_{H\sim\pi}[g(H)]$.

We consider the task of off-policy evaluation, where the goal is to estimate the value of an interested policy π_e, which is called the *target policy*. We execute a different policy π_b, called the behavior policy, to collect data. Because this method evaluate the value of a policy π_e by running a different policy π_b, it is called *off-policy* evaluation. For wide applicability, we consider a general off-policy estimator, OPE, such as the importance sampling estimator IS and the per-decision importance sampling estimator PDIS. We denote such an estimator as $\text{OPE}(\pi_e, \pi_\theta, H)$, which estimates the value of the target policy using trajectory H collected by behavior policy π_θ parameterized by θ.

8.2 Adversarial Off-Policy Evaluation

We begin by studying the adversarial policy evaluation problem for a general off-policy estimator. While standard off-policy evaluation methods assume a fixed transition

model, real-world applications often involve uncertainties in transition dynamics due to model misspecifications, partial observability, or adversarial perturbations. To account for such uncertainties, we formulate the policy evaluation problem as a minimax optimization problem, where the worst-case transition model is identified to assess the robustness of policy evaluation. Our objective is to solve

$$\min_{\theta} \max_{\omega} \mathbb{V}_{H\sim p_\omega,\pi_\theta} \left[\mathrm{OPE}(\pi_e, \pi_\theta, H)\right]. \tag{8.1}$$

Here, the inner maximization over ω finds the transition perturbations that maximize the variance of the evaluation estimator, exposing potential weaknesses in policy evaluation. The outer minimization over θ then optimizes the behavior policy to mitigate these worst-case effects, ensuring that collected data remains informative even under adversarial dynamics. This formulation extends traditional OPE methods by explicitly considering transition uncertainty, making policy evaluation more robust in dynamic and non-stationary environments.

8.3 Solving the Inner Loop

8.3.1 On-Transition Gradient of the Variance

In this section, we focus the inner-loop of the optimization problem (8.1). Given a behavior policy π_θ, we look for the variance-optimizing adversarial transition p_ω. Formally, we need to solve

$$\max_{\omega} \mathbb{V}_{H\sim p_\omega,\pi_\theta} \left[\mathrm{OPE}(\pi_e, \pi_\theta, H)\right].$$

First, let the simulator's transition is the same as the target transition p_ω (namely, an on-transition case). In the following lemma, we give the gradient expression of the evaluation variance for any off-policy estimator OPE.

Lemma 16 (Transition Gradient of the Variance)**.** *For a fixed behavior policy π_θ,*

$$\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, H)] \\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}^2(\pi_e, \pi_\theta, H)\sum_{t=0}^{T-1}\tfrac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t, A_t))\right] \\
&- 2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, H)]\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e, \pi_\theta, H)\sum_{t=0}^{T-1}\tfrac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t, A_t))\right].
\end{aligned}$$

Its proof is in E.1.1. In Algorithm 5, we formalize our method with the important sampling off-policy estimator, IS, as a demonstration. To discuss the convergence

Algorithm 5: On-Transition Gradient (OG) for Variance.

1: **Input:** an initial transition parameter ω_0, a target policy π_e, a fixed behavior policy π_θ, a number of iteration n, a batch size k, a step-size α_i for each i
2: **Output:** a final adversarial transition parameter ω_n
3: **For all** $i \in 0, ..., n-1$ **do**
4: Sample k trajectories $H \sim \pi_\theta, p_{\omega_i}$
5: $\omega_{i+1} = \omega_i + \frac{\alpha_i}{k}\Big[\sum_{j=1}^{k}\Big(\mathrm{IS}^2(\pi_e,\pi_\theta,H_j)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega^j(S_{t+1}|S_t,A_t))\Big) - 2\sum_{j=1}^{\frac{k}{2}}\mathrm{IS}(\pi_e,\pi_\theta,H_j)\sum_{j=\frac{k}{2}+1}^{k}\Big(\mathrm{IS}(\pi_e,\pi_\theta,H_j)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega^j(S_{t+1}|S_t,A_t))\Big)\Big]$
6: **End for**
7: **Return:** ω_n

property of Algorithm 5, we make the widely used assumption on the step-size α_i at each iteration, which is known as the Robbins and Monroe condition (Robbins and Monro, 1951). We assume that the step-size α_i satisfies $\sum_{i=0}^{\infty}\alpha_i = \infty$ and $\sum_{i=0}^{\infty}\alpha_i^2 < \infty$. We also make the standard assumption that the quotient $\frac{\pi_e(a|s)}{\pi_\theta(a|s)}$ is bounded above for all s, a, and θ (Hanna et al., 2024). Then, we have the following lemma for the convergence of Algorithm 5.

Lemma 17 (Transition Gradient Convergence)**.** *For a fixed behavior policy π_θ, Algorithm 5 converges. That is, $\mathbb{V}_{H_i\sim p_{\omega_i},\pi_\theta}[\mathrm{IS}(\pi_e,\pi_\theta,H_i)]$ convereges to a finite value and $\lim_{i\to\infty}\frac{\partial}{\partial\omega}\mathbb{V}_{H_i\sim p_{\omega_i},\pi_\theta}[\mathrm{IS}(\pi_e,\pi_\theta,H_i)] = 0$.*

The proof of Lemma 17 is in Appendix E.1.2.

Although discrepancies often exist between the transition kernel in the deployment environment and the original simulator, the simulator typically remains a reasonable approximation. Thus, to ensure the learned adversarial transition remains realistic rather than overly pessimistic, we incorporate a Kullback–Leibler (KL) divergence penalty that discourages large deviations between p_ω and the initial simulator transition p_{ω_0}. That is, given a behavior policy π_e, we consider the following inner-loop optimization problem under KL regularization:

$$\max_{\omega}\mathbb{V}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,H)\right] - D_{\mathrm{KL}}(\Pr(H|p_\omega)\|\Pr(H|p_{\omega_0})),$$

where $\eta > 0$ is the regularization coefficient and the KL-divergence term is defined as $D_{\mathrm{KL}}(\Pr(H|p_\omega)\|\Pr(H|p_{\omega_0})) = \mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\log\frac{\Pr(H|p_\omega)}{\Pr(H|p_{\omega_0})}\right]$. To simplify notations, we define $\ell_{p_\omega} \doteq \sum_{t=0}^{T-1}\log(p_\omega(S_{t+1}|S_t,A_t))$. We provide the gradient expression of this regularized optimization problem in the following lemma.

Algorithm 6: On-Transition Gradient with KL (OGK) for Variance.

Input: an initial transition parameter ω_0, a target policy π_e, a fixed behavior policy π_θ, a number of iteration n, a batch size k, a step-size α_i for each i, a KL coefficient η

Output: a final adversarial transition parameter ω_n

For all $i \in 0, ..., n-1$ **do**

  Sample k trajectories $H \sim \pi_\theta, p_{\omega_i}$

  $\omega_{i+1} = \omega_i + \frac{\alpha_i}{k}\Big[\sum_{j=1}^{k}\left(\mathrm{IS}^2(\pi_e,\pi_\theta,H_j)\frac{\partial}{\partial\omega}\ell^j_{p_\omega}\right) -$
  $2\sum_{j=1}^{\frac{k}{2}}\mathrm{IS}(\pi_e,\pi_\theta,H_j)\sum_{j=\frac{k}{2}+1}^{k}\left(\mathrm{IS}(\pi_e,\pi_\theta,H_j)\frac{\partial}{\partial\omega}\ell^j_{p_\omega}\right)$
  $-\eta\sum_{j=1}^{k}\left(\frac{\partial}{\partial\omega}\ell^j_{p_\omega}\right)(1+\ell^j_{p_\omega}-\ell^j_{p_{\omega_0}})\Big]$

End for

Return: ω_n

Lemma 18 (Transition Gradient of Variance with KL)**.** *For a fixed behavior policy π_θ and a regularization coefficient $\eta > 0$,*

$$\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)] - D_{\mathrm{KL}}(\Pr(H|p_\omega)\|\Pr(H|p_{\omega_0}))\\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,H)\frac{\partial}{\partial\omega}\ell_{p_\omega}\right]\\
&-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)]\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,H)\frac{\partial}{\partial\omega}\ell_{p_\omega}\right]\\
&-\eta\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\left(\frac{\partial}{\partial\omega}\ell_{p_\omega}\right)\left(1+\ell_{p_\omega}-\ell_{p_{\omega_0}}\right)\right].
\end{aligned}$$

It proof is in Appendix E.1.3. Algorithm 6 formalizes this method with the importance sampling estimator as a demonstration.

8.3.2 Off-Transition Gradient of the Variance

In this section, we consider a setting where the simulator's transition dynamics, defined as p'_ω, remain unchanged. Since our target transition p_ω can differ from the simulator's transition $p_{\omega'}$, we refer to this as the "off-transition" setting.

Since we collect data from the simulator with different "behavior" transition, we have to reweigh the collected samples. For a general off-policy estimator OPE, we overload the notation as follows

$$\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H) \doteq \frac{\prod_{t=0}^{T-1}p_\omega(S_{t+1}|S_t,A_t)}{\prod_{t=0}^{T-1}p_{\omega'}(S_{t+1}|S_t,A_t)}\mathrm{OPE}(\pi_e,\pi_\theta,H).$$

We omit the input $p_{\omega'}$ in $\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)$ to simplify notations. Taking the

importance-sampling off-policy estimator, IS, as an example, we thus have

$$\begin{aligned}
&\mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \\
=&\frac{\prod_{t=0}^{T-1} p_\omega(S_{t+1}|S_t,A_t)}{\prod_{t=0}^{T-1} p_{\omega'}(S_{t+1}|S_t,A_t)}\mathrm{IS}(\pi_e, \pi_\theta, H) \\
=&\frac{\prod_{t=0}^{T-1} p_\omega(S_{t+1}|S_t, A_t)}{\prod_{t=0}^{T-1} p_{\omega'}(S_{t+1}|S_t, A_t)}\frac{\prod_{t=0}^{T-1} \pi_e(A_t|S_t)}{\prod_{t=0}^{T-1} \pi_\theta(A_t|S_t)} g(H).
\end{aligned}$$

We first give the gradient expression of the evaluation variance.

Lemma 19 (Off-Transition Gradient of Variance)**.** *When $p_\omega \neq p_{\omega'}$, for a fixed behavior policy π_θ,*

$$\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)] \\
=&2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}^2(\pi_e, \pi_\theta, p_\omega, H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right] \\
&- 2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)]\cdot\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right].
\end{aligned}$$

Its proof is in Appendix E.1.4. In the following lemma, we incorporate a Kullback-Leibler(KL) divergence term to penalize large deviations between p_ω and $p_{\omega'}$. That is, given a behavior policy π_e, we consider the following inner-loop optimization problem under KL regularization:

$$\max_\omega \mathbb{V}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)\right] - \eta D_{\mathrm{KL}}(\Pr(H|p_{\omega'})\|\Pr(H|p_\omega)),$$

where $\eta > 0$ is the regularization coefficient. We provide the gradient expression of this regularized optimization problem.

Lemma 20 (Off-transition Gradient of Variance with KL)**.** *For a fixed behavior policy π_θ and a regularization coefficient $\eta > 0$,*

$$\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)] - \eta D_{\mathrm{KL}}(\Pr(H|p_{\omega'})\|\Pr(H|p_\omega)) \\
=&2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}^2(\pi_e, \pi_\theta, p_\omega, H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right] \\
&- 2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)]\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right] \\
&- \eta\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[-\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right].
\end{aligned}$$

Its proof is in Appendix E.1.5.

Algorithm 7: Double-Loop Robust Gradient (DRG) for Variance

Input: an initial transition parameter ω_0, a target policy parameter θ_e, a number of iteration n, a batch size k, a step-size α_i for each i
Output: a final robust behavior policy parameter θ_n
$\theta_0 = \theta_e$
For all $i \in 0, ..., n-1$ **do**
    $\omega_{i+1} = \text{OGK}(\pi_e, \pi_{\theta_i}, p_{\omega_i})$
    Sample k trajectories $H \sim \pi_{\theta_i}, p_{\omega_{i+1}}$
    $\theta_{i+1} = \theta_i + \frac{\alpha_i}{k} \sum_{j=1}^{k} \text{IS}^2(\pi_e, \pi_\theta, H_j) \sum_{t=0}^{T-1} \frac{\partial}{\partial \omega} \log \pi_\theta^j(A_t|S_t)$
End for
Return: θ_n

8.4 Solving the Outer Loop

In this section, we propose an off-policy evaluation method that is robust to potential discrepancies in the environment. Specifically, we adopt a policy gradient approach to search for a variance-reducing behavior policy under an adversarial transition kernel. To begin with, we present the gradient expression for variance with respect to the behavior policy adopted from Hanna et al. (2017).

Lemma 21 (Variance Gradient Expression)**.** *With a fixed transition kernel p_ω, $\forall \theta$,*

$$\frac{\partial}{\partial \theta} V_{H \sim p_\omega, \pi_\theta}[\text{IS}(\pi_e, \pi_\theta, H)] = \mathbb{E}_{H \sim p_\omega, \pi_\theta}[-\text{IS}(\pi_e, \pi_\theta, H)^2 \sum_{t=0}^{T-1} \frac{\partial}{\partial \theta} \log \pi_\theta(A_t|S_t)].$$

For simplicity, we denote our inner-loop Algorithm 6 as $\text{OGK}(\pi_e, \pi_\theta, p_{\omega_0})$, where the latter p_{ω_0} serves as the initial transition kernel for Algorithm 6. We have the double-loop method as formalized in Algorithm 7. In the inner loop, $\text{OGK}(\pi_e, \pi_\theta, p_{\omega_0})$ returns an adversarial transition p_ω by performing gradient ascent on the variance objective penalized by a KL divergence. The outer loop then updates the behavior policy parameters via policy gradient to reduce the variance induced by the worst-case transition. Together, this double-loop procedure yields a robust behavior policy for off-policy evaluation under environment uncertainty.

8.5 Empirical Results

In this section, we first show the effectivity of Algorithm 5. Specifically, we demonstrate the variance increasing ability of our algorithm in various Gridworld environments. To the best of our knowledge, this is the first method that adversarially perturbs the transition kernel for evaluation variance in reinforcement learning.

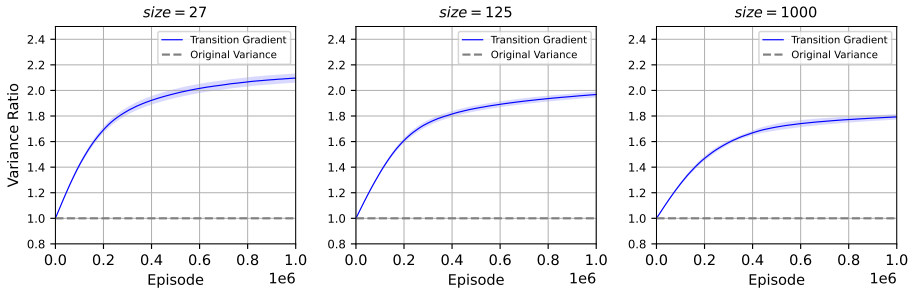

Figure 8.1: Results on Gridworld. Each curve is averaged over 30 training trajectories of transition probability. Shaded regions denote standard errors.

We conduct experiments in Gridworld with n^3 total states, comprising of an $n \times n$ grid with a time horizon of n. Since the number of states scales cubically with the choice of n, such environments provide a great tool to test the scalability of algorithms. We use three different sizes of the environments with $n^3 = 27$, $n^3 = 125$ and $n^3 = 1000$ respectively. In comparison, related works in robust RL conduct experiments with total states of no more than 100 (Grand-Clément and Kroer, 2021; Wang et al., 2023; Sun et al., 2024).

As shown in Figure 8.1, our algorithm consistently increases the evaluation variance across all environment sizes. The variance ratio grows with training episodes and stabilizes at a higher level than the original variance baseline.

Chapter 9

The ODE Method for Stochastic Approximation and Reinforcement Learning with Markovian Noise

This chapter is based on my paper Liu et al. (2025b) published at JMLR 2025.

Stochastic approximation is a class of algorithms that update a vector iteratively, incrementally, and stochastically, including, e.g., stochastic gradient descent and temporal difference learning. One fundamental challenge in analyzing a stochastic approximation algorithm is to establish its stability, i.e., to show that the stochastic vector iterates are bounded almost surely. In this chapter, we extend the celebrated Borkar-Meyn theorem for stability from the Martingale difference noise setting to the Markovian noise setting, which greatly improves its applicability in reinforcement learning, especially in those off-policy reinforcement learning algorithms with linear function approximation and eligibility traces. Central to our analysis is the diminishing asymptotic rate of change of a few functions, which is implied by both a form of the strong law of large numbers and a form of the law of the iterated logarithm.

9.1 Preliminaries

Stochastic approximation (Robbins and Monro, 1951; Benveniste et al., 1990; Kushner and Yin, 2003; Borkar, 2009) is a class of algorithms that update a vector iteratively, incrementally, and stochastically. Successful examples include stochastic gradient descent (Kiefer and Wolfowitz, 1952) and temporal difference learning (Sutton, 1988). Given an initial $x_0 \in \mathbb{R}^d$, stochastic approximation algorithms typically generate a sequence of vectors $\{x_n\}$ recursively as

$$x_{n+1} = x_n + \alpha(n) H(x_n, Y_{n+1}) \quad n = 0, 1, \ldots \tag{9.1}$$

Here $\{\alpha(n)\}_{n=0}^{\infty}$ is a sequence of deterministic learning rates, $\{Y_n\}_{n=1}^{\infty}$ is a sequence of random noise in a general space $\mathcal{Y}$ (not necessarily compact), and $H : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$ is a function that maps the current iterate x_n and noise Y_{n+1} to the actual incremental update.

One way to analyze the asymptotic behavior of $\{x_n\}$ is to regard $\{x_n\}$ as Euler's discretization of the ODE

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = h(x(t)), \tag{9.2}$$

where $h(x) \doteq \mathbb{E}[H(x, y)]$ is the expected updates (the expectation will be rigorously defined shortly). Then the asymptotic behavior of the discrete and stochastic iterates $\{x_n\}$ can be characterized by continuous and deterministic trajectories of the ODE (9.2). To establish this connection between the two, however, requires to establish the stability of $\{x_n\}$ first (Kushner and Yin, 2003; Borkar, 2009). In other words, one needs to first show that

$$\sup_n \|x_n\| < \infty \quad \text{a.s.},$$

which is in general challenging. Once the stability is confirmed, the convergence of $\{x_n\}$ follows easily (Kushner and Yin, 2003; Borkar, 2009). The seminal Borkar-Meyn theorem (Borkar and Meyn, 2000) establishes the desired stability assuming the global asymptotic stability of the following ODE

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = h_\infty(x),$$

where $h_\infty(x) \doteq \frac{h(cx)}{c}$. Despite the celebrated success of the Borkar-Meyn theorem (see, e.g., Abounadi et al. (2001); Maei (2011)), one major limit is that the Borkar-Meyn theorem requires $\{Y_n\}$ to be i.i.d. noise. As a result, $\{H(x_n, Y_{n+1}) - h(x_n)\}_{n=0}^{\infty}$ is then a Martingale difference sequence and the Martingale convergence theorem applies under certain conditions. However, in many Reinforcement Learning (RL, Sutton and Barto (2018)) problems, $\{Y_n\}$ is a Markov chain and is not i.i.d. *Our main contribution is to extend the Borkar-Meyn theorem to the Markovian noise setting with verifiable assumptions.* The extension to Markovian noise has been previously explored by Ramaswamy and Bhatnagar (2018); Borkar et al. (2021). However, their assumptions are way more restrictive than ours so their results are not applicable in many important RL algorithms, particularly, off-policy RL algorithms with eligibility traces (Yu, 2012, 2015, 2017). See Section 9.5 for more discussion on this class of RL algorithms.

In Ramaswamy and Bhatnagar (2018), it is assumed that the Differential Inclusion (DI)

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} \in \overline{\mathrm{co}}\{H_\infty(x(t), y) | y \in \mathcal{Y}\}$$

is stable, where $\overline{\mathrm{co}}(\cdot)$ denotes the convex hull and $H_\infty(x, y) \doteq \lim_{c\to\infty} \frac{H(cx,y)}{c}$. To demonstrate the challenge in verifying this assumption, we consider a special linear case where $H(x, y) = A(y)x + b(y)$ for some matrix-valued function $A(y)$ and vector-valued function $b(y)$. Then one sufficient and commonly used condition (Molchanov and Pyatnitskiy, 1989) for this DI to be stable is that the $A(y)$ is uniformly negative definite, i.e., there exists some strictly positive η such that $x^\top A(y)x \leq -\eta\|x\|^2 \, \forall x \in \mathbb{R}^d, y \in \mathcal{Y}$. However, in many RL algorithms (e.g., Sutton (1988); Sutton et al. (2008b, 2009, 2016), as well as the off-policy RL algorithms with eligibility traces in Section 9.5), we can at most say that $\mathbb{E}[A(y)]$ is negative definite. The individual matrix $A(y)$ does not have any special property. Intuitively, Ramaswamy and Bhatnagar (2018) assume that the function $H_\infty(x, y)$ behaves well almost surely, significantly limiting its application in RL. In fact, we are not aware of any application of Ramaswamy and Bhatnagar (2018) in standard RL algorithms. By contrast, we only need $h_\infty(x)$ to behave well, i.e., we only need $H_\infty(x, y)$ to behave well in expectation. Ramaswamy and Bhatnagar (2018) also assume $\mathcal{Y}$ to be compact. Unfortunately, in many important RL algorithms mentioned above, neither DI's stability nor the compactness holds.

In Borkar et al. (2021), it is assumed that a V4 Laypunov drift condition holds for $\{Y_n\}$ and the eighth moment of some function is bounded. Unfortunately, in many important RL algorithms (see, e.g., those in Section 9.5), neither assumption holds. We instead establish the stability via examining the *asymptotic rate of change* of certain functions, inspired by Kushner and Yin (2003). When V4 does not hold, a form of the strong law of large numbers and a form of the law of the iterated logarithm can be used to establish the desired asymptotic rate of change. When V4 does hold, we only need the second moment, instead of the eighth moment, to be bounded to establish the desired asymptotic rate of change.

We demonstrate in Section 9.5 the wide applicability of our results in RL, especially in off-policy RL algorithms with linear function approximation and eligibility traces, where the Markovian noise $\{Y_n\}$ can easily grow *unbounded almost surely* and have *unbounded second moment*. The key idea of our approach is to apply the Arzela-Ascoli theorem to the scaled iterates. Then the Moore-Osgood theorem computes a double limit, confirming that the scaled iterates converge to the corresponding limiting ODEs along a carefully chosen *subsequence*. This subsequence view is an important technical

innovation of our approach. By contrast, previous works concerning the Borkar-Meyn theorem (Borkar and Meyn, 2000; Bhatnagar, 2011; Lakshminarayanan and Bhatnagar, 2017; Ramaswamy and Bhatnagar, 2017, 2018; Borkar et al., 2021) all seek to establish the convergence along the entire sequence to invoke a proof by contradiction argument to establish the desired stability. This subsequence view is essential for our approach because the Arzela-Ascoli theorem can only guarantee the existence of a convergent subsequence. As a result, we need a variant of the standard proof by contradiction argument to establish the desired stability.

9.2 Main Results

Assumption 1. *The Markov chain $\{Y_n\}$ has a unique invariant probability measure (i.e., stationary distribution), denoted by $d_{\mathcal{Y}}$.*

Technically speaking, the uniqueness and even the existence of the invariant probability measure can be relaxed, as long as the average of certain functions exists. We are, however, not aware of any applications where such relaxation is a must. We, therefore, use Assumption 1 to ease presentation and refer the reader to A1.3 in Chapter 6 of Kushner and Yin (2003) as an example of such relaxation. In light of the update (9.1), we use the convention that $\{Y_n\}$ starts from $n = 1$.

Assumption 2. *The learning rates $\{\alpha(n)\}$ are positive, decreasing, and satisfy*

$$\sum_{i=0}^{\infty} \alpha(i) = \infty, \ \lim_{n\to\infty} \alpha(n) = 0, \text{ and } \frac{\alpha(n) - \alpha(n+1)}{\alpha(n)} = \mathcal{O}\left(\alpha(n)\right). \tag{9.3}$$

Remark 1. *For any $\alpha(n) = \frac{B_1}{(n+B_2)^\beta}$ with $\beta \in (0.5, 1]$, it can be easily computed that*

$$\frac{\alpha(n) - \alpha(n+1)}{\alpha(n)} = \mathcal{O}\left(\frac{\beta}{n}\right) = \mathcal{O}\left(\alpha(n)\right).$$

Next, we make a few assumptions about the function H. For any $c \in [1, \infty)$, define

$$H_c(x, y) \doteq \frac{H(cx, y)}{c}. \tag{9.4}$$

The function H_c is the rescaled version of the function H and will be used to construct rescaled iterates, which are key techniques in proving the Borkar-Meyn theorem (see, e.g., Borkar and Meyn (2000); Borkar (2009)). Similar to Borkar and Meyn (2000); Borkar (2009), we need the limit of H_c to exist in a certain sense when $c \to \infty$.

Assumption 3. *There exists a measurable function $H_\infty(x,y)$, a function $\kappa:\mathbb{R}\to\mathbb{R}$ (independent of x,y), and a measurable function $b(x,y)$ such that for any x,y,*

$$\begin{aligned} H_c(x,y)-H_\infty(x,y) =&\kappa(c)b(x,y), \\ \lim_{c\to\infty}\kappa(c) =&0, \end{aligned} \tag{9.5}$$

Moreover, there exists a measurable function $L_b(y)$ such that $\forall x,x',y$,

$$\|b(x,y)-b(x',y)\|\le L_b(y)\|x-x'\|. \tag{9.6}$$

And the expectation $L_b \doteq \mathbb{E}_{y\sim d_\mathcal{Y}}[L_b(y)]$ is well-defined and finite.

Assumption 3 provides details on how H_c converges to H_∞ when $c\to\infty$. We note that in many RL applications, see, e.g., Section 9.5, the function $b(x,y)$ actually does not depend on x so (9.6) trivially holds. We consider $b(x,y)$ as a function of both x and y for generality. Next, we assume Lipschitz continuity of the functions H_c, which guarantees the existence and uniqueness of the solutions to the corresponding ODEs.

Assumption 4. *There exists a measurable function $L(y)$ such that for any x,x',y,*

$$\|H(x,y)-H(x',y)\|\le L(y)\|x-x'\|, \tag{9.7}$$

$$\|H_\infty(x,y)-H_\infty(x',y)\|\le L(y)\|x-x'\|. \tag{9.8}$$

Moreover, the following expectations are well-defined and finite for any x:

$$\begin{aligned} h(x) &\doteq \mathbb{E}_{y\sim d_\mathcal{Y}}[H(x,y)], \\ h_\infty(x) &\doteq \mathbb{E}_{y\sim d_\mathcal{Y}}[H_\infty(x,y)], \\ L &\doteq \mathbb{E}_{y\sim d_\mathcal{Y}}[L(y)]. \end{aligned}$$

Apparently, the function $x\mapsto H_c(x,y)$ shares the same Lipschitz constant $L(y)$ as the function $x\mapsto H(x,y)$. Similar to (9.4), we define

$$h_c(x)\doteq\frac{h(cx)}{c}.$$

The following assumption is the central assumption in the original proof of the Borkar-Meyn theorem.

Assumption 5. *(Assumption A5 in Chapter 3 of Borkar (2009)) As $c\to\infty$, $h_c(x)$ converges to $h_\infty(x)$ uniformly in x on any compact subsets of $\mathbb{R}^d$. The ODE*

$$\frac{dx(t)}{dt}=h_\infty(x(t)) \tag{ODE@∞}$$

has 0 as its globally asymptotically stable equilibrium.

We refer the reader to Dai (1995); Dai and Meyn (1995); Borkar and Meyn (2000); Borkar (2009); Fort et al. (2008); Meyn (2008, 2022) for the root and history of (ODE@∞).

Assumption 6. *Let g denote any of the following functions:*

$$y \mapsto H(x,y) \quad (\forall x), \tag{9.9}$$

$$y \mapsto L_b(y), \tag{9.10}$$

$$y \mapsto L(y). \tag{9.11}$$

Then for any initial condition Y_1, it holds that

$$\lim_{n\to\infty} \alpha(n) \sum_{i=1}^{n} \left(g(Y_i) - \mathbb{E}_{y\sim d_{\mathcal{Y}}}\left[g(y)\right]\right) = 0 \quad a.s. \tag{9.12}$$

Remark 2. *Consider $\alpha(n) = \frac{B_1}{(n+B_2)^\beta}$ as an example again. For $\beta = 1$, (9.12) is implied by the following Law of Large Numbers (LLN)*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(g(Y_i) - \mathbb{E}_{y\sim d_{\mathcal{Y}}}\left[g(y)\right]\right) = 0 \quad a.s. \tag{LLN}$$

For $\beta \in (0.5, 1]$, (9.12) is implied by the following Law of the Iterated Logarithm (LIL)

$$\left\| \sum_{i=1}^{n} \left(g(Y_n) - \mathbb{E}_{y\sim d_{\mathcal{Y}}}\left[g(y)\right]\right) \right\| \leq \zeta \sqrt{n \log\log n} \quad a.s., \tag{LIL}$$

where ζ is a sample path dependent finite constant.

Remark 3. *If the Markov chain $\{Y_n\}$ is positive[1] Harris[2], then (LLN) holds for any function g whenever $\mathbb{E}[\|g(y)\|] < \infty$ (Theorem 17.0.1 (i) of Meyn and Tweedie (2012)). If $\{Y_n\}$ is further V-uniformly ergodic[3], then (LIL) holds (Theorem 17.0.1 (iii) and (iv) of Meyn and Tweedie (2012)). For the special case where $\mathcal{Y}$ is finite, (LLN) holds when the Markov chain is irreducible and (LIL) holds when it is further aperiodic.*

Remark 4. *We note that (LLN) is stronger than Doob's strong law of large numbers on stationary processes (see, e.g., Theorem 17.1.2 of Meyn and Tweedie (2012), referred to as Doob's LLN hereafter). Doob's LLN concludes (at most) that (LLN) holds for any $Y_1 \in \mathcal{Y}_g$, where $\mathcal{Y}_g$ is an unknown, probably g-dependent set such that $d_{\mathcal{Y}}(\mathcal{Y}_g) = 1$. If we use only Doob's LLN, all the "almost surely" statements in the paper must be*

[1] See page 235 of Meyn and Tweedie (2012) for the definition of positive chains.
[2] See page 204 of Meyn and Tweedie (2012) for the definition of Harris chains.
[3] See page 387 of Meyn and Tweedie (2012) for the definition of V-uniform ergodicity.

replaced by "$\mathcal{Y}_*$-almost surely", where $\mathcal{Y}_* \doteq \bigcap_g \mathcal{Y}_g$. This means that all the statements hold only when $Y_1 \in \mathcal{Y}_*$. However, since the g functions in Assumption 6 depend on x, this $\mathcal{Y}_*$ is an intersection of possibly uncountably many sets $\{\mathcal{Y}_g\}$. It is possible that in some applications $\mathcal{Y}_*$ turns out to be a set of interest, where (LLN) can indeed be relaxed to Doob's LLN. But in general, characterizing $\mathcal{Y}_*$ is pretty challenging.

Remark 5. *The Markov chain $\{Y_n\}$ we consider in our RL applications in Section 9.5 is a general space Markov chain but is not positive Harris. Fortunately, Yu (2012, 2015, 2017) have established that (LLN) holds for those chains. Whether (LIL) holds for those chains remains open.*

To better contrast our work with Borkar et al. (2021), in the following, we provide an alternative to Assumption 6.

Assumption 6′. *The learning rates $\{\alpha(n)\}$ further satisfy $\sum_{n=0}^{\infty} \alpha(n)^2 < \infty$. The Markov chain $\{Y_n\}$ is ψ-irreducible[4]. The Lyapunov drift condition (V4) holds for the Markov chain $\{Y_n\}$.[5] In other words, there exists a Lyapunov function $v : \mathcal{Y} \to [1, \infty]$ such that for any $y \in \mathcal{Y}$,*

$$\mathbb{E}\left[v(Y_{n+1}) - v(Y_n) | Y_n = y\right] \leq -\delta v(y) + \tau \mathbb{I}_C(y). \tag{V4}$$

Here $\delta > 0, \tau < \infty$ are constants, C is a small set[6], and $\mathbb{I}$ is the indicator function. Moreover, let g be any of the functions $H(0, y), L_b(y)$, and $L(y)$. Then $g \in \mathcal{L}^2_{v,\infty}$[7].

Assumption 6′ uses the idea of Borkar et al. (2021) but is weaker than its counterparts. See more detailed comparisons in Section 9.3.

Remark 6. *Assumption 6′ is listed here mostly for better comparison with Borkar et al. (2021). We are not aware of any RL application where Assumption 6′ holds but Assumption 6 does not hold. Instead, in the RL applications in Section 9.5, Assumption 6 holds but Assumption 6′ does not. That being said, the applicability of Assumptions 6 and 6′ outside RL is beyond the scope of this work.*

Having listed all the assumptions, our main theorem confirms the stability of $\{x_n\}$.

[4] See page 91 of Meyn and Tweedie (2012) for the definition of ψ-irreducibility.
[5] See page 371 of Meyn and Tweedie (2012) for in-depth discussion about (V4).
[6] See page 109 of Meyn and Tweedie (2012) for the definition of small sets.
[7] g belongs to $\mathcal{L}^p_{v,\infty}$ if and only if $\sup_{y \in \mathcal{Y}} \frac{\|g(y)\|_p^p}{v(y)} < \infty$, where v is the Lyapunov function in (V4).

Theorem 19. *Let Assumptions 1 - 5 hold. Let Assumption 6 or 6′ hold. Then the iterates* $\{x_n\}$ *generated by* (9.1) *are stable, i.e.,*

$$\sup_n \|x_n\| < \infty \quad a.s.$$

Its proof is in Section 9.4. Once the stability is established, the convergence follows easily.

Corollary 1. *Let Assumptions 1 - 5 hold. Let Assumption 6 or 6′ hold. Then the iterates* $\{x_n\}$ *generated by* (9.1) *converge almost surely to a (sample path dependent) bounded invariant set*[8] *of the ODE*[9]

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = h(x(t)). \tag{9.13}$$

Arguments used in proving Corollary 1 are similar but much simpler than the counterparts in the proof of Theorem 19. We include a proof of Corollary 1 in Appendix F.2.9 with the details of those similar but simpler lemmas omitted to avoid verbatim repetition.

It is worth mentioning that it is easy to extend our results to more general updates

$$x_{n+1} = x_n + \alpha(n)\left(H(x_n, Y_{n+1}) + M_{n+1} + \epsilon_n\right),$$

where M_{n+1} is a Martingale difference sequence and ϵ_n is another additive noise. Similarly, it would require the asymptotic rate of change of $\{M_{n+1}\}$ and $\{\epsilon_n\}$ to diminish. We refer the reader to Kushner and Yin (2003) for more details. Since our main contribution is the stability under the Markovian noise $\{Y_{n+1}\}$, we use the simpler update rule (9.1) for improving clarity.

9.3 Prior Work

General H. In this chapter, the function H can be a general function and we do not make any linearity assumptions. We first compare our results with existing works applicable to general H and Markovian noise $\{Y_n\}$. Since convergence follows easily from stability, we focus on comparison in terms of establishing stability. Notably, the

[8] A set X is an invariant set of the ODE (9.13) if and only if for every $x \in X$, there exists a solution $x(t)$ to the ODE (9.13) such that $x(0) = x$ and $x(t) \in X$ for all $t \in (-\infty, \infty)$. If the ODE (9.13) is globally asymptotically stable, the only bounded invariant set is the singleton $\{x_*\}$, where x_* denotes the unique globally asymptotically stable equilibrium. We refer the reader to page 105 of Kushner and Yin (2003) for more details.

[9] By $\{x_n\}$ converges to a set X, we mean $\lim_{n\to\infty} \inf_{x\in X} \|x_n - x\| = 0$.

related stability results in Borkar and Meyn (2000); Borkar (2009) are superceded by Borkar et al. (2021). We, therefore, discuss only Borkar et al. (2021); Kushner and Yin (2003); Benveniste et al. (1990).

Compared with Borkar et al. (2021), our improvements lie in two aspects. First, central to Borkar et al. (2021) are (i) a V4 Laypunov drift condition, (ii) an aperiodicity assumption of $\{Y_n\}$, and (iii) a boundedness assumption $L(y) \in \mathcal{L}^8_{v,\infty}$. By contrast, our Assumption 6′ only requires $L(y) \in \mathcal{L}^2_{v,\infty}$ and does not need aperiodicity. Second, we further provide an approach that establishes the stability based on Assumption 6 without using (V4), aperiodicity, and the boundedness in $\mathcal{L}^8_{v,\infty}$. As noted in Remark 6, Assumption 6 is more applicable than Assumption 6′ in RL.

Compared with Kushner and Yin (2003), our main improvement is that we prove stability under the asymptotic rate of change conditions. By contrast, Kushner and Yin (2003) mostly use stability as a priori and are concerned with the convergence of projected algorithms in the form of

$$x_{n+1} = \Pi\left(x_n + \alpha(n) H(x_n, Y_{n+1})\right),$$

where Π is a projection to some compact set to ensure stability of $\{x_n\}$. As a result, the corresponding ODE (cf. Corollary 1) becomes

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = h(x(t)) + \xi(t),$$

where $\xi(t)$ is a reflection term resulting from the projection Π. We refer the reader to Section 5.2 of Kushner and Yin (2003) for more details regarding this reflection term. Analyzing these reflection terms typically requires strong domain knowledge, see, e.g., Yu (2015); Zhang et al. (2021b), and Section 5.4 of Borkar (2009).

We argue that this work combines the best of both Borkar and Meyn (2000) and Kushner and Yin (2003), i.e., the ODE@∞ technique for establishing stability from Borkar and Meyn (2000) and the asymptotic rate of change technique for averaging out the Markovian noise $\{Y_n\}$. As a result, our results are more general than both Borkar et al. (2021) and Kushner and Yin (2003) in the aforementioned sense.

Compared with Benveniste et al. (1990), our main improvement is that despite the proof under Assumption 6′ essentially using Poisson's equation[10], the proof under Assumption 6 does not need Poisson's equation at all. Notably, Benveniste et al. (1990)

[10]Let g be a function defined on $\mathcal{Y}$. The Poisson's equation holds for g if there exists a finite function $\hat{g}$ such that $\hat{g}(y) = g(y) - \mathbb{E}_{y\sim d_{\mathcal{Y}}}[g(y)] + \int_{\mathcal{Y}} P(y, y')\hat{g}(y')dy'$ holds for any $y \in \mathcal{Y}$, where P denotes the transition kernel of $\{Y_n\}$. The drift condition (V4), together with some other mild conditions, is sufficient to ensure the existence of Poisson's equation. We refer the reader to Theorem 17.4.2 of Meyn and Tweedie (2012) for more details.

assume Poisson's equation directly without specifying sufficient conditions to establish Poisson's equation. Moreover, to establish stability, Benveniste et al. (1990) require a Lyapunov function for the ODE (9.13) that is always greater than or equal to $\alpha\|\cdot\|^2$ for some $\alpha > 0$ (Condition (ii) of Theorem 17 in Benveniste et al. (1990)). By contrast, our Assumption 5 does not put any restriction on the possible Lyapunov functions. We also note that Borkar et al. (2021) is also based on an error representation similar to Benveniste et al. (1990) enabled by Poisson's equation.

Linear H. If we further assume that the function $H(x, y)$ has a linear form, i.e.,

$$H(x, y) = A(y)x + b(y),$$

there are several other results regarding the stability (and thus convergence), e.g., Konda and Tsitsiklis (1999); Tadic (2001); Yu (2015) and Proposition 4.8 of Bertsekas and Tsitsiklis (1996). They, however, all require that the matrix $A \doteq \mathbb{E}_{y\sim d_{\mathcal{Y}}}[A(y)]$ is negative definite[11]. But contrast, our Assumption 5 only requires A to be Hurwitz[12] (see, e.g., Theorem 4.5 of Khalil (2002)), which is a weaker condition.[13] In Section 9.5, we provide a concrete RL algorithm where the corresponding A matrix is Hurwitz but not negative definite.

Local clock. Another approach to deal with Markovian noise $\{Y_n\}$ is to apply results in asynchronous schemes. We refer the reader to Chapter 7 of Borkar (2009) for details. The major limitation is that it requires count-based learning rates. At the n-th iteration, instead of using $\alpha(n)$, where n can be regarded as a "global lock", the asynchronous schemes use $\alpha(\varpi(n, Y_{n+1}))$ as the learning rate, where $\varpi(n, y)$ counts the number of visits to the state y until time n and can be regarded as a "local clock". The asynchronous schemes also have other assumptions regarding the local clock. Successful examples include Abounadi et al. (2001); Wan et al. (2021). However, we are not aware of any successful applications of such count-based learning rates in RL with function approximation, where an RL algorithm typically only has access to some feature $\phi(Y_n)$ instead of Y_n directly. Unless ϕ is a one-to-one mapping, there will be no way to count the state visitation.

[11] A real matrix A, not necessarily symmetric, is negative definite if and only if all the eigenvalues of the symmetric matrix $A + A^\top$ is strictly negative.

[12] A real matrix A is Hurwitz if and only if the real parts of all its eigenvalues are strictly negative.

[13] All negative definite matrices are Hurwitz, but many Hurwitz matrices are not negative definite. See Chapter 2 of Horn and Johnson (1991) for more details.

Other type of noise. The Borkar-Meyn theorem applies to only Martingale difference noise, which is, later on, relaxed to allow more types of noise, e.g., Bhatnagar (2011); Ramaswamy and Bhatnagar (2017). However, none of those extensions applies to general Markovian noise.

9.4 Main Proof

This section is dedicated to proving Theorem 19. Overall, we prove by contradiction. Section 9.4.1 sets up notations and establishes the desired diminishing asymptotic rate of change of a few functions. Section 9.4.2 establishes the desired equicontinuity. Section 9.4.3 assumes the opposite and thus identifies a subsequence of interest. Section 9.4.4 analyzes the property of the subsequence, helping the *reductio ad absurdum* in Section 9.4.5. Lemmas in this section are derived on an arbitrary sample path $\{x_0, \{Y_i\}_{i=1}^{\infty}\}$ such that the assumptions in Section 9.2 hold. Thus, we omit "$a.s.$" on the lemma statements for simplicity.

9.4.1 Diminishing Asymptotic Rate of Change

We divide the non-negative real axis $[0, \infty)$ into segments of length $\{\alpha(i)\}_{i=0,1,\dots}$. Those segments are then grouped into larger intervals $\{[T_n, T_{n+1})\}_{n=0,1,\dots}$. The sequence $\{T_n\}$ has the property that $T_{n+1} - T_n \approx T$ for some fixed T and as n tends to ∞, the error in this approximation diminishes. Precisely speaking, we define

$$\begin{aligned} t(0) \doteq& 0, \\ t(n) \doteq& \sum_{i=0}^{n-1} \alpha(i) \quad n = 1, 2, \dots \quad . \end{aligned}$$

For any $T > 0$, define

$$m(T) = \max \{i | T \geq t(i)\} \tag{9.14}$$

to be the largest i that has $t(i)$ smaller or equal to T. Intuitively, $t(m(T))$ is "just" left to T in the real axis. Then $t(m(T))$ has the follow properties:

$$t(m(T)) \leq T < t(m(T) + 1) = t(m(T)) + \alpha(m(T)), \tag{9.15}$$

$$t(m(T)) > T - \alpha(m(T)). \tag{9.16}$$

Define

$$\begin{aligned} T_0 &= 0, \\ T_{n+1} &= t(m(T_n + T) + 1). \end{aligned} \tag{9.17}$$

Intuitively, T_{n+1} is "just" right to $T_n + T$ in the real axis. For proving Theorem 19, it suffices to work with solutions of ODEs in only $[0, \infty)$. But for Corollary 1, it is necessary to consider solutions of ODEs in $(-\infty, \infty)$. To this end, we define

$$\begin{aligned} \alpha(i) =& 0 \quad \forall i < 0, \\ m(t) =& 0 \quad \forall t \leq 0, \end{aligned} \tag{9.18}$$

for simplifying notations. For any given function f with domain $\mathcal{Y}$, its asymptotic rate of change is defined as

$$\limsup_n \sup_{-\tau \leq t_1 \leq t_2 \leq \tau} \left\| \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i)[f(Y_{i+1}) - \mathbb{E}_{y \sim d_{\mathcal{Y}}}[f(y)]] \right\|.$$

The asymptotic rate of change characterizes the asymptotic regularity of the sequence $\{f(Y_n)\}$ and is a powerful tool to study stochastic approximation iterates. We refer the reader to Sections 5.3.2 and 6.2 of Kushner and Yin (2003) for an in-depth exposition of this tool. In the following, we demonstrate that the asymptotic rate of change is 0 for the functions in Assumption 6.

Lemma 22. *Let Assumptions 1, 2, and 4 hold. Let Assumption 6 or 6′ hold. Then the asymptotic rate of change of the functions (9.9), (9.10), and (9.11) is 0, i.e., for any fixed $\tau > 0$ and x, it holds that*

$$\limsup_n \sup_{-\tau \leq t_1 \leq t_2 \leq \tau} \left\| \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i) \left[H(x, Y_{i+1}) - h(x) \right] \right\| = 0 \quad a.s.,$$

$$\limsup_n \sup_{-\tau \leq t_1 \leq t_2 \leq \tau} \left\| \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i)[L_b(Y_{i+1}) - L_b] \right\| = 0 \quad a.s., \tag{9.19}$$

$$\limsup_n \sup_{-\tau \leq t_1 \leq t_2 \leq \tau} \left\| \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i)[L(Y_{i+1}) - L] \right\| = 0 \quad a.s. \tag{9.20}$$

Its proof is in Appendix F.4.1. Furthermore, the convergence of H_c to H_∞ in Assumption 3 demonstrates a similar pattern.

Lemma 23. *Let Assumptions 1, 2, 3, and 4 hold. Let Assumption 6 or 6′ hold. It then holds that*

$$\lim_{c \to \infty} \sup_{x \in \mathcal{B}} \sup_n \sup_{t \in [0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) \left[H_c(x, Y_{i+1}) - H_\infty(x, Y_{i+1}) \right] \right\| = 0 \quad a.s.,$$

where $\mathcal{B}$ denote an arbitrary compact set of $\mathbb{R}^d$.

Its proof is in Appendix F.2.1.

9.4.2 Equicontinuity of Scaled Iterates

Fix a sample path $\{x_0, \{Y_n\}\}$. Let $\bar{x}(t)$ be the piecewise constant interpolation [14] of x_n at points $\{t(n)\}_{n=0,1,\ldots}$, i.e.,

$$\bar{x}(t) \doteq \begin{cases} x_0 & t \in [0, t(1)) \\ x_1 & t \in [t(1), t(2)) \\ x_2 & t \in [t(2), t(3)) \\ \vdots \end{cases}$$

Using (9.14) to simplify it, we get

$$\bar{x}(t) \doteq x_{m(t)}. \tag{9.21}$$

Notably, $\bar{x}(t)$ is right continuous and has left limits. By (9.1), $\forall n \geq 0$, we have

$$\bar{x}(t(n+1)) = \bar{x}(t(n)) + \alpha(n) H(\bar{x}(t(n)), Y_{n+1}).$$

Now we scale $\bar{x}(t)$ in each segment $[T_n, T_{n+1})$.

Definition 1. *$\forall n \in \mathbb{N}, t \in [0, T)$, define*

$$\hat{x}(T_n + t) \doteq \frac{\bar{x}(T_n + t)}{r_n} \tag{9.22}$$

where

$$r_n \doteq \max\{1, \|\bar{x}(T_n)\|\}. \tag{9.23}$$

This implies

$$\forall n \in \mathbb{N}, \|\hat{x}(T_n)\| \leq 1. \tag{9.24}$$

Moreover[15], $\forall n \in \mathbb{N}, t \in [0, T)$,

$$\begin{aligned} \hat{x}(T_n + t) &= \frac{\bar{x}(T_n) + \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1})}{r_n}. \\ &= \hat{x}(T_n) + \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H_{r_n}(\hat{x}(t(i)), Y_{i+1}). \end{aligned}$$

[14] It also works if we consider a piecewise linear interpolation following Borkar (2009). The piecewise linear interpolation, however, will significantly complicate the presentation. We, therefore, follow Kushner and Yin (2003) to use piecewise constant interpolation.

[15] In this chapter, we use the convention that $\sum_{k=i}^{j} \alpha(k) = 0$ when $j < i$

The function $t \mapsto \hat{x}(T_n + t)$ is the scaled version of $\bar{x}(t)$ (by r_n) in the interval $[T_n, T_{n+1})$. Its domain is $[0, T_{n+1} - T_n)$. In most of the rest of this work, we will restrict it to $[0, T)$, such that the sequence of functions $\{t \mapsto \hat{x}(T_n + t)\}_{n=0,1,\dots}$ have the same domain $[0, T)$, which is crucial in applying the Arzela-Ascoli Theorem. The excess part $[T, T_{n+1} - T_n)$ diminishes asymptotically (cf. Lemma 49) and thus can be easily processed when necessary. Notably, $\hat{x}(T_n + t)$ can be regarded as the Euler's discretization of $z_n(t)$ defined below.

Definition 2. *$\forall n \in \mathbb{N}, t \in [0, T)$, define $z_n(t)$ as the solution of the ODE*

$$\frac{dz_n(t)}{dt} = h_{r_n}(z_n(t)) \tag{9.25}$$

with initial condition

$$z_n(0) = \hat{x}(T_n). \tag{9.26}$$

Apparently, $z_n(t)$ can also be written as

$$z_n(t) = \hat{x}(T_n) + \int_0^t h_{r_n}(z_n(s))ds. \tag{9.27}$$

Ideally, we would like to see that the error of Euler's discretization diminishes asymptotically. Precisely speaking, the discretization error is defined as

$$f_n(t) \doteq \hat{x}(T_n + t) - z_n(t) \tag{9.28}$$

and we would like that $f_n(t)$ diminishes to 0 as $n \to \infty$ in a certain sense. To this end, we study the following three sequences of functions

$$\{t \mapsto \hat{x}(T_n + t)\}_{n=0}^{\infty}, \{z_n(t)\}_{n=0}^{\infty}, \{f_n(t)\}_{n=0}^{\infty}. \tag{9.29}$$

In particular, we show that they are all equicontinuous in the extended sense. To understand equicontinuity in the extended sense, we first give the definition of equicontinuity.

Definition 3. *A sequence of functions $\left\{g_n : [0, T) \to \mathbb{R}^K\right\}$ is equicontinuous on $[0, T)$ if*
$\sup_n \|g_n(0)\| < \infty$ and $\forall \epsilon > 0$, $\exists \delta > 0$ such that

$$\sup_n \sup_{0 \le |t_1 - t_2| \le \delta, 0 \le t_1 \le t_2 < T} \|g_n(t_1) - g_n(t_2)\| \le \epsilon.$$

One example of equicontinuity is a sequence of bounded Lipschitz continuous functions with a common Lipschitz constant. Obviously, if $\{g_n\}$ is equicontinuous, each g_n must be continuous. However, the functions of interest in this work, i.e., $\hat{x}(T_n + t)$, $f_n(t)$, are not continuous so equicontinuity would not apply. We, therefore, introduce the following equicontinuity in the extended sense[16] akin to Kushner and Yin (2003).

Definition 4. *A sequence of functions* $\{g_n : [0, T) \to \mathbb{R}^K\}$ *is equicontinuous in the extended sense on* $[0, T)$ *if* $\sup_n \|g_n(0)\| < \infty$ *and* $\forall \epsilon > 0$, $\exists \delta > 0$ *such that*

$$\limsup_n \sup_{0 \le |t_1 - t_2| \le \delta, 0 \le t_1 \le t_2 < T} \|g_n(t_1) - g_n(t_2)\| \le \epsilon.$$

Notably, Kushner and Yin (2003) show that $\{t \in (-\infty, \infty) \mapsto \bar{x}(t(n) + t) \in \mathbb{R}^d\}_{n=0}^{\infty}$ is equicontinuous in the extended sense with *a priori* that

$$\sup_n \|x_n\| < \infty.$$

We do *not* have this *a priori*. Instead, we prove *a posteriori* that

$$\sup_{n \ge 0, t \in [0,T)} \|\hat{x}(T_n + t)\| < \infty$$

and show that $\{t \in [0, T) \mapsto \hat{x}(T_n + t) \in \mathbb{R}^d\}_{n=0}^{\infty}$ is equicontinuous in the extended sense. We remark that our function $t \mapsto \hat{x}(T_n + t)$ actually belongs to the J_1 Skorokhod topology (Skorokhod, 1956; Billingsley, 1999; Kern, 2023), although we will not work on this topology explicitly. Nevertheless, the following lemmas establish the desired equicontinuity, where Lemma 22 plays a key role.

Lemma 24. *The three sequences of functions* $\{\hat{x}(T_n + t)\}$, $\{z_n(t)\}$, *and* $\{f_n(t)\}$ *are all equicontinuous in the extended sense on* $t \in [0, T)$.

Its proof is in appendix F.2.2.

9.4.3 A Convergent Subsequence

According to the Arzela-Ascoli theorem in the extended sense (Theorem F.1.4), a sequence of equicontinuous functions always has a subsequence of functions that

[16] We must use this equicontinuity in the extended sense because we have chosen to use piecewise constant instead of piecewise linear interpolation. For piecewise linear interpolation, the standard equicontinuity is enough. However, as also argued in Kushner and Yin (2003), piecewise linear interpolation complicates the presentation much more than the equicontinuity in the extended sense.

uniformly converge to a continuous limit. In the following, we use this to identify a particular subsequence of interest.

We observe the following inequality

$$\forall n, \quad \left\|x_{m(T_n)}\right\| = \|\bar{x}(T_n)\| \leq r_n. \tag{9.30}$$

Thus, to prove Theorem 19, we first show

$$\sup_n r_n < \infty,$$

and which is implied by

$$\limsup_n r_n < \infty. \tag{9.31}$$

In the following, we aim to show (9.31) by contradiction. We first assume the opposite, i.e., $\limsup_n r_n = \infty$. Based on this assumption and applying Gronwall's inequality a few times, we can find a particular subsequence of interest, along which all the three sequences of functions in (9.29) converge uniformly.

Lemma 25. *Suppose* $\limsup_n r_n = \infty$. *Then there exists a subsequence* $\{n_k\}_{k=0}^{\infty} \subseteq \{0, 1, 2, \dots\}$ *that has the following properties:*

$$\begin{aligned} \lim_{k\to\infty} r_{n_k} &= \infty, \\ r_{n_k+1} &> r_{n_k} \quad \forall k. \end{aligned} \tag{9.32}$$

Moreover, there exist some continuous functions $f^{\lim}(t)$ *and* $\hat{x}^{\lim}(t)$ *such that* $\forall t \in [0, T)$,

$$\begin{aligned} \lim_{k\to\infty} f_{n_k}(t) &= f^{\lim}(t), \\ \lim_{k\to\infty} \hat{x}(T_{n_k} + t) &= \hat{x}^{\lim}(t), \end{aligned} \tag{9.33}$$

where both convergences are uniform in t *on* $[0, T)$. *Furthermore, let* $z^{\lim}(t)$ *denote the unique solution to the* (ODE@∞) *with the initial condition*

$$z^{\lim}(0) = \hat{x}^{\lim}(0),$$

in other words,

$$z^{\lim}(t) = \hat{x}^{\lim}(0) + \int_0^t h_\infty(z^{\lim}(s))ds. \tag{9.34}$$

Then $\forall t \in [0, T)$, *we have*

$$\lim_{k\to\infty} z_{n_k}(t) = z^{\lim}(t),$$

where the convergence is uniform in t *on* $[0, T)$.

Its proof is in Appendix F.2.3. We use the subsequence $\{n_k\}$ intensively in the remaining proofs.

9.4.4 Diminishing Discretization Error

Recall that $f_n(t)$ denotes the discretization error of $\hat{x}(T_n + t)$ of $z_n(t)$. We now proceed to prove that this discretization error diminishes along $\{n_k\}$. We note that we are able to improve over Borkar et al. (2021) because we only require the discretization error to diminish along the subsequence $\{n_k\}$, while Borkar et al. (2021) aim to show that the discretization error diminishes along the entire sequence $\{n\}$, which is unnecessary given (9.32).

In particular, we aim to prove that

$$\lim_{k\to\infty} \|f_{n_k}(t)\| = \left\|f^{\lim}(t)\right\| = 0.$$

This means $\hat{x}(T_{n_k} + t)$ is close to $z_{n_k}(t)$ as $k \to \infty$. For any $t \in [0, T)$, we have

$$\begin{aligned}
&\lim_{k\to\infty} \|f_{n_k}(t)\| \\
&= \lim_{k\to\infty} \left\| \hat{x}(T_{n_k}) + \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_k}}(\hat{x}(t(i)), Y_{i+1}) - z_{n_k}(t) \right\| && \text{(by (9.28))} \\
&= \lim_{k\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_k}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds \right\| && \text{(by (9.27))} \\
&\leq \lim_{k\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_k}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds \right\| \\
&\quad + \lim_{k\to\infty} \left\| \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds - \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds \right\|. && (9.35)
\end{aligned}$$

We now prove that the first term in the RHS of (9.35) is 0. Precisely speaking, we aim to prove $\forall t \in [0, T)$,

$$\lim_{k\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_k}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds \right\| = 0. \tag{9.36}$$

To compute the limit above, we first fix any $t \in [0, T)$ and compute the following stronger double limit, which implies the existence of the above limit (cf. Lemma 62).

$$\lim_{\substack{j\to\infty \\ k\to\infty}} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))ds \right\|. \tag{9.37}$$

To compute this double limit, we use the Moore-Osgood theorem (Theorem F.1.5) to make it iterated limits. To invoke the Moore-Osgood theorem, we first prove the uniform convergence in k when $j \to \infty$.

Lemma 26. $\forall t \in [0, T)$,

$$\lim_{j\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))ds \right\|$$
$$= \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{\infty}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{\infty}(\hat{x}^{\lim}(s))ds \right\|$$

uniformly in k.

Its proof is in Appendix F.2.4, where Lemma 23 plays a key role. Next, we prove, for each j, the convergence with $k \to \infty$.

Lemma 27. $\forall t \in [0, T)$, $\forall j$,

$$\lim_{k\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))ds \right\| = 0.$$

The proof of Lemma 27 follows the proof sketch of a similar problem on page 168 of Kushner and Yin (2003) with some minor changes and is the central averaging technique of Kushner and Yin (2003). We expect a reader familiar with Kushner and Yin (2003) should have belief in its correctness. We anyway still include all the details in the Appendix F.4.2 for completeness. We are now ready to compute the limit in (9.36).

Lemma 28. $\forall t \in [0, T)$,

$$\lim_{k\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_k}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds \right\| = 0.$$

Proof. It follows immediately from Lemmas 26 & 27, the Moore-Osgood theorem, and Lemma 62. □

Lemma 28 confirms that the first term in the RHS of (9.35) is 0. Moreover, it also enables us to rewrite $\hat{x}^{\lim}(t)$ from a summation form to an integral form.

$$\begin{aligned}
&\hat{x}^{\lim}(t) \\
=&\lim_{k\to\infty} \hat{x}(T_{n_k}) + \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_k}}(\hat{x}(t(i)), Y_{i+1}) \\
=&\lim_{k\to\infty} \hat{x}(T_{n_k}) + \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds. \qquad \text{(by Lemma 28)(9.38)}
\end{aligned}$$

This, together with a few Gronwall's inequality arguments, confirms that the discretization error indeed diminishes along $\{n_k\}$.

Lemma 29. $\forall t \in [0, T)$,

$$\lim_{k\to\infty} \|f_{n_k}(t)\| = 0.$$

Its proof is in Appendix F.2.6.

9.4.5 Identifying Contradiction and Completing Proof

Having made sure that the error of the discretization $\hat{x}(T_n + t)$ of $z_n(t)$ diminishes along $\{n_k\}$, we now study the behavior $\hat{x}(T_{n_k} + t)$ through $z_{n_k}(t)$ and identify a contradiction. The underlying idea is identical to Borkar (2009). However, the execution is different so we cannot use the arguments from Borkar (2009) directly. Namely, to use the arguments in Chapter 3 of Borkar (2009) directly, we have to prove that the discretization error diminishes along the entire sequence. This is impossible for us because the Arzela-Ascoli theorem only guarantees convergence along the subsequence $\{n_k\}$. Nevertheless, after carefully choosing the subsequence in Lemma 25, we are still able to execute the contradiction idea as documented below.

Lemma 30. *Suppose* $\limsup_n r_n = \infty$. *Then there exists a* k_0 *such that*

$$r_{n_{k_0}+1} \leq r_{n_{k_0}}.$$

Its proof is in Appendix F.2.7. This lemma constructs a contradiction to (9.32). This means the proposition $\limsup_n r_n = \infty$ is impossible. This completes the proof of

$$\sup_n r_n < \infty. \tag{9.39}$$

By decomposition,

$$\begin{aligned}
&\sup_n \|x_n\| \\
=&\sup_n \sup_{i\in\{i|m(T_n)\le m(T_n)+i<m(T_{n+1})\}} \left\|x_{m(T_n)+i}\right\| - \left\|x_{m(T_n)}\right\| + \left\|x_{m(T_n)}\right\| \\
\le&\sup_n \sup_{i\in\{i|m(T_n)\le m(T_n)+i<m(T_{n+1})\}} \left\|x_{m(T_n)+i}\right\| - \left\|x_{m(T_n)}\right\| + \sup_n r_n. \text{ (by (9.30))(9.40)}
\end{aligned}$$

We show the first term above is also bounded.

Lemma 31.

$$\sup_n \sup_{i\in\{i|m(T_n)\le m(T_n)+i<m(T_{n+1})\}} \left\|x_{m(T_n)+i}\right\| - \left\|x_{m(T_n)}\right\| < \infty.$$

Its proof is in Appendix F.2.8. Thus, (9.39), (9.40) and Lemma 31 conclude Theorem 19.

9.5 Applications in Reinforcement Learning

In this section, we discuss broad applications of Corollary 1 in RL. In particular, we both demonstrate state-of-the-art analysis in Section 9.5.3 and greatly simplify existing analysis in Section 9.5.4. We first introduce notations and lay out the background of RL.

All vectors are column vectors. For a vector $d \in \mathbb{R}^N$ with strictly positive entries, we use $\|x\|_d$ to denote the d-weighted ℓ_2 norm, i.e., $\|x\|_d \doteq \sqrt{\sum_{i=1}^N d_i x_i^2}$. We also abuse $\|\cdot\|_d$ to denote the corresponding induced matrix norm. We use $\|\cdot\|$ to denote a general norm that respects sub-multiplicity. We use vectors and functions interchangeably when it does not confuse. For example, for some $g : \mathcal{S} \to \mathbb{R}$, we also interpret g as a vector in $\mathbb{R}^{|\mathcal{S}|}$. We use $\Pi_{\Phi,d}$ to denote a projection operator that projects a vector d to the column space of a matrix Φ, assuming Φ has a full column rank. In other words,

$$\Pi_{\Phi,d} v = \Phi \arg\min_\theta \|\Phi\theta - v\|_d^2.$$

When it is clear from the context, we write $\Pi_{\Phi,d}$ as Π_d for simplifying presentation.

We consider an MDP with a finite state space[17] $\mathcal{S}$, a finite action space $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a transition function $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, an initial distribution $p_0 : \mathcal{S} \to [0, 1]$, and a discount factor $\gamma \in [0, 1)$. At time step 0, an initial

[17]It is worth mentioning that even if the MDP problem itself is finite, the Markov chains used to analyze many RL algorithms still evolve in an uncountable and unbounded space. This will be seen shortly.

state S_0 is sampled from p_0. At time t, given the state S_t, the agent samples an action $A_t \sim \pi(\cdot|S_t)$, where $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the policy being followed by the agent. A reward $R_{t+1} \doteq r(S_t, A_t)$ is then emitted and the agent proceeds to a successor state $S_{t+1} \sim p(\cdot|S_t, A_t)$. The return at time t is defined as $G_t \doteq \sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i}$, using which we define the state-value function $v_\pi(s)$ and action-value function $q_\pi(s)$ as

$$\begin{aligned} v_\pi(s) \doteq& \mathbb{E}_{\pi,p}\left[G_t|S_t = s\right], \\ q_\pi(s,a) \doteq& \mathbb{E}_{\pi,p}\left[G_t|S_t = s, A_t = a\right]. \end{aligned}$$

The value function v_π is the unique fixed point of the Bellman operator

$$\mathcal{T}_\pi v \doteq r_\pi + \gamma P_\pi v,$$

where $r_\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the reward vector induced by the policy π, i.e., $r_\pi(s) \doteq \sum_a \pi(a|s) r(s,a)$, and $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is the transition matrix induced by the policy π, i.e., $P_\pi(s,s') \doteq \pi(a|s)p(s'|s,a)$. With a $\lambda \in [0,1]$, we can rewrite $v_\pi = \mathcal{T}_\pi v_\pi$ using the identity $v_\pi = (1-\lambda)v_\pi + \lambda \mathcal{T}_\pi v_\pi$ as

$$\begin{aligned} v_\pi =& r_\pi + \gamma P_\pi((1-\lambda)v_\pi + \lambda \mathcal{T}_\pi v_\pi) \\ =& r_\pi + \gamma(1-\lambda)P_\pi v_\pi + \gamma\lambda P_\pi(r_\pi + \gamma P_\pi v_\pi) \\ =& r_\pi + \gamma\lambda P_\pi r_\pi + \gamma(1-\lambda)P_\pi v_\pi + \gamma^2 \lambda P_\pi^2((1-\lambda)v_\pi + \lambda \mathcal{T}_\pi v_\pi) \\ =& \dots \\ =& \sum_{i=0}^{\infty} (\gamma\lambda P_\pi)^i r_\pi + (1-\lambda)\sum_{i=1}^{\infty} \lambda^{i-1}\gamma^i P_\pi^i v_\pi, \\ =& (I - \gamma\lambda P_\pi)^{-1} r_\pi + (1-\lambda)\gamma(I-\gamma\lambda P_\pi)^{-1} P_\pi v_\pi. \end{aligned}$$

This suggests that we define a λ-Bellman operator as

$$\mathcal{T}_{\pi,\lambda} v \doteq r_{\pi,\lambda} + \gamma P_{\pi,\lambda} v,$$

where $r_{\pi,\lambda} \doteq (I - \gamma\lambda P_\pi)^{-1} r_\pi$, $P_{\pi,\lambda} \doteq (1-\lambda)(I-\gamma\lambda P_\pi)^{-1} P_\pi$. It is then easy to see that when $\lambda = 0$, $\mathcal{T}_{\pi,\lambda}$ reduces to $\mathcal{T}_\pi$. When $\lambda = 1$, $\mathcal{T}_{\pi,\lambda}$ reduces to a constant function that always output $(I - \gamma P_\pi)^{-1} r_\pi$. It is proved that $\mathcal{T}_{\pi,\lambda}$ is a $\frac{\gamma(1-\lambda)}{1-\gamma\lambda}$-contraction w.r.t. $\|\cdot\|_{d_\pi}$ (see, e.g., Lemma 6.6 of (Bertsekas and Tsitsiklis, 1996)), where we use $d_\pi \in \mathbb{R}^{|\mathcal{S}|}$ to denote the stationary distribution of the Markov chain induced by π. Obviously, v_π is the unique fixed point of $\mathcal{T}_{\pi,\lambda}$.

One fundamental task in RL is prediction, i.e., to estimate v_π, for which temporal difference (TD, Sutton (1988)) learning is the most powerful method. In particular,

Sutton (1988) considers a linear architecture. Let $\phi : \mathcal{S} \to \mathbb{R}^K$ be the feature function that maps a state to a K-dimensional feature. Linear TD(λ) (Sutton, 1988) aims to find a $\theta \in \mathbb{R}^K$ such that $\phi(s)^\top \theta$ is close to $v_\pi(s)$ for every $s \in \mathcal{S}$. To this end, linear TD(λ) updates θ recursively as

$$\begin{aligned} e_t =& \lambda\gamma e_{t-1} + \phi_t, \\ \theta_{t+1} =& \theta_t + \alpha_t \left(R_{t+1} + \gamma\phi_{t+1}^\top \theta_t - \phi_t^\top \theta_t\right) e_t, \end{aligned} \tag{9.41}$$

where we have used $\phi_t \doteq \phi(S_t)$ as shorthand and $e_t \in \mathbb{R}^K$ is the *eligiblity trace* with an arbitrary initial e_{-1}. We use $\Phi \in \mathbb{R}^{|\mathcal{S}| \times K}$ to denote the feature matrix, each row of which is $\phi(s)^\top$. It is proved (Tsitsiklis and Roy, 1996) that, under some conditions, $\{\theta_t\}$ converges to the unique zero of $J_{\text{on}}(\theta) \doteq \|\Pi_{d_\pi} \mathcal{T}_{\pi,\lambda} \Phi\theta - \Phi\theta\|_{d_\pi}^2$. This $J_{\text{on}}(\theta)$ is referred to as the on-policy mean squared projected Bellman error (MSPBE).

In many scenarios, due to the concerns of data efficiency (Lin, 1992; Sutton et al., 2011) or safety (Dulac-Arnold et al., 2019), we would like to estimate v_π but select actions using a different policy, called μ. This is off-policy learning, where π is called the target policy and μ is called the behaivor policy. In the rest of this section, we always consider the off-policy setting, i.e., the action A_t is sampled from $\mu(\cdot|S_t)$. Correspondingly, off-policy linear TD(λ) updates θ recursively as

$$\begin{aligned} e_t =& \lambda\gamma\rho_{t-1} e_{t-1} + \phi_t, \\ \theta_{t+1} =& \theta_t + \alpha_t\rho_t \left(R_{t+1} + \gamma\phi_{t+1}^\top \theta_t - \phi_t^\top \theta_t\right) e_t, \end{aligned} \tag{9.42}$$

where $\rho_t \doteq \rho(S_t, A_t) \doteq \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$ is the importance sampling ratio to account for the discrepancy in action selection between π and μ. Obviously, if $\pi = \mu$, then (9.42) reduces to (9.41). Let $d_\mu \in \mathbb{R}^{|\mathcal{S}|}$ be the stationary distribution of the Markov chain induced by μ. If $\{\theta_t\}$ in (9.42) converged, it would converge to the unique zero of

$$J_{\text{off}}(\theta) \doteq \left\|\Pi_{d_\mu} \mathcal{T}_{\pi,\lambda} \Phi\theta - \Phi\theta\right\|_{d_\mu}^2,$$

which is the off-policy MSPBE.

9.5.1 Eligibility Trace

The eligibility trace is one of the most fundamental ingredients in RL and is deeply rooted in RL since the very beginning of RL (Klopf, 1972; Sutton, 1978; Barto and Sutton, 1981a,b; Barto et al., 1983; Sutton, 1984). The eligibility trace in (9.41) is called the accumulating trace, first introduced in Barto and Sutton (1981a). Later on, this trace is also used in control by Rummery and Niranjan (1994). Its off-policy

version in (9.42) is introduced by Precup et al. (2000b, 2001) and further developed by Bertsekas and Yu (2009); Yu (2012). Other forms of traces include the Dutch trace introduced by Seijen and Sutton (2014) and the followon trace introduced by Sutton et al. (2016). In short, traces are usually used to accelerate credit assignment, which is a fundamental challenge in RL. Intuitively, traces are able to achieve this goal because they function as memory of the past. Empirically, RL algorithms with traces usually outperform those without traces (Sutton and Barto, 2018). Traces are also important in establishing the equivalence between backward and forward views of RL algorithms (Sutton et al., 2014).

Despite the superiority of traces in multiple aspects, they usually complicate the analysis of RL algorithms. Without any trace, to analyze an RL algorithm it is usually sufficient to consider the Markov chain $\{(S_t, A_t)\}$. Under a finite MDP assumption, this augmented Markov chain is still finite. Once trace is introduced, we, however, must consider the Markov chain $\{(S_t, A_t, e_t)\}$, see, e.g., Tsitsiklis and Roy (1996). This augmented Markov chain now immediately evolves in an uncountable space $\mathcal{S} \times \mathcal{A} \times \mathbb{R}^d$. In the on-policy case (cf. (9.41)), this is still managable. It is clear from (9.41) that e_t remains bounded almost surely. So the augmented Markov chain evolves in a compact space. In the off-policy case (cf. (9.42)), the trace e_t can easily be unbounded almost surely due to the importance sampling ratio ρ_{t-1} (Yu, 2012). The augmented Markov chain then evolves in an *unbounded and uncountable* space. Even worse, sometimes the second moment of e_t can also be unbounded (Yu, 2012), further complicating the analysis. Despite that e_t is demonstrated to obey a form of the strong law of large numbers (Yu, 2012), there does not exist a general tool to make use of this in convergence analysis before this work. In other words, this work is the first to provide a general tool to analyze the stability (and thus convergence) of RL algorithms with off-policy traces.

9.5.2 The Deadly Triad

Despite the aforementioned superiority of off-policy learning in safety and data efficiency, it complicates RL algorithms in at least two aspects. The first is that it makes traces extremely hard to analyze, as demonstrated in the section above. Second, it makes the RL algorithm behaves poorly in expectation. In other words, even if there is no noise (cf. replacing $H(x_n, Y_{n+1})$ with $h(x_n)$), the RL algorithm can still behave poorly. A concrete example is that, for a general λ, the iterates $\{\theta_t\}$ in (9.42) can possibly diverge to infinity, as documented in Baird (1995); Tsitsiklis and Roy (1996); Sutton and Barto (2018). This is the notorious *deadly triad*, which

refers to the instability of an RL algorithm when it combines bootstrapping, function approximation, and off-policy learning simultaneously while maintaining a constant $\mathcal{O}(K)$ computational complexity each step.

The deadly triad has been one of the central challenges of RL in the past three decades and numerous works have been done in this topic (Precup et al., 2000b, 2001; Sutton et al., 2008b, 2009; Maei et al., 2009, 2010; Maei and Sutton, 2010; Maei, 2011; Sutton et al., 2011; Yu, 2012; Mahadevan et al., 2014; Liu et al., 2015; Yu, 2015; White and White, 2016; Mahmood et al., 2017; Yu, 2017; Wang et al., 2017; Touati et al., 2018; Liu et al., 2018; Zhang et al., 2020b; Nachum et al., 2019; Xu et al., 2019; Zhang et al., 2021a, 2020c; Ghiassian et al., 2020; Wang and Zou, 2020; Zhang et al., 2020a; Guan et al., 2021; Zhang et al., 2021b; Zhang and Whiteson, 2022; Qian and Zhang, 2025; Liu et al., 2025d). We refer the reader to Chapter 11 of Sutton and Barto (2018) and Zhang (2022) for more detailed exposition.

Among all those works, gradient temporal difference learning (GTD, Sutton et al. (2008b)) and emphatic temporal difference learning (ETD, Sutton et al. (2016)) are the two most important solutions to the deadly triad in terms of policy evaluation. GTD and ETD are also important building blocks for other algorithms. They can be used in convergent off-policy actor-critic algorithms for control, see, e.g., Imani et al. (2018); Maei (2018); Zhang et al. (2020b); Xu et al. (2021); Graves et al. (2023). They can also be used to learn value functions w.r.t. some augmented reward function to construct behavior policies for efficient and unbiased Monte Carlo policy evaluation, see, e.g., Liu and Zhang (2024); Liu et al. (2025c); Chen et al. (2025); Liu et al. (2025a). But surprisingly, the convergence analysis of their ultimate form with eligibility trace, i.e., GTD(λ) and ETD(λ), is still not fully settled down. In the next, we shall analyze GTD(λ) and ETD(λ) in the sequel. Throughout the rest of Section 9.5, we make the following assumptions.

Assumption 9.5.1. *Both $\mathcal{S}$ and $\mathcal{A}$ are finite. The Markov chain $\{S_t\}$ induced by the behavior policy μ is irreducible. And $\mu(a|s) > 0$ for all s, a.*

We note again that in light of Section 9.5.1, even if the MDP itself is finite, the augmented Markov chain used to analyze GTD(λ) and ETD(λ) still evolves in an unbounded and uncountable space. The analysis is, therefore, very challenging. Assumption 9.5.1 is a standard assumption in off-policy RL to ensure enough exploration, see, e.g., Precup et al. (2001); Sutton et al. (2016). The condition $\mu(a|s) > 0$ can be easily relaxed to $\pi(a|s) > 0 \implies \mu(a|s) > 0$, at the price of complicating the presentation.

Assumption 9.5.2. *The learning rates* $\{\alpha_t\}$ *have the form* $\alpha_t = \frac{B_1}{t+B_2}$.

Assumption 9.5.2 is also used in existing works, see, e.g., Yu (2012, 2015, 2017).

Assumption 9.5.3. *The feature matrix* Φ *has a full column rank.*

Assumption 9.5.3 is a standard assumption in RL with linear function approximation to ensure the existence and uniqueness of the solution, see, e.g., Tsitsiklis and Roy (1996).

9.5.3 Gradient Temporal Difference Learning

The idea of GTD is to perform stochastic gradient descent on $J_{\text{off}}(\theta)$ directly and use a weight duplication trick or Fenchel's duality to address a double sampling issue in estimating $\nabla J_{\text{off}}(\theta)$. We provide the key concepts of GTD in Section 2.10, and we re-elaborate on them in this section to facilitate reading. We refer the reader to Sutton et al. (2009); Liu et al. (2015) for detailed derivation. GTD has many different variants, see, e.g., Sutton et al. (2008b, 2009); Maei (2011); Yu (2017); Zhang et al. (2021a); Qian and Zhang (2025). In this chapter, we present and analyze the following arguably most representative one, referred to as GTD(λ) for simplicity.[18] In particular, GTD(λ) employs an additional weight vector $\nu \in \mathbb{R}^K$ and update θ and ν simultaneously in a recursive way as

$$\begin{aligned} e_t =& \lambda\gamma\rho_{t-1}e_{t-1} + \phi_t, \\ \delta_t =& R_{t+1} + \gamma\phi_{t+1}^\top\theta_t - \phi_t^\top\theta_t, \\ \nu_{t+1} =& \nu_t + \alpha_t\left(\rho_t\delta_t e_t - \phi_t\phi_t^\top\nu_t\right), \\ \theta_{t+1} =& \theta_t + \alpha_t\rho_t(\phi_t - \gamma\phi_{t+1})e_t^\top\nu_t. \end{aligned} \tag{9.43}$$

This additional weight vector results from the weight duplication or Fenchel's duality. To analyze (9.43), we first express the update to ν and θ in a compact form as

$$\begin{bmatrix}\nu_{t+1}\\ \theta_{t+1}\end{bmatrix} = \begin{bmatrix}\nu_t\\ \theta_t\end{bmatrix} + \alpha_t\left(\begin{bmatrix}-\phi_t\phi_t^\top & \rho_t e_t(\gamma\phi_{t+1}-\phi_t)^\top\\ -(\gamma\phi_{t+1}-\phi_t)\rho_t e_t^\top & 0\end{bmatrix}\begin{bmatrix}\nu_t\\ \theta_t\end{bmatrix} + \begin{bmatrix}\rho_t R_{t+1}e_t\\ 0\end{bmatrix}\right).$$

To further simplify it, we define an augmented Markov chain $\{Y_t\}$ as

$$Y_{t+1} \doteq (S_t, A_t, S_{t+1}, e_t), \quad t = 0, 1, \ldots.$$

[18] This is the GTDa in Yu (2017) and is the GTD2 in Sutton et al. (2009) with eligibility trace.

We also define shorthands

$$\begin{aligned} x \doteq& \begin{bmatrix} \nu \\ \theta \end{bmatrix}, x_t \doteq \begin{bmatrix} \nu_t \\ \theta_t \end{bmatrix}, \\ y \doteq& (s, a, s', e), \\ A(y) \doteq& \rho(s,a) e (\gamma \phi(s') - \phi(s))^\top, \\ b(y) \doteq& \rho(s,a) r(s,a) e, \\ C(y) \doteq& \phi(s) \phi(s)^\top, \\ H(x,y) \doteq& \begin{bmatrix} -C(y) & A(y) \\ -A(y)^\top & 0 \end{bmatrix} x + \begin{bmatrix} b(y) \\ 0 \end{bmatrix}. \end{aligned}$$

Then GTD(λ) can be expressed as

$$x_{t+1} = x_t + \alpha_t H(x_t, Y_{t+1}),$$

which reduces to the form of (9.1). We now proceed to prove the almost sure convergence of $\{x_t\}$ using Corollary 1. Apparently, $\{Y_t\}$ evolves in the state space

$$\mathcal{Y} \doteq \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R}^K.$$

Despite that both $\mathcal{S}$ and $\mathcal{A}$ are finite, $\mathcal{Y}$ can still be unbounded and uncountable. It is shown in Proposition 3.1 of Yu (2012) that as long as there is a cycle in $\{S_t\}$, e_t is unbounded almost surely in arguably almost all natural problems. Nevertheless, Yu (2012) shows that $\{Y_t\}$ has the following property.

Lemma 32. *(Theorems 3.2 & 3.3 of Yu (2012)) Let Assumption 9.5.1 hold. Then*

(i) $\{Y_t\}$ *has a unique invariant probability measure, referred to as* $d_\mathcal{Y}$.

(ii) For any matrix/vector-valued function $g(s, a, s', e)$ *on* $\mathcal{Y}$ *which is Lipschitz continuous in* e *with a Lipschitz constant* L_g*, i.e.,*

$$\|g(s,a,s',e) - g(s,a,s',e')\| \le L_g \|e - e'\|, \quad \forall s, a, s', e, e',$$

the expectation $\mathbb{E}_{y \sim d_\mathcal{Y}}[g(y)]$ *exists and is finite, and the* (LLN) *holds for the* g *function.*

Yu (2012) also shows that

$$\begin{aligned} A \doteq& \mathbb{E}_{y \sim d_\mathcal{Y}}[A(y)] = \Phi^\top D_\mu (\gamma P_{\pi,\lambda} - I) \Phi, \\ b \doteq& \mathbb{E}_{y \sim d_\mathcal{Y}}[b(y)] = \Phi^\top D_\mu r_{\pi,\lambda}, \\ C \doteq& \mathbb{E}_{y \sim d_\mathcal{Y}}[C(y)] = \Phi^\top D_\mu \Phi, \end{aligned}$$

where we use D_μ to denote the diagonal matrix whose diagonal entry is d_μ.

Theorem 20. *Let Assumptions 9.5.1 - 9.5.3 hold. Assume A is nonsingular. Then the iterates $\{\theta_t\}$ generated by GTD(λ) (9.43) satisfy*

$$\lim_{t\to\infty} \theta_t = -A^{-1}b \quad a.s.$$

Its proof is in Appendix F.2.10. It can be shown easily that $-A^{-1}b$ is the unique zero of $J_{\text{off}}(\theta)$, see, e.g., Sutton et al. (2009). Notably, Theorem 20 is the first almost sure convergence analysis of GTD with eligibility trace without adding additional bias terms. Most existing convergence analyses of GTD (see, e.g., Sutton et al. (2008b, 2009); Maei (2011); Liu et al. (2015); Wang et al. (2017); Qian and Zhang (2025)) do not have eligibility trace. To our knowledge, the only previous analysis of GTD with eligibility trace is Yu (2017), which, however, relies on additional projection operators or regularization to ensure the stability and unavoidably introduces bias into the final limiting point. As a result, Yu (2017) cannot establish the almost sure convergence of GTD(λ) to the unique zero of $J_{\text{off}}(\theta)$. Yu (2017) also introduces extensions to λ. Instead of being a constant, it can be a state-dependent function $\lambda : \mathcal{S} \to [0, 1]$. The almost sure convergence of GTD(λ) with a state-dependent λ function follows similarly. We present the simplest constant λ case for clarity. Yu (2017) also introduces history-dependent λ function, which we leave for future work.

9.5.4 Emphatic Temporal Difference Learning

The idea of ETD is to reweight the off-policy linear TD update (9.42) by an additional factor. Similar to GTD, ETD also has many different variants, see, e.g., Yu (2015); Sutton et al. (2016); Hallak et al. (2016); Zhang et al. (2020b); Zhang and Whiteson (2022); Guan et al. (2021). Variants of ETD have also been applied in deep RL, see, e.g., Jiang et al. (2021, 2022); Mathieu et al. (2023).

We introduce the key concepts of ETD in Section 2.11, and we re-elaborate on them here to facilitate reading. In this section, we consider the original ETD(λ) in Yu (2015); Sutton et al. (2016). ETD(λ) updates θ recursively in the following way

$$\begin{aligned} F_t =& \gamma\rho_{t-1}F_{t-1} + i(S_t), \\ M_t =& \lambda i(S_t) + (1-\lambda)F_t, \\ e_t =& \lambda\gamma\rho_{t-1}e_{t-1} + M_t\phi_t, \\ \theta_{t+1} =& \theta_t + \alpha_t\rho_t\left(R_{t+1} + \gamma\phi_{t+1}^\top\theta_t - \phi_t^\top\theta_t\right)e_t, \end{aligned} \tag{9.44}$$

where $i : \mathcal{S} \to (0, \infty)$ is an arbitrary "interest" function (Sutton et al., 2016), specifying user's preference for different states, despite that in most applications, $i(s)$ is a constant

function which is always 1. See Zhang et al. (2019) for an example where the interest function is not trivially 1. Comparing the eligibility trace e_t in (9.44) with that in (9.42), one can find that there is an additional scalar multiplier M_t proceeding ϕ_t. This M_t is called "emphasis" (Sutton et al., 2016), which is the accumulation of F_t, called "followon trace" (Sutton et al., 2016). We refer the reader to Sutton et al. (2016) for the intuition behind ETD. Nevertheless, Yu (2015) proves that, under mild conditions, $\{\theta_t\}$ in (9.44) converges almost surely to the unique zero of

$$J_{\text{emphatic}}(\theta) = \|\Pi_m \mathcal{T}_{\pi,\lambda}\Phi\theta - \Phi\theta\|_m^2,$$

where $m \doteq (I - \gamma P_{\pi,\lambda}^\top)^{-1} D_\mu i$. We remark that the zero of $J_{\text{emphatic}}(\theta)$ has better theoretical guarantees than the zero of $J_{\text{off}}(\theta)$ in terms of the approximation error for v_π (Hallak et al., 2016). ETD, however, usually suffers from a larger variance than GTD (Sutton and Barto, 2018).

To analyze ETD(λ), Yu (2015) considers the following augmented Markov chain

$$Y_{t+1} = (S_t, A_t, S_{t+1}, e_t, F_t).$$

Again, $\{Y_t\}$ behaves poorly in that (e_t, F_t) can be unbounded almost surely and its variance can grow to infinity as time progresses. We refer the reader to Remark A.1 in Yu (2015) for an in-depth discussion regarding this poor behavior. Nevertheless, Yu (2015) shows that $\{Y_t\}$ has the following property.

Lemma 33. *(Theorems 3.2 & 3.3 of Yu (2015)) Let Assumption 9.5.1 hold. Then*

(i) *$\{Y_t\}$ has a unique invariant probability measure, referred to as $d_\mathcal{Y}$.*

(ii) *For any matrix / vector-valued function $g(s, a, s', e, f)$ on $\mathcal{Y}$ which is Lipschitz continuous in (e, f) with a Lipschitz constant L_g, i.e.,*

$$\|g(s, a, s', e, f) - g(s, a, s', e', f')\| \le L_g \|e - e'\|, \quad \forall s, a, s', e, e', f, f',$$

the expectation $\mathbb{E}_{y\sim d_\mathcal{Y}}[g(y)]$ exists and is finite, and the (LLN) holds for the function g.

We now discuss how Yu (2015) establishes the almost sure convergence of $\{\theta_t\}$. First, we define shorthands

$$\begin{aligned} y &\doteq (s, a, s', e, f), \\ A(y) &= \rho(s, a) e (\gamma\phi(s') - \phi(s))^\top, \\ b(y) &= \rho(s, a) r(s, a) e, \\ H(\theta, y) &= A(y)\theta + b(y). \end{aligned}$$

Then the ETD(λ) update can be expressed as

$$\theta_{t+1} = \theta_t + \alpha_t H(\theta_t, Y_{t+1}).$$

Yu (2015) also shows that

$$\begin{aligned} A \doteq &\mathbb{E}_{y\sim d_{\mathcal{Y}}}\left[A(y)\right] = \Phi^\top D_m(\gamma P_{\pi,\lambda} - I)\Phi, \\ b \doteq &\mathbb{E}_{y\sim d_{\mathcal{Y}}}\left[b(y)\right] = \Phi^\top D_m r_{\pi,\lambda}, \end{aligned}$$

and $-A^{-1}b$ is the unique zero of $J_{\text{emphatic}}(\theta)$. Despite that A is negative definite (see, e.g., Section 4 of Sutton et al. (2016)) and the corresponding ODE@∞ is, therefore, globally asymptotically stable, Yu (2015) is not able to establish the stability of $\{\theta_t\}$ directly, simply because the results in the stochastic approximation community are not ready yet. See Section 9.3 for a comprehensive review. As a workaround, Yu (2015) analyzes a constrained variant of ETD(λ) first:

$$\theta'_{t+1} = \Pi\left(\theta'_t + \alpha_t H(\theta'_t, Y_{t+1})\right),$$

where Π is a projection to a centered ball of properly chosen radius w.r.t. ℓ_2 norm. Yu (2015) then proves that the difference between $\{\theta_t\}$ and $\{\theta'_t\}$ diminishes almost surely and therefore establishes the convergence of $\{\theta_t\}$ indirectly. To establish the convergence of $\{\theta'_t\}$, Yu (2015) invokes Theorem 1.1 in Chapter 6 of Kushner and Yin (2003). Now with our Corollary 1, the same arguments Yu (2015) use to invoke Kushner and Yin (2003) can lead to the convergence of $\{\theta_t\}$ directly. Our contribution is, therefore, a greatly simplified almost sure convergence analysis of ETD(λ). In particular, we have

Theorem 21. *Let Assumptions 9.5.1 - 9.5.3 hold. Then the iterates $\{\theta_t\}$ generated by ETD(λ) (9.44) satisfy*

$$\lim_{t\to\infty} \theta_t = -A^{-1}b \quad a.s.$$

The proof of Theorem 21 is a verbatim repetition of the proof of Theorem 20 in Appendix F.2.10 after noticing that A is negative definite and Lemma 33 and is thus omitted. Notably, this proof does not involve the comparison between $\{\theta_t\}$ and $\{\theta'_t\}$.

We remark that the comparison technique between $\{\theta_t\}$ and $\{\theta'_t\}$ used by Yu (2015) heavily relies on the fact that A is negative definite (see Lemma 4.1 of Yu (2015)). But in GTD(λ), the corresponding matrix is $\begin{bmatrix} -C & A \\ -A^\top & 0 \end{bmatrix}$, which is Hurwitz but not negative definite. In fact, it is only negative semidefinite. As a result, the comparison technique in Yu (2015) does not apply to GTD(λ).

9.6 Discussion

In this chapter, we develop a novel stability result of stochastic approximations, extending the celebrated Borkar-Meyn theorem from the Martingale difference noise setting to the Markovian noise setting. Our result is built on the diminishing asymptotic rate of change of a few functions, which is implied by both a form of the strong law of larger numbers and a form of the law of the iterated logarithm. We demonstrate the wide applicability of our results in RL, generating state-of-the-art analysis for important RL algorithms in breaking the notorious deadly triad. There are many possible directions for future work. One direction is to characterize the behavior of the iterates in (9.1) in more aspects. For example, it is possible to establish a (functional) central limit theorem following Borkar et al. (2021). It is also possible to establish an almost sure convergence rate, a high probability concentration bound, and an L^p convergence rate following Qian et al. (2024). Another direction is to weaken the required assumptions further. In the context of RL, Assumption 5 is typically obtained by assuming h is related to some contraction operator and the feature matrix Φ has a full column rank. It is possible to weaken h to nonexpansive operators following Blaser and Zhang (2024). It is also possible to allow Φ to have arbitrary ranks following Wang and Zhang (2024).

Chapter 10

Conclusion

Policy evaluation remains one of the most fundamental challenges in reinforcement learning, particularly when aiming for both efficiency and reliability in high-stakes environments. In this thesis, we focused on improving the efficiency, robustness, and theoretical understanding of policy evaluation. While we made substantial progress—achieving significant improvements over prior methods—many important questions remain open for future investigation.

One possible line of future work is to extend our variance reduction methods beyond Monte Carlo evaluation. Temporal difference (TD) learning is a widely used alternative that requires less data but suffers from stability issues, particularly in off-policy settings. Adapting our optimal data collection and processing frameworks to TD-based objectives could lead to more sample-efficient algorithms.

Another important direction is to extend our methods to multi-agent settings. Currently, our methods are developed for single-agent environments, but many real-world applications—such as autonomous driving, smart grids, and financial markets—involve multiple interacting agents whose policies evolve over time. Extending our variance reduction and safety frameworks to multi-agent systems presents new challenges, particularly due to the non-stationarity introduced by the changing behaviors of other agents. This may require developing new approaches that account for strategic interactions, as well as designing data collection strategies that remain informative in the presence of such dynamics.

While we introduced robustness to transition model uncertainty, future work could explore robustness to other forms of environmental perturbation, such as reward misspecification or partial observability in state information. These forms of uncertainty are common in deployed systems and present unique challenges for off-policy evaluation. New formulations of robust off-policy evaluation that explicitly

handle reward and state uncertainty could yield safer and more reliable estimators in noisy real-world environments.

Another future direction is to investigate how to stabilize transition-gradient-based methods, which we introduced for robust policy evaluation. While powerful in practice, these methods may suffer from instability when applied with deep function approximators. Techniques from modern policy optimization—such as trust region methods, clipping strategies, or adaptive baselines—could be adapted to stabilize the transition gradient updates. Understanding the theoretical and practical properties of these extensions could significantly enhance the robustness of evaluation in RL.

On the theoretical front, our ODE-based framework lays the foundation for analyzing the stability of stochastic approximation algorithms under Markovian noise, but several extensions are worth pursuing. One direction is to characterize the behavior of the learning iterates more precisely. For instance, it may be possible to establish a functional central limit theorem, capturing the asymptotic distribution of the iterates. Further refinements such as high-probability concentration bounds and almost sure convergence rates may also be derived. Pursuing these directions could broaden the applicability of our theory to a wider range of RL algorithms.

The tension between efficiency, robustness, safety, and stability is at the heart of policy evaluation—and, more broadly, reinforcement learning itself. The algorithms and analyses in this thesis do not offer a complete solution, but they provide new tools and perspectives that we hope will be useful for tackling the next set of challenges. We believe that progress in reinforcement learning depends not just on new algorithms, but on deepening our understanding of when, how, and why they work. In this spirit, we hope this thesis can serve as a stepping stone toward building reinforcement learning systems that are not only powerful, but also deployable and reliable.

Bibliography

Abounadi, J., Bertsekas, D., and Borkar, V. S. (2001). Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*.

Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR.

Agarwal, A., Basu, S., Schnabel, T., and Joachims, T. (2017). Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. (2018). A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60.

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

Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the International Conference on Machine Learning*.

Barto, A. G. and Sutton, R. S. (1981a). Goal seeking components for adaptive intelligence: An initial assessment. air force wright aeronautical laboratories. Technical report, Avionics Laboratory Technical Report AFWAL-TR-81-1070, Wright-Patterson AFB.

Barto, A. G. and Sutton, R. S. (1981b). Landmark learning: An illustration of associative search. *Biological Cybernetics*.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*.

Benjamin Melamed, R. Y. R. (1998). *Modern Simulation and Modeling (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer.

Berkenkamp, F., Turchetta, M., Schoellig, A., and Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific Belmont, MA.

Bertsekas, D. P. and Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642.

Bertsekas, D. P. and Yu, H. (2009). Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics*.

Bhatnagar, S. (2011). The borkar-meyn theorem for asynchronous stochastic approximations. *Systems & Control Letters*.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics.

Blaser, E. and Zhang, S. (2024). Asymptotic and finite sample analysis of nonexpansive stochastic approximations with markovian noise. *arXiv preprint arXiv:2409.19546*.

Borkar, V., Chen, S., Devraj, A., Kontoyiannis, I., and Meyn, S. (2021). The ode method for asymptotic statistics in stochastic approximation and reinforcement learning. *arXiv preprint arXiv:2110.14427*.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*. Springer.

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*.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). OpenAI Gym. *ArXiv Preprint*.

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. (2022). Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*.

Carpentier, A., Munos, R., and Antos, A. (2015). Adaptive strategy for stratified monte carlo sampling. *Journal of Machine Learning Research*.

Chen, C., Liu, S., and Zhang, S. (2025). Efficient policy evaluation with safety constraint for reinforcement learning. In *Proceedings of the International Conference on Learning Representations*.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*.

Chervonyi, Y., Dutta, P., Trochim, P., Voicu, O., Paduraru, C., Qian, C., Karagozler, E., Davis, J. Q., Chippendale, R., Bajaj, G., Witherspoon, S., and Luo, J. (2022). Semi-analytical industrial cooling system model for reinforcement learning.

Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. (2018). A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31.

Chua, K., Calandra, R., McAllister, R., and Levine, S. (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*.

Dai, J. G. (1995). On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability*.

Dai, J. G. and Meyn, S. P. (1995). Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control*.

Dann, C., Ghavamzadeh, M., and Marinov, T. V. (2023). Multiple-policy high-confidence policy evaluation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Deisenroth, M. P. and Rasmussen, C. E. (2011). PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the International Conference on Machine Learning*.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*.

Dunford, N. and Schwartz, J. T. (1988). *Linear operators, part 1: general theory*. John Wiley & Sons.

Farahmand, A.-m. and Szepesvári, C. (2011). Model selection in reinforcement learning. *Machine learning*, 85(3):299–332.

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*.

Fort, G., Meyn, S., Moulines, E., and Priouret, P. (2008). The ODE method for stability of skip-free Markov chains with applications to MCMC. *The Annals of Applied Probability*.

Fujimoto, S., Meger, D., Precup, D., Nachum, O., and Gu, S. S. (2022). Why should i trust you, bellman? the bellman error is a poor replacement for value error. *arXiv preprint arXiv:2201.12417*.

Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Geweke, J. (1988). Antithetic acceleration of monte carlo integration in bayesian inference. *Journal of Econometrics*.

Ghiassian, S., Patterson, A., Garg, S., Gupta, D., White, A., and White, M. (2020). Gradient temporal-difference learning with regularized corrections. In *International Conference on Machine Learning*.

Grand-Clément, J. and Kroer, C. (2021). Scalable first-order methods for robust mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12086–12094.

Graves, E., Imani, E., Kumaraswamy, R., and White, M. (2023). Off-policy actor-critic with emphatic weightings. *Journal of Machine Learning Research*.

Greensmith, E., Bartlett, P. L., and Baxter, J. (2004). Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*.

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., and Knoll, A. (2022). A review of safe reinforcement learning: Methods, theory and applications. *ArXiv Preprint*.

Guan, Z., Xu, T., and Liang, Y. (2021). Per-etd: A polynomially efficient emphatic temporal difference learning method. *arXiv preprint arXiv:2110.06906*.

Hallak, A., Tamar, A., Munos, R., and Mannor, S. (2016). Generalized emphatic temporal difference learning: Bias-variance analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hanna, J. P., Chandak, Y., Thomas, P. S., White, M., Stone, P., and Niekum, S. (2024). Data-efficient policy evaluation through behavior policy search. *Journal of Machine Learning Research*, 25(313):1–58.

Hanna, J. P., Thomas, P. S., Stone, P., and Niekum, S. (2017). Data-efficient policy evaluation through behavior policy search. In *Proceedings of the International Conference on Machine Learning*.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*.

Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge University Press.

Huang, S., Dossa, R. F. J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., and Araújo, J. G. (2022). Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*.

Imani, E., Graves, E., and White, M. (2018). An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems*.

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280.

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

Jiang, R., Zahavy, T., White, A., Xu, Z., Hessel, M., Blundell, C., and van Hasselt, H. (2021). Emphatic algorithms for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*.

Jiang, R., Zhang, S., Chelu, V., White, A., and van Hasselt, H. (2022). Learning expected emphatic traces for deep RL. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*.

Kahn, H. and Marshall, A. W. (1953). Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*.

Kakutani, S. (1945). Markoff process and the dirichlet problem. *Proceedings of the Japan Academy*.

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*.

Kallus, N., Saito, Y., and Uehara, M. (2021). Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*.

Kern, J. (2023). Skorokhod topologies: What they are and why we should care. *Mathematische Semesterberichte*.

Khalil, H. K. (2002). *Nonlinear Systems*. Prentice Hall.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Klopf, A. H. (1972). *Brain function and adaptive systems: a heterostatic theory*. Air Force Cambridge Research Laboratories, Air Force Systems Command, United States Air Force.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Mit Press.

Konda, V. R. and Tsitsiklis, J. N. (1999). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*.

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

Lai, J., Zou, L., and Song, J. (2020). Optimal mixture weights for off-policy evaluation with multiple behavior policies. *arXiv preprint arXiv:2011.14359*.

Lakshminarayanan, C. and Bhatnagar, S. (2017). A stability criterion for two timescale stochastic approximation schemes. *Automatica*.

Le, H. M., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *Proceedings of the International Conference on Machine Learning*.

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*.

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*.

Li, L. (2019). A perspective on off-policy evaluation in reinforcement learning. *Frontiers of Computer Science*.

Lin, L. J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*.

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

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*.

Liu, S., Chen, C., and Zhang, S. (2025a). Doubly optimal policy evaluation for reinforcement learning. In *Proceedings of the International Conference on Learning Representations*.

Liu, S., Chen, S., and Zhang, S. (2025b). The ode method for stochastic approximation and reinforcement learning with markovian noise. *Journal of Machine Learning Research*.

Liu, S., Chen, Y., and Zhang, S. (2025c). Efficient multi-policy evaluation for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, S., Shen, W., and Xu, H. (2021a). Optimal pricing of information. In *Proceedings of the ACM Conference on Economics and Computation*.

Liu, S. and Zhang, S. (2024). Efficient policy evaluation with offline data informed behavior policy design. In *Proceedings of the International Conference on Machine Learning*.

Liu, X., Xie, Z., and Zhang, S. (2025d). Linear Q-learning does not diverge: Convergence rates to a bounded set. *arXiv preprint arXiv:2501.19254*.

Liu, Y., Halev, A., and Liu, X. (2021b). Policy learning with constraints in model-free reinforcement learning: A survey. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Maei, H., Szepesvari, C., Bhatnagar, S., Precup, D., Silver, D., and Sutton, R. S. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. *Advances in Neural Information Processing Systems*.

Maei, H. R. (2011). *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta.

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

Maei, H. R. and Sutton, R. S. (2010). Gq ($lambda$): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the Conference on Artificial General Intelligence*.

Maei, H. R., Szepesvári, C., Bhatnagar, S., and Sutton, R. S. (2010). Toward off-policy learning control with function approximation. In *Proceedings of the International Conference on Machine Learning*.

Mahadevan, S., Liu, B., Thomas, P. S., Dabney, W., Giguere, S., Jacek, N., Gemp, I., and Liu, J. (2014). Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces. *arXiv preprint arXiv:1405.6757*.

Mahmood, A. R., Yu, H., and Sutton, R. S. (2017). Multi-step off-policy learning without importance sampling ratios. *arXiv preprint arXiv:1702.03006*.

Manek, G. and Kolter, J. Z. (2022). The pitfalls of regularization in off-policy td learning. *Advances in Neural Information Processing Systems*, 35:35621–35631.

Marivate, V. N. (2015). *Improved empirical methods in reinforcement-learning evaluation*. Rutgers The State University of New Jersey, School of Graduate Studies.

Mathieu, M., Ozair, S., Srinivasan, S., Gulcehre, C., Zhang, S., Jiang, R., Paine, T. L., Powell, R., Żołna, K., Schrittwieser, J., Choi, D., Georgiev, P., Toyama, D., Huang, A., Ring, R., Babuschkin, I., Ewalds, T., Bordbar, M., Henderson, S., Colmenarejo, S. G., van den Oord, A., Czarnecki, W. M., de Freitas, N., and Vinyals, O. (2023). Alphastar unplugged: Large-scale offline reinforcement learning.

Meyn, S. (2008). *Control techniques for complex networks*. Cambridge University Press.

Meyn, S. (2022). *Control systems and reinforcement learning*. Cambridge University Press.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Molchanov, A. P. and Pyatnitskiy, Y. S. (1989). Criteria of asymptotic stability of differential and difference inclusions encountered in control theory. *Systems & Control Letters*.

Moldovan, T. M. and Abbeel, P. (2012). Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*.

Mukherjee, S., Hanna, J. P., and Nowak, R. D. (2022). Revar: Strengthening policy evaluation via reduced variance sampling. In *Proceedings of the Conference in Uncertainty in Artificial Intelligence*.

Munos, R. (2003). Error bounds for approximate policy iteration. In *Proceedings of the International Conference on Machine Learning*.

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

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*.

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*. Springer.

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. (2018). The uncertainty bellman equation and exploration. In *Proceedings of the International Conference on Machine Learning*.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples*. Stanford.

Precup, D., Sutton, R. S., and Dasgupta, S. (2001). Off-policy temporal difference learning with function approximation. In *Proceedings of the International Conference on Machine Learning*.

Precup, D., Sutton, R. S., and Singh, S. P. (2000a). Eligibility traces for off-policy policy evaluation. In *Proceedings of the International Conference on Machine Learning*.

Precup, D., Sutton, R. S., and Singh, S. P. (2000b). Eligibility traces for off-policy policy evaluation. In *Proceedings of the International Conference on Machine Learning*.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qian, X., Xie, Z., Liu, X., and Zhang, S. (2024). Almost sure convergence rates and concentration of stochastic approximation and reinforcement learning with markovian noise. *arXiv preprint arXiv:2411.13711*.

Qian, X. and Zhang, S. (2025). Revisiting a design choice in gradient temporal difference learning. In *Proceedings of the International Conference on Learning Representations*.

Ramaswamy, A. and Bhatnagar, S. (2017). A generalization of the borkar-meyn theorem for stochastic recursive inclusions. *Mathematics of Operations Research*.

Ramaswamy, A. and Bhatnagar, S. (2018). Stability of stochastic approximations with "controlled markov" noise and temporal difference learning. *IEEE Transactions on Automatic Control*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*.

Royden, H. L. and Fitzpatrick, P. (1968). *Real analysis*. Macmillan New York.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. Wiley.

Rummery, G. A. and Niranjan, M. (1994). *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering Cambridge, UK.

Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. (2016). High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations*.

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

Seijen, H. and Sutton, R. (2014). True online TD(λ). In *Proceedings of the International Conference on Machine Learning*.

Sherstan, C., Bennett, B., Young, K., Ashley, D. R., White, A., White, M., and Sutton, R. S. (2018). Directly estimating the variance of the λ-return using temporal-difference methods. *arXiv preprint arXiv:1801.08287*.

Skorokhod, A. V. (1956). Limit theorems for stochastic processes. *Theory of Probability & Its Applicationss*.

Sun, Z., He, S., Miao, F., and Zou, S. (2024). Policy optimization for robust average reward mdps. *Advances in Neural Information Processing Systems*, 37:17348–17372.

Sutton, R., Mahmood, A. R., Precup, D., and Hasselt, H. (2014). A new Q(λ) with interim forward view and monte carlo equivalence. In *International Conference on Machine Learning*.

Sutton, R. S. (1978). Single channel theory: A neuronal theory of learning. *Brain Theory Newsletter*.

Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*. University of Massachusetts Amherst.

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

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the International Conference on Machine Learning*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press.

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

Sutton, R. S., Mahmood, A. R., and White, M. (2016). An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., and Precup, D. (2011). Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*.

Sutton, R. S., Szepesvári, C., Geramifard, A., and Bowling, M. H. (2008a). Dyna-style planning with linear function approximation and prioritized sweeping. In *Proceedings of the Conference in Uncertainty in Artificial Intelligence*.

Sutton, R. S., Szepesvári, C., and Maei, H. R. (2008b). A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*.

Tadic, V. (2001). On the convergence of temporal-difference learning with linear function approximation. *Machine Learning*.

Tamar, A., Castro, D. D., and Mannor, S. (2016). Learning the variance of the reward-to-go. *Journal of Machine Learning Research*.

Thomas, P. S. (2015). *Safe reinforcement learning*. PhD thesis, University of Massachusetts Amherst.

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

Thomas, P. S. and Brunskill, E. (2017). Policy gradient methods for reinforcement learning with function approximation and action-dependent baselines. *arXiv preprint arXiv:1706.06643*.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *Proceedings of the International Conference on Intelligent Robots and Systems*.

Touati, A., Bacon, P.-L., Precup, D., and Vincent, P. (2018). Convergent tree backup and retrace with function approximation. In *Proceedings of the International Conference on Machine Learning*.

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al. (2024). Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*.

Tsitsiklis, J. N. and Roy, B. V. (1996). Analysis of temporal-diffference learning with function approximation. In *IEEE Transactions on Automatic Control*.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gülçehre, Ç., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T. P., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nature*.

Wachi, A. and Sui, Y. (2020). Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR.

Wan, Y., Naik, A., and Sutton, R. S. (2021). Learning and planning in average-reward markov decision processes. In *Proceedings of the International Conference on Machine Learning*.

Wang, J. and Zhang, S. (2024). Almost sure convergence of linear temporal difference learning with arbitrary features. *arXiv preprint arXiv:2409.12135*.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2020). On the global optimality of model-agnostic meta-learning. In *International conference on machine learning*, pages 9837–9846. PMLR.

Wang, Q., Ho, C. P., and Petrik, M. (2023). Policy gradient in robust mdps with global convergence guarantee. In *International Conference on Machine Learning*, pages 35763–35797. PMLR.

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

Wang, Y. and Zou, S. (2020). Finite-sample analysis of greedy-gq with linear function approximation under markovian noise. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD thesis, King's College, Cambridge.

White, A. and White, M. (2016). Investigating practical linear temporal difference learning. *arXiv preprint arXiv:1602.08771*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*.

Wu, C., Rajeswaran, A., Duan, Y., Kumar, V., Bayen, A. M., Kakade, S., Mordatch, I., and Abbeel, P. (2018). Variance reduction for policy gradient with action-dependent factorized baselines. *arXiv preprint arXiv:1803.07246*.

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

Xu, T., Zou, S., and Liang, Y. (2019). Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples. *Advances in Neural Information Processing Systems*, 32.

Yu, H. (2012). Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*.

Yu, H. (2015). On convergence of emphatic temporal-difference learning. In *Proceedings of the Conference on Learning Theory*.

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

Zhang, S. (2022). *Breaking the deadly triad in reinforcement learning*. PhD thesis, University of Oxford.

Zhang, S. (2023). A new challenge in policy evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, S., Boehmer, W., and Whiteson, S. (2019). Generalized off-policy actor-critic. In *Advances in Neural Information Processing Systems*.

Zhang, S., Liu, B., and Whiteson, S. (2020a). GradientDICE: Rethinking generalized offline estimation of stationary values. In *Proceedings of the International Conference on Machine Learning*.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. (2020b). Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proceedings of the International Conference on Machine Learning*.

Zhang, S., Veeriah, V., and Whiteson, S. (2020c). Learning retrospective knowledge with reverse reinforcement learning. In *Advances in Neural Information Processing Systems*.

Zhang, S., Wan, Y., Sutton, R. S., and Whiteson, S. (2021a). Average-reward off-policy policy evaluation with function approximation. In *Proceedings of the International Conference on Machine Learning*.

Zhang, S. and Whiteson, S. (2022). Truncated emphatic temporal difference methods for prediction and control. *Journal of Machine Learning Research*.

Zhang, S., Yao, H., and Whiteson, S. (2021b). Breaking the deadly triad with a target network. In *Proceedings of the International Conference on Machine Learning*.

Zhong, R., Zhang, D., Schäfer, L., Albrecht, S. V., and Hanna, J. P. (2022). Robust on-policy sampling for data-efficient policy evaluation in reinforcement learning. In *Advances in Neural Information Processing Systems*.

Appendix A

Appendix for Chapter 4

A.1 Proofs

A.1.1 Proof of Lemma 1

Proof.

$$\begin{aligned}
\mathbb{E}_{A\sim\mu}\left[\rho(A)q(A)\right] =& \sum_{a\in\{a|\mu(a)>0\}} \mu(a)\frac{\pi(a)}{\mu(a)}q(a) \\
=& \sum_{a\in\{a|\mu(a)>0\}} \pi(a)q(a) \\
=& \sum_{a\in\{a|\mu(a)>0\}} \pi(a)q(a) + \sum_{a\in\{a|\mu(a)=0\}} \pi(a)q(a) \qquad (\mu\in\Lambda) \\
=& \sum_{a} \pi(a)q(a) \\
=& \mathbb{E}_{A\sim\pi}\left[q(A)\right].
\end{aligned}$$

The intuition in the third equation is that the sample a where μ does not cover π must satisfy $q(a) = 0$, i.e., this sample does not contribute to the expectation anyway. □

A.1.2 Proof of Lemma 2

Proof.
For a given π and q, define

$$\mathcal{A}_+ \doteq \{a \mid \pi(a)q(a) \neq 0\}.$$

For any $\mu \in \Lambda$, we expand the variance as

$$\begin{aligned}
&\mathbb{V}_{A\sim\mu}(\rho(A)q(A)) \\
=&\mathbb{E}_{A\sim\mu}[(\rho(A)q(A))^2] - \mathbb{E}^2_{A\sim\mu}[\rho(A)q(A)] \\
=&\mathbb{E}_{A\sim\mu}[(\rho(A)q(A))^2] - \mathbb{E}^2_{A\sim\pi}[q(A)] && \text{(Lemma 1)} \\
=&\sum_{a\in\{a|\mu(a)>0\}} \frac{\pi^2(a)q^2(a)}{\mu(a)} - \mathbb{E}^2_{A\sim\pi}[q(A)] \\
=&\sum_{a\in\{a|\mu(a)>0\}\cap\mathcal{A}_+} \frac{\pi^2(a)q^2(a)}{\mu(a)} - \mathbb{E}^2_{A\sim\pi}[q(A)] && (\pi(a)q(a)=0, \forall a \notin \mathcal{A}_+) \\
=&\sum_{a\in\mathcal{A}_+} \frac{\pi^2(a)q^2(a)}{\mu(a)} - \mathbb{E}_{A\sim\pi}[q(A)]^2 && (\mu \in \Lambda)
\end{aligned}$$

The second term is a constant and is unrelated to μ. Solving the optimization problem (4.7) is, therefore, equivalent to solving

$$\min_{\mu\in\Lambda} \quad \sum_{a\in\mathcal{A}_+} \frac{\pi^2(a)q^2(a)}{\mu(a)}. \tag{A.1}$$

Case 1: $|\mathcal{A}_+| = 0$

In this case, the variance is always 0 so any $\mu \in \Lambda$ is optimal. In particular, $\mu^*(a) = \frac{1}{\mathcal{A}}$ is optimal.

Case 2: $|\mathcal{A}_+| > 0$

The definition of Λ in (4.6) can be equivalently expressed, using contraposition, as

$$\Lambda = \{\mu \in \Delta(\mathcal{A}) \mid \forall a, a \in \mathcal{A}_+ \implies \mu(a) > 0\}.$$

The optimization problem (A.1) can then be equivalently written as

$$\begin{aligned}
\min_{\mu\in\Delta(\mathcal{A})} \quad & \sum_{a\in\mathcal{A}_+} \frac{\pi^2(a)q^2(a)}{\mu(a)} \\
\text{s.t.} \quad & \mu(a) > 0 \quad \forall a \in \mathcal{A}_+.
\end{aligned} \tag{A.2}$$

If for some μ we have $\sum_{a\in\mathcal{A}_+} \mu(a) < 1$, then there must exist some $a_0 \notin \mathcal{A}_+$ such that $\mu(a_0) > 0$. Since a_0 does not contribute to the summation in the objective function of (A.2), we can move the probability mass on a_0 to some other $a_1 \in \mathcal{A}_+$ to increase $\mu(a_1)$ to further decrease the objective. In other words, any optimal solution μ to (A.2) must put all its mass on $\mathcal{A}_+$. This motivates the following problem

$$\begin{aligned}
\min_{z\in\Delta(\mathcal{A}_+)} \quad & \sum_{a\in\mathcal{A}_+} \frac{\pi^2(a)q^2(a)}{z(a)} \\
\text{s.t.} \quad & z(a) > 0 \quad \forall a \in \mathcal{A}_+.
\end{aligned} \tag{A.3}$$

In particular, if z_* is an optimal solution to (A.3), then an optimal solution to (A.2) can be constructed as

$$\mu_*(a) = \begin{cases} z_*(a) & a \in \mathcal{A}_+ \\ 0 & \text{otherwise.} \end{cases} \tag{A.4}$$

Let $\mathbb{R}_{++} \doteq (0, +\infty)$.

According to the Cauchy-Schwarz inequality, for any $z \in \mathbb{R}_{++}^{|\mathcal{A}_+|}$, we have

$$\left(\sum_{a\in\mathcal{A}_+} \frac{\pi^2(a)q^2(a)}{z(a)}\right)\left(\sum_{a\in\mathcal{A}_+} z(a)\right) \geq \left(\sum_{a\in\mathcal{A}_+} \frac{\pi(a)|q(a)|}{\sqrt{z(a)}}\sqrt{z(a)}\right)^2 = \left(\sum_{a\in\mathcal{A}_+} \pi(a)|q(a)|\right)^2.$$

It can be easily verified that the equality holds for

$$z^*(a) \doteq \frac{\pi(a)|q(a)|}{\sum_b \pi(b)|q(b)|} > 0.$$

Since $\sum_{a\in\mathcal{A}_+} z^*(a) = 1$, we conclude that z^* is an optimal solution to (A.3). An optimal solution μ_* to (4.7) can then be constructed according to (A.4). Making use of the fact that $\pi(a)|q(a)| = 0$ for $a \notin \mathcal{A}_+$, this μ_* can be equivalently expressed as

$$\mu_*(a) = \frac{\pi(a)|q(a)|}{\sum_{b\in\mathcal{A}} \pi(b)q(b)},$$

which completes the proof. □

A.1.3 Proof of Lemma 3

Proof. We start by showing $\Lambda = \Lambda_+$. Lemma 1 ensures that $\mu \in \Lambda \implies \mu \in \Lambda_+$. We now show that $\mu \in \Lambda_+ \implies \mu \in \Lambda$. For any $\mu \in \Lambda_+$, we have

$$\sum_{a\in\{a|\mu(a)>0\}} \mu(a)\frac{\pi(a)}{\mu(a)}q(a) = \sum_a \pi(a)q(a).$$

This indicates that

$$\sum_{a\in\{a|\mu(a)=0\}} \pi(a)q(a) = 0.$$

Since $\pi(a) \geq 0$ and all $q(a)$ has the same sign, we must have

$$\pi(a)q(a) = 0,\ \forall a \in \{a \mid \mu(a) = 0\}.$$

This is exactly $\mu(a) = 0 \implies \pi(a)q(a) = 0$, yielding $\mu \in \Lambda$. This completes the proof of $\Lambda_+ = \Lambda$.

We now show the zero variance. When $\forall a \in \mathcal{A}, q(a) \geq 0$, if $\exists a_0, \pi_0(a_0)q(a_0) \neq 0$, we have $\forall a \in \mathcal{A}$

$$\mu^*(a) = \frac{\pi(a)|q(a)|}{c}$$

and $c > 0$ is a normalizing constant. Plugging μ^* to $\rho(A)q(A)$, we get $\forall a \in \mathcal{A}$

$$\rho(a)q(a) = \frac{\pi(a)}{\mu^*(a)}q(a) = \frac{\pi(a)}{\frac{\pi(a)|q(a)|}{c}}q(a) = c.$$

This means in this setting, with the optimal distribution μ^*, the random variable $\rho(\cdot)q(\cdot)$ is a constant function. Thus,

$$\mathbb{V}_{A\sim\mu^*}(\rho(A)q(A)) = 0.$$

When $\forall a \in \mathcal{A}, q(a) \geq 0$, if $\forall a_0, \pi_0(a_0)q(a_0) = 0$, we have $\forall a \in \mathcal{A}$

$$\mu^*(a) = \frac{1}{|\mathcal{A}|}.$$

Plugging μ^* to $\rho(A)q(A)$, we get $\forall a \in \mathcal{A}$

$$\rho(a)q(a) = \frac{\pi(a)}{\mu^*(a)}q(a) = \frac{\pi(a)q(a)}{\frac{1}{|\mathcal{A}|}} = 0.$$

This shows $\rho(A)q(A)$ is also a constant. Thus,

$$\mathbb{V}_{A\sim\mu^*}(\rho(A)q(A)) = 0.$$

The proof is similar for $\forall a \in \mathcal{A}, q(a) \leq 0$ and is thus omitted.

□

A.1.4 Proof of Theorem 1

Proof. We proceed via induction. For $t = T - 1$, we have

$$\begin{aligned}
\mathbb{E}\left[G^{\text{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right] =& \mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] = \mathbb{E}\left[\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right] \\
=& \mathbb{E}_{A_t\sim\pi_t(\cdot|S_t)}\left[q_{\pi,t}(S_t, A_t) \mid S_t\right] && \text{(Lemma 1)} \\
=& v_{\pi,t}(S_t).
\end{aligned}$$

For $t \in [T-2]$, we have

$$
\begin{aligned}
&\mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t R_{t+1} + \rho_t G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] + \mathbb{E}\left[\rho_t G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t \sim \mu_t(\cdot|S_t), S_{t+1} \sim p(\cdot|S_t, A_t)}\left[\mathbb{E}\left[\rho_t G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right] \mid S_t\right] \\
& \qquad \text{(Law of total expectation)} \\
=&\mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t \sim \mu_t(\cdot|S_t), S_{t+1} \sim p(\cdot|S_t, A_t)}\left[\rho_t \mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right] \mid S_t\right] \\
& \qquad \text{(Conditional independence and Markov property)} \\
=&\mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t \sim \mu_t(\cdot|S_t), S_{t+1} \sim p(\cdot|S_t, A_t)}\left[\rho_t v_{\pi,t+1}(S_{t+1}) \mid S_t\right] \\
& \qquad \text{(Inductive hypothesis)} \\
=&\mathbb{E}_{A_t \sim \mu_t(\cdot|S_t)}\left[\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right] \qquad \text{(Definition of } q_{\pi,t}\text{)} \\
=&\mathbb{E}_{A_t \sim \pi_t(\cdot|S_t)}\left[q_{\pi,t}(S_t, A_t) \mid S_t\right] \qquad \text{(Lemma 1)} \\
=&v_{\pi,t}(S_t),
\end{aligned}
$$

which completes the proof. □

A.1.5 Proof of Theorem 2

To prove Theorem 2, we rely on a recursive expression of the PDIS Monte Carlo estimator summarized by the following lemma.

Lemma 34 (Recursive Expression of Variance)**.** *For any* $\mu \in \Lambda$, *for* $t = T-1$,

$$
\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) = \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^2 q_{\pi,t}^2(S_t, A_t) \mid S_t\right] - v_{\pi,t}^2(S_t),
$$

for $t \in [T-2]$,

$$
\begin{aligned}
&\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^2 \left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{PDIS}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t\right) \mid S_t, A_t\right] + \nu_{\pi,t}(S_t, A_t) + q_{\pi,t}^2(S_t, A_t)\right) \mid S_t\right] \\
&- v_{\pi,t}^2(S_t).
\end{aligned}
$$

Proof. When $t \in [T-2]$, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) && \text{(A.5)}\\
=&\mathbb{E}_{A_t}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t, A_t\right] \mid S_t\right) \\
& && \text{(Law of total variance)}\\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(r(S_t, A_t) + G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\left(\rho_t \mathbb{E}\left[r(S_t, A_t) + G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right] \mid S_t\right) && \text{(Using (7.1))}\\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\rho_t \mathbb{E}\left[r(S_t, A_t) + G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right] \mid S_t\right) \\
& && \text{(Deterministic reward } r\text{)}\\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right).
\end{aligned}
$$

Further decomposing the first term, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) && \text{(A.6)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right) \mid S_t, A_t\right] \\
&+ \mathbb{V}_{S_{t+1}}\left(\mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right] \mid S_t, A_t\right) && \text{(Law of total variance)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}\left(\mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right] \mid S_t, A_t\right) \\
& && \text{(Markov property)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}\left(v_{\pi,t+1}(S_{t+1}) \mid S_t, A_t\right). \\
& && \text{(Theorem 14)}
\end{aligned}
$$

With $\nu_{\pi,t}$ defined in (4.10), plugging (A.6) back to (A.5) yields

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{E}_{A_t}\left[\rho_t^2 q_{\pi,t}^2(S_t, A_t) \mid S_t\right] - \left(\mathbb{E}_{A_t}\left[\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right]\right)^2 \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{E}_{A_t}\left[\rho_t^2 q_{\pi,t}^2(S_t, A_t) \mid S_t\right] - v_{\pi,t}^2(S_t). && \text{(Lemma 1)}
\end{aligned}
$$

When $t = T-1$, we have

$$
\begin{aligned}
\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) =& \mathbb{V}\left(\rho_t r(S_t, A_t) \mid S_t\right) \\
=& \mathbb{V}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right) \\
=& \mathbb{E}_{A_t}\left[\rho_t^2 q_{\pi,t}^2(S_t, A_t) \mid S_t\right] - v_{\pi,t}^2(S_t),
\end{aligned}
$$

which completes the proof. □

We restate and present the main proof of Theorem 2.

Theorem 2 (Optimal Behavior Policy)**.** *For any t and s, the behavior policy $\mu_t^*(a|s)$ defined above is an optimal solution to the following problem*

$$\min_{\mu_t \in \Lambda_t, \ldots, \mu_{T-1} \in \Lambda_{T-1}} \quad \mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right),$$

where $\Lambda_t \doteq \{\mu_t \in \Delta(\mathcal{A}) \mid \forall s, a, \mu_t(a|s) = 0 \implies \pi_t(a|s) u_{\pi,t}(s,a) = 0\}$.

Proof. We proceed via induction. When $t = T-1$, we have

$$\begin{aligned}
&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{T-1:T-1}^{\mu_{T-1:T-1}}) \mid S_{T-1} = s\right) \\
=&\mathbb{V}_{A_{T-1}}\left(\rho_{T-1} r(s, A_{T-1}) \mid S_{T-1} = s\right) \\
=&\mathbb{V}_{A_{T-1}}\left(\rho_{T-1} q_{\pi,T-1}(s, A_{T-1}) \mid S_{T-1} = s\right).
\end{aligned}$$

The definition of μ_{T-1}^* in (4.11) and Lemma 2 ensure that μ_{T-1}^* is an optimal solution to

$$\min_{\mu_{T-1} \in \Lambda_{T-1}} \quad \mathbb{V}\left(G^{\text{PDIS}}\left(\tau_{T-1}^{\mu_{T-1}}\right) \mid S_{T-1} = s\right).$$

Now, suppose for some $t \in [T-2]$, $\mu_{t+1:T-1}^*$ is an optimal solution to

$$\min_{\mu_{t+1} \in \Lambda_{t+1}, \ldots, \mu_{T-1} \in \Lambda_{T-1}} \quad \mathbb{V}\left(G^{\text{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_{t+1} = s\right).$$

To complete induction, we proceed to proving that $\mu_{t:T-1}^*$ is an optimal solution to

$$\min_{\mu_t \in \Lambda_t, \ldots, \mu_{T-1} \in \Lambda_{T-1}} \quad \mathbb{V}\left(G^{\text{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t = s\right). \tag{A.7}$$

In the rest of this proof, we omit the domain $\Lambda_t, \ldots, \Lambda_{T-1}$ for simplifying notations.

For any $\mu_{t:T-1}$, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) + q_{\pi,t}^2(S_t, A_t)\right) \mid S_t\right] \\
&- v_{\pi,t}^2(S_t) && \text{(By Lemma 34)} \\
\overset{(a)}{\geq}&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\min_{\mu'_{t+1:T-1}} \mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu'_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) + q_{\pi,t}^2(S_t, A_t)\right) \mid S_t\right] \\
&- v_{\pi,t}^2(S_t) && \text{(Monotonically non-increasing in } \mathbb{V}(\cdot)) \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) + q_{\pi,t}^2(S_t, A_t)\right) \mid S_t\right] \\
&- v_{\pi,t}^2(S_t) && \text{(Inductive hypothesis)} \\
=&\mathbb{E}_{A_t}\left[\rho_t^2 u_{\pi,t}(S_t, A_t) \mid S_t\right] - v_{\pi,t}^2(S_t) && \text{(By (4.12))} \\
=&\mathbb{V}_{A_t}\left(\rho_t\sqrt{u_{\pi,t}(S_t, A_t)} \mid S_t\right) + \mathbb{E}_{A_t}\left[\rho_t\sqrt{u_{\pi,t}(S_t, A_t)} \mid S_t\right]^2 - v_{\pi,t}^2(S_t) \\
& && \text{(Definition of variance)} \\
=&\mathbb{V}_{A_t}\left(\rho_t\sqrt{u_{\pi,t}(S_t, A_t)} \mid S_t\right) + \mathbb{E}_{A_t\sim\pi_t(\cdot|S_t)}\left[\sqrt{u_{\pi,t}(S_t, A_t)} \mid S_t\right]^2 - v_{\pi,t}^2(S_t) \\
& && \text{(Lemma 1 and } \mu_t \in \Lambda_t) \\
\overset{(b)}{\geq}&\mathbb{E}_{A_t\sim\pi_t(\cdot|S_t)}\left[\sqrt{u_{\pi,t}(S_t, A_t)} \mid S_t\right]^2 - v_{\pi,t}^2(S_t). && \text{(Non-negativity of variance)}
\end{aligned}
$$

According to the inductive hypothesis, the equality in (a) can be achieved when $\mu_{t+1:T-1} = \mu^*_{t+1:T-1}$. According to the construction of μ_t^* in (4.11) and Lemma 3, the equality in (b) can be achieved when $\mu_t = \mu_t^*$. This suggests that $\mu^*_{t:T-1}$ achieves the lower bound and is thus an optimal solution to (A.7), which completes the induction and thus completes the proof. □

A.1.6 Proof of Theorem 3

To prove the variance reduction property of $\hat{\mu}$, we express $\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right)$, the variance of the on-policy Monte Carlo estimator, in the form of a Bellman equation (Tamar et al., 2016; O'Donoghue et al., 2018; Sherstan et al., 2018). Define

$$
\tilde{r}_{\pi,t}(s,a) \doteq \nu_{\pi,t}(s,a) + q_{\pi,t}^2(s,a) - v_{\pi,t}^2(s) \quad \forall t \in [T-1], \tag{A.8}
$$

$$
\tilde{q}_{\pi,t}(s,a) \doteq \begin{cases} \tilde{r}_{\pi,t}(s,a) + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')\tilde{q}_{\pi,t+1}(s',a') & \text{if } t \in [T-2] \\ \tilde{r}_{\pi,t}(s,a) & \text{if } t = T-1 \end{cases} \tag{A.9}
$$

We have

Lemma 35 (Variance Equality)**.**

$$\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) = \sum_a \pi_t(a|s)\tilde{q}_{\pi,t}(s,a) \quad \forall t, s.$$

Proof. We proceed via induction. When $t = T-1$, we have

$$\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{V}_{A_t}\left(\rho_t r(S_t, A_t) \mid S_t\right) \\
=&\mathbb{V}_{A_t}\left(r(S_t, A_t) \mid S_t\right) && \text{(By on-policy)} \\
=&\mathbb{V}_{A_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[q_{\pi,t}^2(S_t, A_t) \mid S_t\right] - v_{\pi,t}^2(S_t) \\
=&\sum_a \pi_t(a|S_t)\tilde{q}_{\pi,t}(S_t, a). && \text{(By (A.9) and } \nu_{\pi,T-1}(s,a) = 0)
\end{aligned}$$

For $t \in [T-2]$, we have

$$\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + q_{\pi,t}^2(S_t, A_t) + \nu_{\pi,t}(S_t, A_t) \mid S_t\right] - v_{\pi,t}^2(S_t) \\
& && \text{(Lemma 34 and on-policy)} \\
=&\sum_a \pi_t(a|S_t)\left(\sum_{s'} p(s'|S_t, a)\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1} = s'\right) + \tilde{r}(S_t, a)\right) \\
=&\sum_a \pi_t(a|S_t)\left(\sum_{s'} p(s'|S_t, a)\sum_{a'} \pi_{t+1}(a'|s')\tilde{q}_{\pi,t+1}(s', a') + \tilde{r}(S_t, a)\right) \\
& && \text{(Inductive hypothesis)} \\
=&\sum_a \pi_t(a|S_t)\tilde{q}_{\pi,t}(S_t, a), && \text{(By (A.9))}
\end{aligned}$$

which completes the proof. □

Here, this $\tilde{q}$ is exactly the state-action value function of the target policy π in the MDP w.r.t. to a new reward function $\tilde{r}$. Manipulating (4.14) then yields

$$\begin{aligned}
\hat{q}_{\pi,t}(s,a) =& \sum_{s'} p(s'|s,a)\sum_{a'} \pi_{t+1}(a'|s')\tilde{q}_{\pi,t+1}(s',a') + \nu_t(s,a) + q_{\pi,t}^2(s,a) \\
=&\tilde{q}_{\pi,t}(s,a) + v_{\pi,t}^2(s). && \text{(A.10)}
\end{aligned}$$

Now, we restate and present the main proof of Theorem 3.

Theorem 3 (Variance Reduction)**.** *For any t and s,*

$$\begin{aligned}
&\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}) \mid S_t = s\right) \\
\leq&\mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) - \epsilon_t(s).
\end{aligned}$$

To define $\epsilon_t(s)$*, first define* $c_t(s) =$

$$\sum_a \pi_t(a|s)\hat{q}_{\pi,t}(s,a) - \left(\sum_a \pi_t(a|s)\sqrt{\hat{q}_{\pi,t}(s,a)}\right)^2.$$

Then we define $\epsilon_t(s) \doteq c_t(s)$ *for* $t = T-1$ *and otherwise*

$$\epsilon_t(s) \doteq c_t(s) + \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2 \mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1})|s,A_t\right]\right]. \tag{4.18}$$

Proof. We proceed via induction. For $t = T-1$, we have

$$\begin{aligned}
&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}) \mid S_t\right)\\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2 q_{\pi,t}^2(S_t,A_t) \mid S_t\right] - v_{\pi,t}^2(S_t) && \text{(Lemma 34)}\\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2 \hat{q}_{\pi,t}(S_t,A_t) \mid S_t\right] - v_{\pi,t}^2(S_t) && \text{(Definition of } \hat{q} \text{ (4.13))}\\
=&\mathbb{V}_{A_t\sim\hat{\mu}_t}\left(\rho_t\sqrt{\hat{q}_{\pi,t}(S_t,A_t)}|S_t\right) + \mathbb{E}^2_{A_t\sim\hat{\mu}_t}\left[\rho_t\sqrt{\hat{q}_{\pi,t}(S_t,A_t)}|S_t\right] - v_{\pi,t}^2(S_t)\\
& && \text{(Definition of variance and non-negativity of } \hat{q})\\
=&\mathbb{V}_{A_t\sim\hat{\mu}_t}\left(\rho_t\sqrt{\hat{q}_{\pi,t}(S_t,A_t)}|S_t\right) + \left(\sum_a \pi_t(a|S_t)\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)^2 - v_{\pi,t}^2(S_t) && \text{(Lemma 1)}\\
=&\left(\sum_a \pi_t(a|S_t)\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)^2 - v_{\pi,t}^2(S_t) && \text{(Definition of } \hat{\mu} \text{ (4.17) and Lemma 3)}\\
=&\sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a) + \left(\sum_a \pi_t(a|S_t)\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)^2 - \sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a) - v_{\pi,t}^2(S_t)\\
=&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t\right) + \left(\sum_a \pi_t(a|S_t)\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)^2 - \sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a)\\
& && \text{(By (A.10) and Lemma 35)}\\
=&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t\right) - \epsilon_t(S_t). && \text{(Definition of } \epsilon \text{ (4.18))}
\end{aligned}$$

For $t \in [T-2]$, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\hat{\mu}_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_{\pi,t}(S_t,A_t) + q_{\pi,t}^2(S_t,A_t)\right) \mid S_t\right] \\
&- v_{\pi,t}^2(S_t) && \text{(Lemma 34)} \\
\leq&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\sum_{a'}\pi_{t+1}(a'|S_{t+1})\tilde{q}_{\pi,t+1}(S_{t+1},a') \mid S_t, A_t\right] + \nu_{\pi,t}(S_t,A_t)\right.\right. \\
&\left.\left.+ q_{\pi,t}^2(S_t,A_t)\right) \mid S_t\right] - v_{\pi,t}^2(S_t) - \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
&&& \text{(Inductive hypothesis and Lemma 35)} \\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\left(\tilde{q}_{\pi,t}(S_t,A_t) + v_{\pi,t}^2(S_t)\right) \mid S_t\right] - v_{\pi,t}^2(S_t) - \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
&&& \text{(Definition of } \tilde{q} \text{ (A.9))} \\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\hat{q}_{\pi,t}(S_t,A_t) \mid S_t\right] - v_{\pi,t}^2(S_t) - \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
&&& \text{(Definition of } \hat{q} \text{ (A.10))} \\
=&\mathbb{V}_{A_t\sim\hat{\mu}_t}\left(\rho_t\sqrt{\hat{q}_{\pi,t}(S_t,A_t)}|S_t\right) + \mathbb{E}_{A_t\sim\hat{\mu}_t}^2\left[\rho_t\sqrt{\hat{q}_{\pi,t}(S_t,A_t)}|S_t\right] - v_{\pi,t}^2(S_t) \\
&- \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
&&& \text{(Definition of variance and non-negativity of } \hat{q}) \\
=&\mathbb{V}_{A_t\sim\hat{\mu}_t}\left(\rho_t\sqrt{\hat{q}_{\pi,t}(S_t,A_t)}|S_t\right) + \left(\sum_a \pi_t(a|S_t)\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)^2 - v_{\pi,t}^2(S_t) \\
&- \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] && \text{(Lemma 1)} \\
=&\left(\sum_a \pi_t(a|S_t)\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)^2 - v_{\pi,t}^2(S_t) - \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
&&& \text{(Definition of } \hat{\mu} \text{ (4.17) and Lemma 3)} \\
=&\sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a) - v_{\pi,t}^2(S_t) + \left(\sum_a \pi_t(a|S_t)\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)^2 - \sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a) \\
&- \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
=&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t\right) + \left(\sum_a \pi_t(a|S_t)\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right)^2 - \sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a) \\
&- \mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] && \text{(By (A.10) and Lemma 35)} \\
=&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t\right) - \epsilon_t(S_t). && \text{(Definition of } \epsilon \text{ (4.18))}
\end{aligned}
$$

□

A.1.7 Proof of Theorem 4

Proof. For $t = T-1$, we have

$$\begin{aligned}\hat{q}_{\pi,t}(s,a) &= q^2_{\pi,t}(s,a) && \text{(Definition of } \hat{q}_{\pi,t} \text{ (4.13))}\\ &= \hat{r}_{\pi,t}(s,a). && \text{(By } q_{\pi,T-1}(s,a) = r(s,a) \text{ and Theorem 4)}\end{aligned}$$

For $t \in [T-2]$, we have

$$\begin{aligned}&\hat{q}_{\pi,t}(s,a)\\ =&\tilde{q}_{\pi,t}(s,a) + v^2_{\pi,t}(s) && \text{(By (A.10))}\\ =&\tilde{r}_{\pi,t}(s,a) + v^2_{\pi,t}(s) + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')\tilde{q}_{\pi,t+1}(s',a') && \text{(Definition of } \tilde{q} \text{ (A.9))}\\ =&\tilde{r}_{\pi,t}(s,a) + v^2_{\pi,t}(s) + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')(\tilde{q}_{\pi,t+1}(s',a') + v^2_{\pi,t+1}(s') - v^2_{\pi,t+1}(s'))\\ =&\tilde{r}_{\pi,t}(s,a) + v^2_{\pi,t}(s) + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')(\hat{q}_{\pi,t+1}(s',a') - v^2_{\pi,t+1}(s')) && \text{(By (A.10))}\\ =&\nu_{\pi,t}(s,a) + q^2_{\pi,t}(s,a) - \sum_{s'} p(s'|s,a)v^2_{\pi,t+1}(s') + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')\hat{q}_{\pi,t+1}(s',a')\\ & && \text{(Definition of } \tilde{r} \text{ (A.8))}\\ =&-(\mathbb{E}[v_{\pi,t+1}(S_{t+1}) \mid S_t = s, A_t = a])^2 + q^2_{\pi,t}(s,a) + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')\hat{q}_{\pi,t+1}(s',a')\\ & && \text{(Definition of } \nu \text{ (4.10))}\\ =&-(q_{\pi,t}(s,a) - r(s,a))^2 + q^2_{\pi,t}(s,a) + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')\hat{q}_{\pi,t+1}(s',a')\\ =&2r(s,a)q_{\pi,t}(s,a) - r^2(s,a) + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')\hat{q}_{\pi,t+1}(s',a')\\ =&\hat{r}_{\pi,t}(s,a) + \sum_{s',a'} p(s'|s,a)\pi_{t+1}(a'|s')\hat{q}_{\pi,t+1}(s',a'), && \text{(By Theorem 4)}\end{aligned}$$

which completes the proof. □

A.1.8 Proof of Theorem 5

Proof. We first derive an important equality. $\forall t$,

$$
\begin{aligned}
&\mathbb{E}_{A_t\sim\hat{\mu}_t^+}\left[{\rho_t^+}^2\hat{q}_{\pi,t}(S_t,A_t)\mid S_t\right] && \text{(A.11)}\\
=&\sum_a \frac{\pi_t^2(a|S_t)}{\hat{\mu}_t^+(a|S_t)}\hat{q}_{\pi,t}(S_t,a)\\
=&\sum_a \frac{\pi_t^2(a|S_t)}{\frac{\pi_t(a|S_t)\sqrt{\hat{q}^+_{\pi,t}(S_t,a)}}{\sum_b \pi_t(b|S_t)\sqrt{\hat{q}^+_{\pi,t}(S_t,b)}}}\hat{q}_{\pi,t}(S_t,a) && \text{(by (4.22))}\\
=&\left[\sum_a \pi_t(a|S_t)\sqrt{\hat{q}^+_{\pi,t}(S_t,a)}\right]\left[\sum_a \pi_t(a|S_t)\frac{\hat{q}_{\pi,t}(S_t,a)}{\sqrt{\hat{q}^+_{\pi,t}(S_t,a)}}\right]\\
=&\left[\sum_a \pi_t(a|S_t)\sqrt{\eta_{\pi,t}(S_t,a)}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]\left[\sum_a \pi_t(a|S_t)\frac{1}{\sqrt{\eta_{\pi,t}(S_t,a)}}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]. && \text{(By (4.21))}
\end{aligned}
$$

We proceed via induction. For $t=T-1$, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\hat{\mu}^+_{t:T-1}})\mid S_t\right)\\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t^+}\left[{\rho_t^+}^2 q^2_{\pi,t}(S_t,A_t)\mid S_t\right]-v^2_{\pi,t}(S_t) && \text{(Lemma 34)}\\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t^+}\left[{\rho_t^+}^2 \hat{q}_{\pi,t}(S_t,A_t)\mid S_t\right]-v^2_{\pi,t}(S_t) && \text{(Definition of } \hat{q} \text{ (4.13))}\\
=&\left[\sum_a \pi_t(a|S_t)\sqrt{\eta_{\pi,t}(S_t,a)}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]\left[\sum_a \pi_t(a|S_t)\frac{1}{\sqrt{\eta_{\pi,t}(S_t,a)}}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]-v^2_{\pi,t}(S_t)\\
& && \text{(By (A.11))}\\
=&\sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a)+\left[\sum_a \pi_t(a|S_t)\sqrt{\eta_{\pi,t}(S_t,a)}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]\\
&\left[\sum_a \pi_t(a|S_t)\frac{1}{\sqrt{\eta_{\pi,t}(S_t,a)}}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]-\sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a)-v^2_{\pi,t}(S_t)\\
=&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}})\mid S_t\right)\\
&-\left(\sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a)-\left[\sum_a \pi_t(a|S_t)\sqrt{\eta_{\pi,t}(S_t,a)}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]\right.\\
&\left.\left[\sum_a \pi_t(a|S_t)\frac{1}{\sqrt{\eta_{\pi,t}(S_t,a)}}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]\right) && \text{(By (A.10) and Lemma 35)}\\
=&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}})\mid S_t\right)-\epsilon_t^+(S_t). && \text{(Definition of } \epsilon^+ \text{ (4.23))}
\end{aligned}
$$

For $t \in [T-2]$, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\hat{\mu}^+_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\hat{\mu}^+_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_{\pi,t}(S_t,A_t) + q^2_{\pi,t}(S_t,A_t)\right) \mid S_t\right] \\
&- v^2_{\pi,t}(S_t) && \text{(Lemma 34)} \\
\leq&\mathbb{E}_{A_t\sim\hat{\mu}^+_t}\Big[\rho_t^2\Big(\mathbb{E}_{S_{t+1}}\Big[\sum_{a'}\pi_{t+1}(a'|S_{t+1})\tilde{q}_{\pi,t+1}(S_{t+1},a') \mid S_t, A_t\Big] + \nu_{\pi,t}(S_t,A_t) \\
&+ q^2_{\pi,t}(S_t,A_t)\Big) \mid S_t\Big] - v^2_{\pi,t}(S_t) - \mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon^+_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
&&\text{(Inductive hypothesis and Lemma 35)} \\
=&\mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\left(\tilde{q}_{\pi,t}(S_t,A_t) + v^2_{\pi,t}(S_t)\right) \mid S_t\right] - v^2_{\pi,t}(S_t) - \mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon^+_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
&&\text{(Definition of } \tilde{q} \text{ (A.9))} \\
=&\mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\hat{q}_{\pi,t}(S_t,A_t) \mid S_t\right] - v^2_{\pi,t}(S_t) - \mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon^+_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
&&\text{(Definition of } \hat{q} \text{ (4.14))} \\
=&\left[\sum_a \pi_t(a|S_t)\sqrt{\eta_{\pi,t}(S_t,a)}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]\left[\sum_a \pi_t(a|S_t)\frac{1}{\sqrt{\eta_{\pi,t}(S_t,a)}}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right] - v^2_{\pi,t}(S_t) \\
&- \mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon^+_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] && \text{(By (A.11))} \\
=&\sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a) - v^2_{\pi,t}(S_t) \\
&+ \left[\sum_a \pi_t(a|S_t)\sqrt{\eta_{\pi,t}(S_t,a)}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right]\left[\sum_a \pi_t(a|S_t)\frac{1}{\sqrt{\eta_{\pi,t}(S_t,a)}}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right] \\
&- \sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a) \\
&- \mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon^+_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right] \\
=&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t\right) + \left[\sum_a \pi_t(a|S_t)\sqrt{\eta_{\pi,t}(S_t,a)}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right] \\
&\left[\sum_a \pi_t(a|S_t)\frac{1}{\sqrt{\eta_{\pi,t}(S_t,a)}}\sqrt{\hat{q}_{\pi,t}(S_t,a)}\right] \\
&- \sum_a \pi_t(a|S_t)\hat{q}_{\pi,t}(S_t,a) - \mathbb{E}_{A_t\sim\hat{\mu}^+_t}\left[\rho_t^2\mathbb{E}_{S_{t+1}}\left[\epsilon^+_{t+1}(S_{t+1}) \mid S_t, A_t\right]\right]] \\
&&\text{(By (A.10) and Lemma 35)} \\
=&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t\right) - \epsilon^+_t(S_t). && \text{(Definition of } \epsilon^+ \text{ (4.23))}
\end{aligned}
$$

□

A.2 Experiment Details

A.2.1 GridWorld

For a Gridworld with size n, its width, height, and time horizon T are all set to n. There are four possible actions: up, down, left, and right. After taking an action, the agent has a 0.9 probability of moving accordingly and a 0.1 probability of moving uniformly at random. If the agent runs into a boundary, the agent stays in its current location. The reward function $r(s, a)$ is randomly generated and fixed after generation. We normalize the rewards across all (s, a) such that $\max_{s,a} r(s, a) = 1$. We consider a set of randomly generated target policies. The ground truth policy performance is estimated using the on-policy Monte Carlo method by running each target policy for 10^6 episodes. We test two different sizes of the Gridworld with a number of $1,000$ and $27,000$ states. The offline dataset contains $m = 10^5$ randomly generated tuples. For a Gridworld of size n, the total amount of possible (s, t, a, r, s') tuples is $n \times n \times n \times 4 \times 4 = 16n^3$. The offline data coverages for the Gridworld of size $1,000$ and $27,000$ are then 62.5% and 2.3%.

We use a one-hot vector representing the position of the agent and a real number representing the current time step as features for the state. We execute Algorithm 1 to approximate function r, q, and $\hat{q}$. As shown in Algorithm 1, we train r using supervised learning by batch stochastic gradient descent. We train q and $\hat{q}$ using fitted Q-learning. We split the offline data into a training set and a test set. We tune all hyperparameters offline based on the supervised learning loss and fitted Q-learning loss on the test set. With the Adam optimizer (Kingma and Ba, 2015), we search the learning rates in $\{2^{-20}, 2^{-18}, \cdots, 2^0\}$ to minimize the loss on the offline data and use the learning rate 2^{-10} on all learning processes. For the behavior policy search (BPS, Hanna et al. (2017)) and robust on-policy sampling (ROS, Zhong et al. (2022)) algorithms, we use the reported parameters from Hanna et al. (2017) and Zhong et al. (2022), since it is not clear how to do hyperparameter turning for BPS and ROS with only offline data.

A.2.2 MuJoCo

Figure A.1 is an introduction to the MuJoCo environments. We construct 150 policies (30 policies in each environment) with a wide range of performance using the proximal policy optimization (PPO) algorithm (Schulman et al., 2017) and the default PPO implementation in Huang et al. (2022). Since our methods are designed for discrete action space, we discretize the first dimension of MuJoCo action space

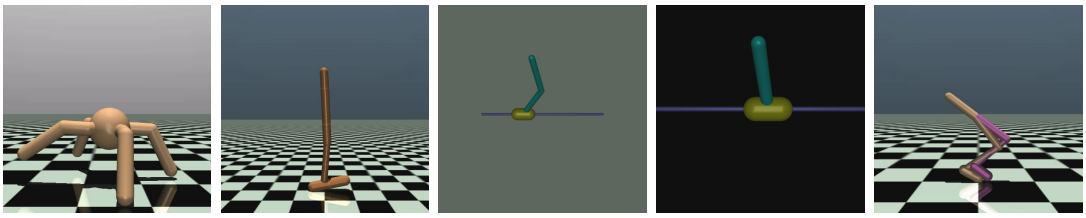

Figure A.1: MuJoCo (Todorov et al., 2012) robot simulation tasks. MuJoCo is a physics engine for robotics simulation and contains various stochastic environments. The goal in each environment is to control a robot to achieve different behaviors such as walking, jumping, and balancing. Environments from the left to the right are Ant, Hopper, InvertedDoublePendulum, InvertedPendulum, and Walker. We conducted experiments on those five environments with results reported in Section 4.5.

in our experiments. The remaining dimensions are controlled by the PPO policy and are deemed as part of the environment. We run each compared algorithm 30 times for each policy and compute the average and standard error to plot curves in Figure 4.2. To generate offline data, we add different levels of noise to the target policy and run noisy target policies for 2000 episodes. The noise is in the form of a uniformly random policy, and its weight is uniformly randomly sampled from $(0, 0.1]$. This data generation process simulates the data generated during the training of a policy. Notably, compared with previous works, we do not need data to be complete trajectories or generated by known policies. We leave the investigation of entirely irrelevant offline data in the MuJoCo domain for future work. Our algorithm is robust on hyperparameters. All learning rates in Algorithm 1 are tuned offline and are the same 2^{-10} across all MuJoCo and Gridworld experiments.

In MuJcCo, the episode length varies because of stochasticity in policies and environments. Because the length of each episode is not fixed, episodes in off-policy estimation may be longer than episodes in on-policy estimation. In the main text, we use episodes instead of steps as the x-axis mainly to improve readability. Because after running 100 steps, we might already have a good estimate for a target policy with a length of 10 but may still not finish a single episode for a target policy with a length of 250. Due to the diversity of our target policies, averaging using steps as the x-axis makes the plot conceptually hard to interpret.

We anyway show the figure with steps as the x-axis in Figure A.2. Setting steps as the x-axis, we linearly interpolate the estimation error across episodes. At each step, we average the estimation error for all tests that have completed the first episode and, thus, have an estimate. The estimation error is divided by the first estimate of the on-policy estimation to get the normalized estimation error. Although the normalized

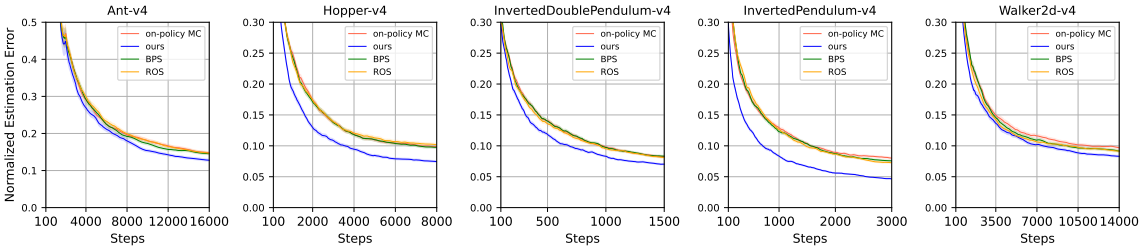

Figure A.2: MuJoCo results using steps as the x-axis. We draw each curve from step 100 because some policies need more than 100 steps to finish the first episode. All curves are averaged over 900 trials (30 target policies, each having 30 independent runs). The shaded regions denote standard errors and are invisible because they are too small.

estimation error for the on-policy estimation starts from 1, it may be unstable until around 1000 steps because different policies get the first estimate at different steps. However, it is still clear that our off-policy estimator achieves the same accuracy with fewer online steps.

Appendix B

Appendix for Chapter 6

B.1 Proofs

B.1.1 Proof of Lemma 8

Proof. Given any baseline function b, $\forall \mu \in \Lambda^-$, $\forall t, s, a$,

$$\begin{aligned} & \mu_t(a|s) = 0 \\ \implies & \pi_t(a|s) = 0 && \text{(Definition of } \Lambda^-\text{)} \\ \implies & \pi_t(a|s) u_{\pi,t}(s,a) = 0. \end{aligned}$$

This shows $\mu \in \Lambda$. Thus, $\Lambda^- \subseteq \Lambda$. □

B.1.2 Proof of Lemma 9

To prove Lemma 9, we first prove the following auxiliary lemma.

Lemma 36. $\forall b, \forall \mu \in \Lambda, \forall t, s,$

$$\mathbb{E}_{A_t \sim \mu_t(\cdot|S_t)} \left[\rho_t [q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t = s \right] = \mathbb{E}_{A_t \sim \pi_t(\cdot|S_t)} \left[q_{\pi,t}(S_t, A_t) \mid S_t = s \right].$$

Proof. We fix any baseline function b. Because $\mu \in \Lambda$, $\forall t, s, a$,

$$\begin{aligned} & \mu_t(a|s) = 0 \\ \implies & \pi_t(a|s) u_{\pi,t}(s,a) = 0 \\ \implies & \pi_t(a|s) \Big[(q_{\pi,t}(s,a) - b_t(s,a))^2 + \nu_{\pi,t}(s,a) + \textstyle\sum_{s'} p(s'|s,a) \mathbb{V} \left(G^b(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1} = s' \right) \Big] = 0 \\ & && \text{(By (6.13))} \\ \implies & \pi_t(a|s) (q_{\pi,t}(s,a) - b_t(s,a))^2 = 0 \\ & && (\nu_{\pi,t}(s,a) \text{ and } \textstyle\sum_{s'} p(s'|s,a) \mathbb{V} \left(G^b(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1} = s' \right) \text{ are non-negative}) \\ \implies & \pi_t(a|s) (q_{\pi,t}(s,a) - b_t(s,a)) = 0. && \text{(B.1)} \end{aligned}$$

Then, we have

$$
\begin{aligned}
&\mathbb{E}_{A_t\sim\mu_t(\cdot|S_t)}\left[\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)]+\bar{b}_t(S_t)\mid S_t=s\right]\\
=&\mathbb{E}_{A_t\sim\mu_t(\cdot|S_t)}\left[\frac{\pi_t(A_t|S_t)}{\mu_t(A_t|S_t)}[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)]+\bar{b}_t(S_t)\mid S_t=s\right]\\
=&\sum_{a\in\{a|\mu_t(a|s)>0\}}\mu_t(a|s)\left[\frac{\pi_t(a|s)}{\mu_t(a|s)}[q_{\pi,t}(s,a)-b_t(s,a)]+\bar{b}_t(s)\right]\\
=&\sum_{a\in\{a|\mu_t(a|s)>0\}}\pi_t(a|s)[q_{\pi,t}(s,a)-b_t(s,a)]+\sum_{a\in\{a|\mu_t(a|s)>0\}}\mu_t(a|s)\bar{b}_t(s)\\
=&\sum_{a\in\{a|\mu_t(a|s)>0\}}\pi_t(a|s)[q_{\pi,t}(s,a)-b_t(s,a)]+\bar{b}_t(s)\sum_{a\in\{a|\mu_t(a|s)>0\}}\mu_t(a|s)\\
=&\sum_{a\in\{a|\mu_t(a|s)>0\}}\pi_t(a|s)[q_{\pi,t}(s,a)-b_t(s,a)]+\bar{b}_t(s)\\
=&\sum_{a\in\{a|\mu_t(a|s)>0\}}\pi_t(a|s)[q_{\pi,t}(s,a)-b_t(s,a)]\\
&+\sum_{a\in\{a|\mu_t(a|s)=0\}}\pi_t(a|s)[q_{\pi,t}(s,a)-b_t(s,a)]+\bar{b}_t(s) && \text{(By (B.1))}\\
=&\sum_a\pi_t(a|s)[q_{\pi,t}(s,a)-b_t(s,a)]+\bar{b}_t(s)\\
=&\sum_a\pi_t(a|s)q_{\pi,t}(s,a)-\bar{b}_t(s)+\bar{b}_t(s) && \text{(By (6.2))}\\
=&\mathbb{E}_{A\sim\pi}[q_{\pi,t}(S_t,A_t)\mid S_t=s].
\end{aligned}
$$

□

Now, we are ready to prove Lemma 9.

Lemma 9 (Unbiasedness)**.** $\forall b, \forall \mu\in\Lambda, \forall t, \forall s,\ \mathbb{E}\left[G^b(\tau_{t:T-1}^{\mu_{t:T-1}})\mid S_t=s\right]=v_{\pi,t}(s)$.

Proof. Fix any baseline function b. We proceed via induction.

For $t=T-1$, $\forall\mu\in\Lambda$, $\forall s$, we have

$$
\begin{aligned}
&\mathbb{E}\left[G^b(\tau_{t:T-1}^{\mu_{t:T-1}})\mid S_t=s\right]\\
=&\mathbb{E}_{A_t\sim\mu_t(\cdot|S_t)}\left[\rho_t[R_{t+1}-b_t(S_t,A_t)]+\bar{b}_t(S_t)\mid S_t\right]\\
=&\mathbb{E}_{A_t\sim\mu_t(\cdot|S_t)}\left[\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)]+\bar{b}_t(S_t)\mid S_t\right]\\
=&\mathbb{E}_{A_t\sim\pi_t(\cdot|S_t)}\left[q_{\pi,t}(S_t,A_t)\mid S_t\right] && \text{(Lemma 36)}\\
=&v_{\pi,t}(S_t).
\end{aligned}
$$

For $t\in[T-2]$, assuming that Lemma 9 holds for $t+1$, we have $\forall\mu\in\Lambda$, $\forall s$,

$$
\mathbb{E}\left[G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}})\mid S_{t+1}=s\right]=v_{\pi,t+1}(s).
$$

Then, $\forall t$,

$$
\begin{aligned}
&\mathbb{E}\left[G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right] \\
=&\mathbb{E}\left[\rho_t\left(R_{t+1} + G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) - b_t(S_t, A_t)\right) + \bar{b}_t(S_t)) \mid S_t\right] && \text{(By (6.1))} \\
=&\mathbb{E}\left[\rho_t(R_{t+1} - b_t(S_t, A_t)) + \bar{b}_t(S_t) \mid S_t\right] + \mathbb{E}\left[\rho_t G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t(R_{t+1} - b_t(S_t, A_t)) + \bar{b}_t(S_t) \mid S_t\right] \\
&+ \mathbb{E}_{A_t\sim\mu_t(\cdot|S_t), S_{t+1}\sim p(\cdot|S_t,A_t)}\left[\mathbb{E}\left[\rho_t G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right] \mid S_t\right] \\
& && \text{(Law of Iterated Expectation)} \\
=&\mathbb{E}\left[\rho_t(R_{t+1} - b_t(S_t, A_t)) + \bar{b}_t(S_t) \mid S_t\right] \\
&+ \mathbb{E}_{A_t\sim\mu_t(\cdot|S_t), S_{t+1}\sim p(\cdot|S_t,A_t)}\left[\rho_t \mathbb{E}\left[G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right] \mid S_t\right] \\
& && \text{(Conditional independence and Markov property)} \\
=&\mathbb{E}\left[\rho_t(R_{t+1} - b_t(S_t, A_t)) + \bar{b}_t(S_t) \mid S_t\right] + \mathbb{E}_{A_t\sim\mu_t(\cdot|S_t), S_{t+1}\sim p(\cdot|S_t,A_t)}\left[\rho_t v_{\pi,t+1}(S_{t+1}) \mid S_t\right] \\
& && \text{(Inductive hypothesis)} \\
=&\mathbb{E}_{A_t\sim\mu_t(\cdot|S_t)}\left[\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\right] && \text{(Definition of } q_{\pi,t}) \\
=&\mathbb{E}_{A_t\sim\pi_t(\cdot|S_t)}\left[q_{\pi,t}(S_t, A_t) \mid S_t\right] && \text{(Lemma 36)} \\
=&v_{\pi,t}(s),
\end{aligned}
$$

which completes the proof. □

B.1.3 Proof of Theorem 10

To prove Theorem 10, we first characterize the variance of the off-policy estimator in a recursive form.

Lemma 37. $\forall b$, $\forall \mu \in \Lambda$, *for* $t = T-1$,

$$
\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) = \mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2;
$$

For $t \in [T-2]$,

$$
\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right.\right. \\
&\left.\left.+[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2\right) \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2.
\end{aligned}
$$

Proof. When $t \in [T-2]$, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\mathbb{E}\left[G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t, A_t\right] \mid S_t\right) \\
& \qquad \text{(Law of total variance)} \\
=&\mathbb{E}_{A_t}\left[\mathbb{V}\left(\rho_t\left[r(S_t, A_t) + G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) - b_t(S_t, A_t)\right] + \bar{b}_t(S_t) \mid S_t, A_t\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\left(\mathbb{E}\left[\rho_t\left[r(S_t, A_t) + G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) - b_t(S_t, A_t)\right] + \bar{b}_t(S_t)\right) \mid S_t, A_t\right] \mid S_t\right) \\
& \qquad \text{(By (6.1))} \\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\left(\rho_t[r(S_t, A_t) + \mathbb{E}\left[G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right] - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\right) \\
& \qquad (r(S_t, A_t), b_t(S_t, A_t), \bar{b}_t(S_t) \text{ are constant given } S_t, A_t) \\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\left(\rho_t[r(S_t, A_t) + \mathbb{E}\left[v_{\pi,t+1}(S_{t+1}) \mid S_t, A_t\right] - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\right) \\
& \qquad \text{(Lemma 9)} \\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\left(\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\right). \qquad \text{(Defintion of } q_{\pi,t} \text{(B.2)}
\end{aligned}
$$

For the inner part of the first term, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right) \mid S_t, A_t\right] \\
&+ \mathbb{V}_{S_{t+1}}\left(\mathbb{E}\left[G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right] \mid S_t, A_t\right) \qquad \text{(Law of total variance)} \\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}\left(\mathbb{E}\left[G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right] \mid S_t, A_t\right) \\
& \qquad \text{(Markov property)} \\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}\left(v_{\pi,t+1}(S_{t+1}) \mid S_t, A_t\right) \quad \text{(Lemma 9)} \\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t). \qquad \text{(By (6.11))} \qquad \text{(B.3)}
\end{aligned}
$$

For the second term, we have

$$
\begin{aligned}
&\nu_t(S_t, A_t) \\
=&\mathbb{V}_{A_t}\big(\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\big) && \text{(By (6.11))} \\
=&\mathbb{E}_{A_t}\left[\big(\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t)\big)^2 \mid S_t\right] \\
&- \left(\mathbb{E}_{A_t}\left[\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\right]\right)^2 \\
=&\mathbb{E}_{A_t}\left[\big(\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t)\big)^2 \mid S_t\right] - v_{\pi,t}(S_t)^2. && \text{(Lemma 36)} \\
=&\mathbb{E}_{A_t}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] + 2\bar{b}_t(S_t)\mathbb{E}_{A_t}\left[\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] \mid S_t\right] \\
&+ \bar{b}_t(S_t)^2 - v_{\pi,t}(S_t)^2 \\
=&\mathbb{E}_{A_t}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] + 2\bar{b}_t(S_t)\mathbb{E}_{A_t}\left[\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\right] \\
&- 2\bar{b}_t(S_t)^2 + \bar{b}_t(S_t)^2 - v_{\pi,t}(S_t)^2 \\
=&\mathbb{E}_{A_t}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] + 2\bar{b}_t(S_t)v_{\pi,t}(S_t) \\
&- \bar{b}_t(S_t)^2 - v_{\pi,t}(S_t)^2 && \text{(Lemma 9)} \\
=&\mathbb{E}_{A_t}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2. && \text{(B.4)}
\end{aligned}
$$

Plugging (B.3) and (B.4) back to (B.2) gives

$$
\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\big(\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\big) && \text{(By (B.2))} \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\big(\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\big) && \text{(By (B.3))} \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{E}_{A_t}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 && \text{(By (B.4))} \\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^*}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) + [q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2\right) \mid S_t\right] \\
&- [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2.
\end{aligned}
$$

When $t = T - 1$, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) \\
=&\mathbb{V}\left(\rho_t[r(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\right) && \text{(By (6.1))} \\
=&\mathbb{V}\left(\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] + \bar{b}_t(S_t) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2,
\end{aligned}
$$

which completes the proof. □

We now restate Theorem 10 and give its proof.

Theorem 10. *For a baseline function b, the behavior policy μ^* defined in (6.12) is an optimal solution to the optimization problems $\forall t, s$,*

$$\begin{aligned} \min_{\mu} \quad & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\ s.t. \quad & \mu \in \Lambda. \end{aligned}$$

Proof. Fix a baseline function b. $\forall t, s, a$,

$$\begin{aligned} & \mu_t^*(a|s) = 0 \\ \Longrightarrow\ & \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)} = 0 \qquad \text{(By (6.12))} \\ \Longrightarrow\ & \pi_t(a|s)u_{\pi,t}(s,a) = 0. \end{aligned}$$

Thus, $\mu^* \in \Lambda$.

$\forall t$, $\forall \mu \in \Lambda$, we have an unbiasedness on $\sqrt{u_{\pi,t}(s,a)}$.

$$\begin{aligned} & \mathbb{E}_{A_t\sim\mu_t}\left[\rho_t\sqrt{u_{\pi,t}(S_t,A_t)} \mid S_t = s\right] \\ =& \sum_{a\in\{a|\mu_t(a|s)>0\}} \mu_t(a|s)\frac{\pi_t(a|s)}{\mu_t(a|s)}\sqrt{u_{\pi,t}(s,a)} \\ =& \sum_{a\in\{a|\mu_t(a|s)>0\}} \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)} \\ =& \sum_{a\in\{a|\mu_t(a|s)>0\}} \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)} + \sum_{a\in\{a|\mu_t(a|s)=0\}} \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)} \\ & \qquad (\forall \mu \in \Lambda, \mu_t(a|s) = 0 \implies \pi_t(a|s)u_{\pi,t}(s,a) = 0 \text{ by (6.14)}) \\ =& \mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_{\pi,t}(S_t,A_t)} \mid S_t = s\right]. \end{aligned} \tag{B.5}$$

We prove the optimality of the behavior policy μ^* via induction.

When $t = T-1$, $\forall \mu \in \Lambda$, $\forall s$, the variance of the off-policy estimator has the

following lower bound

$$
\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\
=&\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^2[q_{\pi,t}(S_t,A_t) - b_t(S_t,A_t)]^2 \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 && \text{(Lemma 37)} \\
=&\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^2 u_{\pi,t}(S_t,A_t) \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 && \text{(By (6.13))} \\
\geq&\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t\sqrt{u_{\pi,t}(S_t,A_t)} \mid S_t\right]^2 - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 && \text{(By Jensen's Inequality)} \\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_{\pi,t}(S_t,A_t)} \mid S_t\right]^2 - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2. && \text{(By (B.5))}
\end{aligned}
$$

For any state s, the variance of the off-policy estimator with the behavior policy μ^* achieves this lower bound by the following derivations.

$$
\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) && \text{(B.6)} \\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2[q_{\pi,t}(S_t,A_t) - b_t(S_t,A_t)]^2 \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 && \text{(Lemma 37)} \\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2 u_{\pi,t}(S_t,A_t) \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2. && \text{(By (6.13))}
\end{aligned}
$$

For the first term, we have

$$
\begin{aligned}
&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2 u_{\pi,t}(S_t,A_t) \mid S_t\right] \\
=&\sum_a \frac{\pi_t(a|S_t)^2}{\mu_t^*(a|S_t)} u_{\pi,t}(S_t,a) \\
=&\sum_a \pi_t(a|S_t)\sqrt{u_{\pi,t}(S_t,a)}\sum_b \pi_t(S_t,b)\sqrt{u_{\pi,t}(S_t,b)} && \text{(By (6.12))} \\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_{\pi,t}(S_t,A_t)} \mid S_t\right]^2. && \text{(B.7)}
\end{aligned}
$$

Plugging (B.7) back to (B.6), we obtain

$$
\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) \\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2 u_{\pi,t}(S_t,A_t) \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 \\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_{\pi,t}(S_t,A_t)} \mid S_t\right]^2 - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2.
\end{aligned}
$$

Thus, the behavior policy μ^* defined in (6.12) is an optimal solution to the optimization problems

$$
\begin{aligned}
\min_{\mu} \quad & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\
\text{s.t.} \quad & \mu \in \Lambda
\end{aligned}
$$

for $t = T-1$ and all s.

When $t \in [T-2]$, we proceed via induction. The inductive hypothesis is that the behavior policy μ^* is an optimal solution to the optimization problems

$$\begin{aligned} \min_{\mu} \quad & \mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t = s\right) \\ \text{s.t.} \quad & \mu \in \Lambda \end{aligned}$$

for all s.

To complete the induction, we prove that the behavior policy μ^* is an optimal solution to the optimization problems

$$\begin{aligned} \min_{\mu} \quad & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\ \text{s.t.} \quad & \mu \in \Lambda \end{aligned}$$

for all s.

$\forall \mu \in \Lambda$, $\forall s$, the variance of the off-policy estimator has the following lower bound

$$\begin{aligned} & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\ =& \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^2 \left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) + [q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2\right) \mid S_t\right] \\ & - [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2 && \text{(Lemma 37)} \\ \geq & \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^2 \left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^*}) \mid S_{t+1}\right)\right] \mid S_t, A_t\right]\right. \\ & \left. + \nu_t(S_t, A_t) + [q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2\right) \mid S_t\Big] - [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2 \\ & && \text{(Indutive Hypothesis)} \\ =& \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^2 u_{\pi,t}(S_t, A_t) \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 && \text{(By (6.13))} \\ \geq & \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t \sqrt{u_{\pi,t}(S_t, A_t)} \mid S_t\right]^2 - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 && \text{(By Jensen's Inequality)} \\ =& \mathbb{E}_{A_t \sim \pi_t}\left[\sqrt{u_{\pi,t}(S_t, A_t)} \mid S_t\right]^2 - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2. && \text{(By (B.5))} \end{aligned}$$

For any state s, the variance of the off-policy estimator with the behavior policy μ^* achieves the lower bound by the following derivations.

$$\begin{aligned} & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}^*}) \mid S_t = s\right) \\ =& \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^2 \left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^*}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right.\right. \\ & \left.\left. + [q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2\right) \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2 && \text{(Lemma 37)} \\ =& \mathbb{E}_{A_t \sim \mu_t^*}\left[\rho_t^2 u_{\pi,t}(S_t, A_t) \mid S_t\right] - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2 && \text{(By (6.13))} \\ =& \mathbb{E}_{A_t \sim \pi_t}\left[\sqrt{u_{\pi,t}(S_t, A_t)} \mid S_t\right]^2 - (v_{\pi,t}(S_t) - \bar{b}_t(S_t))^2. && \text{(By (B.7))} \end{aligned}$$

Thus, the behavior policy μ^* defined in (6.12) is an optimal solution to the optimization problems

$$\begin{aligned} \min_{\mu} \quad & \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t = s\right) \\ \text{s.t.} \quad & \mu \in \Lambda \end{aligned}$$

for t and all s.

This completes the induction. □

B.1.4 Proof of Theorem 11

In this proof, to differentiate different μ^* with different baseline functions b, we use $\mu^{*,b}$ to denote the corresponding μ^* when using a function b as the baseline function. G^b, $u_{\pi,t}^b$, and Λ^b are defined following the same convention. We first present an auxiliary lemma.

Lemma 38. $\forall b$, $\forall \mu \in \Lambda^b$, $\forall t$,

$$\begin{aligned} &\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^2[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t)-\bar{b}_t(S_t)]^2 \\ =&\mathbb{V}_{A_t\sim\mu_t}\left(\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right) \end{aligned} \tag{B.8}$$

Proof. $\forall b$, $\forall \mu \in \Lambda^b$, $\forall t$,

$$\begin{aligned} &\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^2[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t)-\bar{b}_t(S_t)]^2 \\ =&\mathbb{V}_{A_t\sim\mu_t}\left(\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right) + \mathbb{E}_{A_t\sim\mu_t}\left[\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right]^2 \\ &- [v_{\pi,t}(S_t)-\bar{b}_t(S_t)]^2 \\ =&\mathbb{V}_{A_t\sim\mu_t}\left(\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right) + \mathbb{E}_{A_t\sim\mu_t}\left[\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right]^2 \\ &- \left[\mathbb{E}_{A_t\sim\mu_t(\cdot|S_t)}\left[\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)]+\bar{b}_t(S_t) \mid S_t\right] - \bar{b}_t(S_t)\right]^2 \\ &\qquad\qquad \text{(Definition of } q_{\pi,t}\text{, Lemma 36)} \\ =&\mathbb{V}_{A_t\sim\mu_t}\left(\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right) + \mathbb{E}_{A_t\sim\mu_t}\left[\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right]^2 \\ &- \mathbb{E}_{A_t\sim\mu_t(\cdot|S_t)}\left[\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right]^2 \\ =&\mathbb{V}_{A_t\sim\mu_t}\left(\rho_t[q_{\pi,t}(S_t,A_t)-b_t(S_t,A_t)] \mid S_t\right). \end{aligned}$$

□

We now restate Theorem 11 and give its proof.

Theorem 11. *b^* is the optimal solution to the optimization problems $\forall t, s$,*

$$\min_{b} \quad \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right). \tag{6.16}$$

Proof. We prove this by induction on the time step t.

When $t = T-1$, $\forall s$, the optimization problem (6.16) has the following lower bound

$$\begin{aligned}
&\mathbb{V}(G^b(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b}}) \mid S_0 = s) \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t^2[q_{\pi,t}(S_t,A_t) - b_t(S_t,A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2 && \text{(Lemma 37)} \\
=&\mathbb{V}_{A_t\sim\mu_t^{*,b}}\left(\rho_t[q_{\pi,t}(S_t,A_t) - b_t(S_t,A_t)] \mid S_t\right) && \text{(Lemma 38)} \\
\geq&0. && \text{(Variance non-negativity)}
\end{aligned}$$

When using b^* as the baseline, we achieve this lower bound.

$$\begin{aligned}
&\mathbb{V}(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b^*}}) \mid S_0 = s) \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2[q_{\pi,t}(S_t,A_t) - b_t^*(S_t,A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b}_t^*(S_t)]^2 && \text{(Lemma 37)} \\
=&\mathbb{V}_{A_t\sim\mu_t^{*,b^*}}\left(\rho_t[q_{\pi,t}(S_t,A_t) - b_t^*(S_t,A_t)] \mid S_t\right) && \text{(Lemma 38)} \\
=&\mathbb{V}_{A_t\sim\mu_t^{*,b^*}}\left(\rho_t[q_{\pi,t}(S_t,A_t) - q_{\pi,t}(S_t,A_t)] \mid S_t\right) && \text{(Definition of } b^* \text{ (6.15))} \\
=&0.
\end{aligned}$$

When $t \in [T-2]$, we proceed via induction. The inductive hypothesis is that the baseline function b^* is an optimal solution to the optimization problems

$$\min_b \quad \mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b}}) \mid S_t = s\right)$$

for all s. Notice that we have

$$\Lambda^b \subseteq \Lambda^{b^*}. \tag{B.9}$$

This is because $\forall s, a$,

$$\begin{aligned}
&u_{\pi,t}^b(s,a) \\
=&(q_{\pi,t}(s,a) - b_t(s,a))^2 + \nu_{\pi,t}(s,a) + \textstyle\sum_{s'} p(s'|s,a)\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b}}) \mid S_{t+1} = s'\right) \\
& && \text{(By (6.13))} \\
\geq&\nu_{\pi,t}(s,a) + \textstyle\sum_{s'} p(s'|s,a)\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b}}) \mid S_{t+1} = s'\right) \\
\geq&\nu_{\pi,t}(s,a) + \textstyle\sum_{s'} p(s'|s,a)\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1} = s'\right) && \text{(Inductive Hypothesis)} \\
\geq&u_{\pi,t}^{b^*}(s,a).
\end{aligned}$$

Thus, $\forall \mu \in \Lambda^b$, we have $\forall s, a$

$$\begin{aligned}
&\mu(a|s) = 0 \\
\implies &\pi(a|s)u_{\pi,t}^b(s,a) = 0 \\
\implies &\pi(a|s)u_{\pi,t}^{b^*}(s,a) = 0.
\end{aligned}$$

This shows

$$\Lambda^b \subseteq \Lambda^{b^*}.$$

$\forall b$, the optimization problem (6.16) has the following lower bound

$$\begin{aligned}
&\mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b}}) \mid S_0 = s\right) \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2 \quad \text{(Lemma 37)} \\
\geq&\mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2 \\
&\qquad\qquad \text{(Inductive hypothesis)} \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t\sim\mu_t^{*,b}}\left(\rho_t[q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)] \mid S_t\right) \quad \text{(Lemma 38)} \\
\geq&\mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&\qquad\qquad \text{(Variance non-negativity)} \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t^2 u_{\pi,t}^{b^*}(S_t, A_t) \mid S_t\right] \quad \text{(By (6.13))} \\
\geq&\mathbb{E}_{A_t\sim\mu_t^{*,b}}\left[\rho_t\sqrt{u_{\pi,t}^{b^*}(S_t, A_t)} \mid S_t\right]^2 \quad \text{(Jensen's inequality)} \\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_{\pi,t}^{b^*}(S_t, A_t)} \mid S_t\right]^2 \quad \text{(By (B.5) and (B.9))}
\end{aligned}$$

When setting $\forall s, \forall a, b_t^*(s, a) \doteq q_{\pi,t}(s, a)$ as the baseline, we achieve this lower

bound.

$$
\begin{aligned}
&\mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b^*}}) \mid S_0 = s\right) \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2[q_{\pi,t}(S_t, A_t) - b_t^*(S_t, A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b}_t^*(S_t)]^2 \quad \text{(Lemma 37)} \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t\sim\mu_t^{*,b^*}}\left(\rho_t[q_{\pi,t}(S_t, A_t) - b_t^*(S_t, A_t)] \mid S_t\right) \quad \text{(Lemma 38)} \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t\sim\mu_t^{*,b^*}}\left(\rho_t[q_{\pi,t}(S_t, A_t) - q_{\pi,t}(S_t, A_t)] \mid S_t\right) \quad \text{(Definition of } b^* \text{ (6.15))} \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2 u_{\pi,t}^{b^*}(S_t, A_t) \mid S_t\right] \quad \text{(By (6.13))} \\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_{\pi,t}^{b^*}(S_t, A_t)} \mid S_t\right]^2 \quad \text{(By (6.12))}
\end{aligned}
$$

Thus, b^* is the optimal solution to the optimization problem

$$\min_b \quad \mathbb{V}\left(G^b(\tau_{t:T-1}^{\mu_{t:T-1}^*}) \mid S_t = s\right)$$

for all t and s.

□

B.1.5 Proof of Theorem 13

Proof. Use $u_t^{b^*}$ to denote u_t (6.13) using b^* as the baseline function. Then, by (6.13), for $t = T-1$,

$$u_t^{b^*}(s, a) = [q_{\pi,t}(s, a) - b_t(s, a)]^2 = 0. \tag{B.10}$$

For $t \in [T-2]$,

$$
\begin{aligned}
&u_t^{b^*}(s, a) \\
=&(q_{\pi,t}(s, a) - b_t(s, a))^2 + \nu_{\pi,t}(s, a) + \textstyle\sum_{s'} p(s'|s, a)\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1} = s'\right) \\
&\qquad \text{(By (6.13))} \\
=&\nu_{\pi,t}(s, a) + \textstyle\sum_{s'} p(s'|s, a)\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1} = s'\right). \quad \text{(By (6.15))} \quad \text{(B.11)}
\end{aligned}
$$

The variance of $G^{b^*}(\tau_{t:T-1}^{\mu^*_{t:T-1}})$ has $\forall s$, for $t = T-1$,

$$\begin{aligned}
&\mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) && \text{(B.12)}\\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2[q_{\pi,t}(S_t,A_t) - b_t^*(S_t,A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b^*}_t(S_t)]^2 && \text{(Lemma 37)}\\
=&0 && \text{(Definition of } b^* \text{ (6.15))}\\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2 u_t^{b^*}(S_t,A_t) \mid S_t\right] && \text{(By (B.10))}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_t^{b^*}(S_t,A_t)} \mid S_t\right]^2. && \text{(By (B.7))}
\end{aligned}$$

For $t \in [T-2]$,

$$\begin{aligned}
&\mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) && \text{(B.13)}\\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}})|S_{t+1}\right)|S_t,A_t\right] + \nu_t(S_t,A_t) + [q_{\pi,t}(S_t,A_t) - b_t(S_t,A_t)]^2\right)|S_t\right]\\
& && \text{(Lemma 37)}\\
&- [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2\\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t,A_t\right] + \nu_t(S_t,A_t)\right) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\mu_t^*}\left(\rho_t[q_{\pi,t}(S_t,A_t) - b_t^*(S_t,A_t)] \mid S_t\right) && \text{(Lemma 38)}\\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t,A_t\right] + \nu_t(S_t,A_t)\right) \mid S_t\right]\\
& && \text{(By (6.15))}\\
=&\mathbb{E}_{A_t\sim\mu_t^*}\left[\rho_t^2 u_t^{b^*}(S_t,A_t) \mid S_t\right] && \text{(By (B.11))}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_t^{b^*}(S_t,A_t)} \mid S_t\right]^2. && \text{(By (B.7))}
\end{aligned}$$

The variance of $G^{\text{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}})$ has $\forall s$, for $t - T - 1$,

$$\begin{aligned}
&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) && \text{(B.14)}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[q_{\pi,t}(S_t,A_t)^2 \mid S_t\right] - v_{\pi,t}(S_t)^2 && \text{(Lemma 37 with } b=0 \text{ and on-policy)}\\
=&\mathbb{V}_{A_t\sim\pi_t}\left(q_{\pi,t}(S_t,A_t) \mid S_t\right) && \text{(Lemma 38 with } b=0 \text{ and on-policy)}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[u_t^{b^*}(S_t,A_t) \mid S_t\right] + \mathbb{V}_{A_t\sim\pi_t}\left(q_{\pi,t}(S_t,A_t) \mid S_t\right). && \text{(By (B.11))}
\end{aligned}$$

For $t \in [T-2]$,

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) && \text{(B.15)}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) + q_{\pi,t}(S_t, A_t)^2 \mid S_t\right]\\
&- v_{\pi,t}(S_t)^2 && \text{(Lemma 37 with } b=0)\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\pi_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t\right) && \text{(Lemma 38)}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\pi_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t\right)\\
&+ \mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] \mid S_t\right]\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[u_t^{b^*}(S_t, A_t) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\pi_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t\right)\\
&+ \mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] \mid S_t\right]. \\
& && \text{(By (B.11))}
\end{aligned}
$$

Thus, for $t = T-1$, their difference is

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^{*}}) \mid S_t = s\right)\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[u_t^{b^*}(S_t, A_t) \mid S_t\right] - \mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t\right]^2\\
&+ \mathbb{V}_{A_t\sim\pi_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t\right) && \text{(By (B.12) and (B.14))}\\
=&\mathbb{V}_{A_t\sim\pi_t}\left(\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t\right) + \mathbb{V}_{A_t\sim\pi_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t\right).
\end{aligned}
$$

For $t \in [T-2]$,

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) \\
=&\mathbb{E}_{A_t \sim \pi_t}\left[u_t^{b^*}(S_t, A_t) \mid S_t\right] - \mathbb{E}_{A_t \sim \pi_t}\left[\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t\right]^2 \\
&+ \mathbb{V}_{A_t \sim \pi_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t\right) \\
&+ \mathbb{E}_{A_t \sim \pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] \mid S_t\right] \\
&\qquad\qquad \text{(By (B.13) and (B.15))} \\
=&\mathbb{V}_{A_t \sim \pi_t}\left(\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t\right) \\
&+ \mathbb{V}_{A_t \sim \pi_t}\left(q_{\pi,t}(S_t, A_t) \mid S_t\right) \\
&+ \mathbb{E}_{A_t \sim \pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] \mid S_t\right].
\end{aligned}
$$

We use induction to prove $\forall t, s, \delta_t^{\mathrm{ON,\ ours}}(s) \geq 0$. For $t = T-1$,

$$\delta_t^{\mathrm{ON,\ ours}}(s) = 0 \geq 0.$$

For $t \in [T-2]$, the induction hypothesis is $\forall s$,

$$\delta_{t+1}^{\mathrm{ON,\ ours}}(s) \geq 0.$$

This implies $\forall s$,

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1} = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1} = s\right) \\
=&\mathbb{V}_{A_{t+1} \sim \pi_{t+1}}\left(\sqrt{u_{t+1}^{b^*}(S_{t+1}, A_{t+1})} \mid S_{t+1} = s\right) \\
&+ \mathbb{V}_{A_{t+1} \sim \pi_{t+1}}\left(q_{\pi,t+1}(S_{t+1}, A_{t+1}) \mid S_{t+1} = s\right) + \delta_{t+1}^{\mathrm{ON,\ ours}}(s) \\
\geq&0. \qquad\qquad \text{(B.16)}
\end{aligned}
$$

Thus, $\forall s$,

$$
\begin{aligned}
&\delta_t^{\mathrm{ON,\ ours}}(s) \\
=&\mathbb{E}_{A_t \sim \pi_t, S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t = s\right] \\
\geq&0. \qquad\qquad \text{(by (B.16))} \quad \text{(B.17)}
\end{aligned}
$$

Thus, $\forall t, s, \delta_t^{\mathrm{ON,\ ours}}(s) \geq 0$.

□

B.1.6 Proof of Theorem 14

Proof. The variance of $G^{\text{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,\text{PDIS}}})$ has $\forall s$, for $t = T-1$,

$$
\begin{aligned}
&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,\text{PDIS}}}) \mid S_t = s\right) && \text{(B.18)}\\
=&\mathbb{E}_{A_t\sim\mu_t^{*,\text{PDIS}}}\left[\rho_t^2 q_{\pi,t}(S_t,A_t)^2 \mid S_t\right] - v_{\pi,t}(S_t)^2 && \text{(Lemma 37 and } b=0)\\
=&\mathbb{V}_{A_t\sim\mu_t^{*,\text{PDIS}}}\left(\rho_t q_{\pi,t}(S_t,A_t) \mid S_t\right) && \text{(By (B.8))}\\
=&\mathbb{E}_{A_t\sim\mu_t^{*,\text{PDIS}}}\left[\rho_t^2 u_t^{b^*}(S_t,A_t) \mid S_t\right] + \mathbb{V}_{A_t\sim\mu_t^{*,\text{PDIS}}}\left(\rho_t q_{\pi,t}(S_t,A_t) \mid S_t\right) && \text{(By (B.11))}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_t^{b^*}(S_t,A_t)} \mid S_t\right]^2 + \mathbb{V}_{A_t\sim\mu_t^{*,\text{PDIS}}}\left(\rho_t q_{\pi,t}(S_t,A_t) \mid S_t\right). && \text{(By (B.7))}
\end{aligned}
$$

For $t \in [T-2]$,

$$\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,\mathrm{PDIS}}}) \mid S_t = s\right) && \text{(B.19)}\\
=&\mathbb{E}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,\mathrm{PDIS}}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t,A_t) + q_{\pi,t}(S_t,A_t)^2\right) \mid S_t\right]\\
&- v_{\pi,t}(S_t)^2 && \text{(Lemma 37 with } b=0)\\
=&\mathbb{E}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,\mathrm{PDIS}}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t,A_t)\right) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left(\rho_t q_{\pi,t}(S_t,A_t) \mid S_t\right) && \text{(Lemma 38 with } b=0)\\
=&\mathbb{E}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t,A_t)\right) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left(\rho_t q_{\pi,t}(S_t,A_t) \mid S_t\right)\\
&+ \mathbb{E}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,\mathrm{PDIS}}}) \mid S_{t+1}\right)\right.\right.\right.\\
&\left.\left.\left.-\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right]\right) \mid S_t\right]\\
\geq&\mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t,A_t)\right) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left(\rho_t q_{\pi,t}(S_t,A_t) \mid S_t\right)\\
&+ \mathbb{E}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,\mathrm{PDIS}}}) \mid S_{t+1}\right)\right.\right.\right.\\
&\left.\left.\left.-\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right]\right) \mid S_t\right]\\
&\qquad(\forall \mu_t^{*,\mathrm{PDIS}} \in \Lambda_t,\ \mu_t^{*,b^*} \text{ achieves the minimum value of the first term in } \Lambda_t)\\
=&\mathbb{E}_{A_t\sim\mu_t^{*,b^*}}\left[\rho_t^2 u_t^{b^*}(S_t,A_t) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left(\rho_t q_{\pi,t}(S_t,A_t) \mid S_t\right)\\
&+ \mathbb{E}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,\mathrm{PDIS}}}) \mid S_{t+1}\right)\right.\right.\right.\\
&\left.\left.\left.-\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right]\right) \mid S_t\right] && \text{(By (B.11))}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_t^{b^*}(S_t,A_t)} \mid S_t\right]^2\\
&+ \mathbb{V}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left(\rho_t q_{\pi,t}(S_t,A_t) \mid S_t\right)\\
&+ \mathbb{E}_{A_t\sim\mu_t^{*,\mathrm{PDIS}}}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,\mathrm{PDIS}}}) \mid S_{t+1}\right)\right.\right.\right.\\
&\left.\left.\left.-\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right]\right) \mid S_t\right]. && \text{(By (B.7))}
\end{aligned}$$

Thus, for $t = T-1$,

$$\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,\mathrm{PDIS}}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b^*}}) \mid S_t = s\right) \\
=&\mathbb{V}_{A_t \sim \mu_t^{*,\mathrm{PDIS}}}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right). \qquad \text{(By (B.12) and (B.18))}
\end{aligned}$$

For $t \in [T-2]$,

$$\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,\mathrm{PDIS}}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b^*}}) \mid S_t = s\right) \\
=&\mathbb{V}_{A_t \sim \mu_t^{*,\mathrm{PDIS}}}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right) \\
&+ \mathbb{E}_{A_t \sim \mu_t^{*,\mathrm{PDIS}}}\left[\rho_t^2 \left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,\mathrm{PDIS}}}) \mid S_{t+1}\right)\right.\right.\right. \\
&\left.\left.\left. - \mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right]\right) \mid S_t\right]. \qquad \text{(By (B.13)and (B.19))}
\end{aligned}$$

The proof of $\forall t, s, \delta_t^{\mathrm{ODI,\ ours}}(s) \geq 0$ is similar to (B.17) and is omitted.

□

B.1.7 Proof of Theorem 15

Proof. We begin the proof by manipulating the variance of $G^{b^*}(\tau_{t:T-1}^{\pi_{t:T-1}})$. $\forall s$, for $t = T-1$,

$$\begin{aligned}
&\mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) \qquad \text{(B.20)} \\
=&\mathbb{E}_{A_t \sim \pi_t}\left[[q_{\pi,t}(S_t, A_t) - b_t^*(S_t, A_t)]^2 \mid S_t\right] - [v_{\pi,t}(S_t) - \bar{b^*}_t(S_t)]^2 \\
&\qquad \text{(Lemma 37 and on-policy)} \\
=&0 \qquad \text{(Definition of } b^* \text{ (6.15))} \\
=&\mathbb{E}_{A_t \sim \pi_t}\left[u_t^{b^*}(S_t, A_t) \mid S_t\right]. \qquad \text{(By (B.11))}
\end{aligned}$$

For $t \in [T-2]$,

$$\begin{aligned}
&\mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) && \text{(B.21)}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) + [q_{\pi,t}(S_t, A_t) - b_t(S_t, A_t)]^2 \mid S_t\right]\\
&- [v_{\pi,t}(S_t) - \bar{b}_t(S_t)]^2 && \text{(By Lemma 37)}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) \mid S_t\right]\\
&+ \mathbb{V}_{A_t\sim\pi_t}\left([q_{\pi,t}(S_t, A_t) - b_t^*(S_t, A_t)] \mid S_t\right) && \text{(Lemma 38)}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) \mid S_t\right] && \text{(By (6.15))}\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t) \mid S_t\right]\\
&+ \mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] \mid S_t\right]\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[u_t^{b^*}(S_t, A_t) \mid S_t\right]\\
&+ \mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] \mid S_t\right].\\
&&& \text{(By (B.11))}
\end{aligned}$$

Thus, for $t = T-1$, their difference is

$$\begin{aligned}
&\mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b^*}}) \mid S_t = s\right)\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[u_t^{b^*}(S_t, A_t) \mid S_t\right] - \mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t\right]^2 && \text{(By (B.12) and (B.20))}\\
=&\mathbb{V}_{A_t\sim\pi_t}\left(\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t\right).
\end{aligned}$$

For $t \in [T-2]$,

$$\begin{aligned}
&\mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right) - \mathbb{V}\left(G^{b^*}(\tau_{t:T-1}^{\mu_{t:T-1}^{*,b^*}}) \mid S_t = s\right)\\
=&\mathbb{E}_{A_t\sim\pi_t}\left[u_t^{b^*}(S_t, A_t) \mid S_t\right] - \mathbb{E}_{A_t\sim\pi_t}\left[\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t\right]^2\\
&+ \mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] \mid S_t\right]\\
&&& \text{(By (B.13) and (B.21))}\\
=&\mathbb{V}_{A_t\sim\pi_t}\left(\sqrt{u_t^{b^*}(S_t, A_t)} \mid S_t\right)\\
&+ \mathbb{E}_{A_t\sim\pi_t}\left[\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}) \mid S_{t+1}\right) - \mathbb{V}\left(G^{b^*}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}^{*,b^*}}) \mid S_{t+1}\right) \mid S_t, A_t\right] \mid S_t\right].
\end{aligned}$$

The proof of $\forall t, s, \delta_t^{\text{DR, ours}}(s) \geq 0$ is similar to (B.17) and is omitted.

□

B.1.8 Proof of Lemma 10

Lemma 10 (Recursive form of u). *With $b = b^*$, when $t = T-1$, $\forall s, a$, $u_{\pi,t}(s,a) = 0$, when $t \in [T-2]$, $\forall s, a$,*

$$u_{\pi,t}(s,a) = \nu_{\pi,t}(s,a) + \textstyle\sum_{s',a'} \rho_{t+1} p(s'|s,a)\pi_{t+1}(a'|s')u_{\pi,t+1}(s',a').$$

Proof. When $t = T-1$, $\forall s, a$,

$$\begin{aligned}
&u_{\pi,t}(s,a) \\
=&(q_{\pi,t}(s,a) - b_t^*(s,a))^2 + \nu_{\pi,t}(s,a) && \text{(By (B.11))} \\
=&0. && \text{(By (6.15))}
\end{aligned}$$

When $t \in [T-2]$, $\forall s, a$,

$$\begin{aligned}
&u_{\pi,t}(s,a) \\
=&\nu_{\pi,t}(s,a) + \sum_{s'} p(s'|s,a)\mathbb{V}\left(G^b(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1} = s'\right) && \text{(By (B.11))} \\
=&\nu_{\pi,t}(s,a) + \sum_{s'} p(s'|s,a)\Big[\mathbb{E}_{A_{t+1}\sim\mu^*_{t+1}}\left[\rho_{t+1}^2 u_{\pi,t+1}(S_{t+1},A_{t+1}) \mid S_{t+1} = s'\right] \\
&-[v_{\pi,t+1}(s') - \bar{b}^*_{t+1}(s')]^2\Big] && \text{(Lemma 37)} \\
=&\nu_{\pi,t}(s,a) + \sum_{s'} p(s'|s,a)\mathbb{E}_{A_{t+1}\sim\mu^*_{t+1}}\left[\rho_{t+1}^2 u_{\pi,t+1}(S_{t+1},A_{t+1}) \mid S_{t+1} = s'\right] && \text{(By (6.15))} \\
=&\nu_{\pi,t}(s,a) + \sum_{s',a'} \rho_{t+1} p(s'|s,a)\pi_{t+1}(a'|s')u_{\pi,t+1}(s',a').
\end{aligned}$$

□

B.2 Experiment Details

We utilize the behavior policy-agnostic offline learning setting (Nachum et al., 2019), in which the offline data consists of $\{(t_i, s_i, a_i, r_i, s_i')\}_{i=1}^m$, with m previously logged data tuples. Those tuples can be generated by one or multiple behavior policies, regardless of whether these policies are known or unknown, and they are not required to form a complete trajectory. In the i-th data tuple, t_i represents the time step, s_i is the state at time step t_i, a_i is the action executed, r_i is the sampled reward, and s_i' is the successor state.

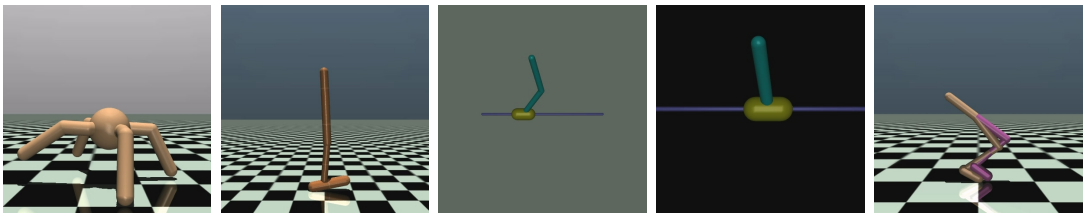

Figure B.1: MuJoCo robot simulation tasks (Todorov et al., 2012). The pictures are adapted from (Liu and Zhang, 2024). Environments from the left to the right are Ant, Hopper, InvertedDoublePendulum, InvertedPendulum, and Walker.

In this work, we first learn the action-value function $q_{\pi,t}$ from offline data using Fitted Q-Evaluation algorithms (FQE, Le et al. (2019)), but our method can integrate state-of-the-art offline policy evaluation techniques. Notably, Fitted Q-Evaluation (FQE, Le et al. (2019)) is a different algorithm from Fitted Q-Improvement (FQI). Fitted Q-Evaluation is not prone to overestimate the action-value function $q_{\pi,t}$ because Fitted Q-Evaluation does not have any max operator and does not change the policy.

Then, by the following derivation

$$
\begin{aligned}
&\nu_{\pi,t}(s,a) \\
=&\mathbb{V}_{S_{t+1}}\left(v_{\pi,t+1}(S_{t+1}) \mid S_t = s, A_t = a\right) && \text{(By (6.11))} \\
=&\mathbb{E}_{S_{t+1}}\left[v_{\pi,t+1}(S_{t+1})^2 \mid S_t = s, A_t = a\right] - \mathbb{E}_{S_{t+1}}\left[v_{\pi,t+1}(S_{t+1}) \mid S_t = s, A_t = a\right]^2 \\
=&\mathbb{E}_{S_{t+1}}\left[v_{\pi,t+1}(S_{t+1})^2 \mid S_t = s, A_t = a\right] - (q_{\pi,t}(s,a) - r(s,a))^2, && \text{(B.22)}
\end{aligned}
$$

the first term is an expectation of S_{t+1}. Because we have $(t_i, s_i, a_i, r_i, s_i')$ data tuples, we construct ν using s_i' in $(t_i, s_i, a_i, r_i, s_i')$ data tuples as the sample of the first term and compute the rest quantity using the learned action-value function $q_{\pi,t}$ and reward data r_i. Therefore, we construct $\mathcal{D}_\nu \doteq \{(t_i, s_i, a_i, \nu_i, s_i')\}_{i=1}^m$. Finally, by passing data tuples in $\mathcal{D}_\nu$ from $t = T-1$ to 0, we fit the function $u_{\pi,t}$ using FQE in a dynamic programming way with respect to the recursive form of $u_{\pi,t}$ derived in Lemma 10. For each time step, we take a copy of the neural network as an approximation of function $u_{\pi,t}$ at time step t. After learning the functions $u_{\pi,t}$ and $q_{\pi,t}$, we return the learned behavior policy $\mu_t^*(a|s) \propto \pi_t(a|s)\sqrt{u_{\pi,t}(s,a)}$ and the learned baseline function $b_t^*(s,a) = q_{\pi,t}(s,a)$. The pseudocode of this procedure is presented in Algorithm 3.

B.2.1 GridWorld

For a Gridworld with size n, we set its width, height, and time horizon all to be n. The number of states in this Gridworld environment scales cubically with n, offering a suitable tool to test algorithm scalability. We choose Gridworld with $n^3 = 1,000$

and $n^3 = 27,000$, the largest Gridworld environment tested among related works (Jiang and Li, 2016; Hanna et al., 2017; Liu and Zhang, 2024). There are four possible actions: left, right, up, and down. After the agent takes an action, it has a probability of 0.9 to move accordingly and a probability of 0.1 to move uniformly at random. If runs into a boundary, the agent stays in its current location. The reward function $r(s, a)$ is randomly generated. We consider 30 randomly generated target policies. We generate the ground truth policy performance using the on-policy Monte Carlo method, running each target policy for 10^6 episodes. We test two different environment sizes of the Gridworld, one with $1,000$ states and $27,000$ states. The offline dataset of both environments contains $1,000$ episodes generated by a set of random policies. To learn functions $q_{\pi,t}$ and $u_{\pi,t}$, we split the offline data into a training set and a test set. We tune all hyperparameters offline based on Fitted Q-learning loss on the test set. We choose a one-hidden-layer neural network and test the neural network size with $[64, 128, 256]$ and choose 64 as the final size. We test the learning rate for Adam optimizer with $[1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}]$ and choose to use the default learning rate $1e^{-3}$ as learning rate for Adam optimizer (Kingma and Ba, 2015). All benchmark algorithms are learned using their reported hyperparameters (Jiang and Li, 2016; Liu and Zhang, 2024). Each policy has 30 independent runs, resulting in $30 \cdot 30 = 900$ total runs. Therefore, each curve in Figure 6.1 is averaged from 900 different runs over a wide range of policies, showing a strong statistical significance.

B.2.2 MuJoCo

MuJoCo is a physics engine containing various stochastic environments, where the goal is to control a robot to achieve different behaviors such as walking, jumping, and balancing. Environments in Figure B.1 from the left to the right are Ant, Hopper, InvertedDoublePendulum, InvertedPendulum, and Walker. We construct 30 policies in each environment (resulting a total of 150 policies), incorporating a wide range of performance generated by the proximal policy optimization (PPO) algorithm (Schulman et al., 2017). We use the the default PPO implementation in Huang et al. (2022). We set each MuJoCo environment to have a fixed time horizon 100 in OpenAI Gymnasium (Towers et al., 2024). As our methods are designed for discrete action space, we discretize the first dimension of MuJoCo action space. The remaining dimensions are controlled by the PPO policies, and they are deemed as part of the environment. The offline dataset of each environment contains $1,000$ episodes generated by a set of policies with various performances. Functions $q_{\pi,t}$ and $u_{\pi,t}$ are learned the same way as in Gridworld environments. Our algorithm is robust on

hyperparameters. All hyperparameters in Algorithm 3 are tuned offline and are the same across all MuJoCo and Gridworld experiments. Each policy in MuJoCo also has 30 independent runs, resulting in $30 \cdot 30 = 900$ total runs. Therefore, each curve in Figure 6.2 and each number in Table 4.2 are averaged from 900 different runs over a wide range of policies indicating strong statistical significance.

Appendix C

Appendix for Chapter 5

C.1 Proofs

C.1.1 Proof of Lemma 4

Proof. $\forall k$,

$$\begin{aligned}
\mathbb{E}_{A\sim\mu}\left[\rho^{\pi^{(k)},\mu}(A)q(A)\right] &= \sum_{a\in\{a|\mu(a)>0\}} \mu(a)\frac{\pi^{(k)}(a)}{\mu(a)}q(a) \\
&= \sum_{a\in\{a|\mu(a)>0\}} \pi^{(k)}(a)q(a) \\
&= \sum_{a\in\{a|\mu(a)>0\}} \pi^{(k)}(a)q(a) + \sum_{a\in\{a|\mu(a)=0\}} \pi^{(k)}(a)q(a) \qquad (\mu\in\Lambda) \\
&= \sum_{a} \pi^{(k)}(a)q(a) \\
&= \mathbb{E}_{A\sim\pi^{(k)}}\left[q(A)\right].
\end{aligned}$$

The intuition in the third equation is that the sample a where μ does not cover $\pi^{(k)}$ must satisfy $q(a) = 0$, i.e., this sample does not contribute to the expectation anyway. □

C.1.2 Proof of Lemma 5

Proof.
Define

$$\mathcal{A}_+ \doteq \left\{a \mid \exists k, \pi^{(k)}(a)q(a) \neq 0\right\}. \tag{C.1}$$

For any $\mu \in \Lambda$, we expand the variance in (5.4) as

$$\begin{aligned}
&\sum_{k\in[K]} \mathbb{V}_{A\sim\mu}(\rho^{\pi^{(k)},\mu}(A)q(A)) && \text{(C.2)}\\
&= \sum_{k\in[K]} \mathbb{E}_{A\sim\mu}[(\rho^{\pi^{(k)},\mu}(A)q(A))^2] - \mathbb{E}_{A\sim\mu}[\rho^{\pi^{(k)},\mu}(A)q(A)]^2\\
&= \sum_{k\in[K]} \mathbb{E}_{A\sim\mu}[(\rho^{\pi^{(k)},\mu}(A)q(A))^2] - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 && \text{(Lemma 4)}\\
&= \sum_{k\in[K]} \sum_{a\in\{a|\mu(a)>0\}} \frac{\pi^{(k)}(a)^2 q(a)^2}{\mu(a)} - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2\\
&= \sum_{k\in[K]} \sum_{a\in\{a|\mu(a)>0\}\cap\mathcal{A}_+} \frac{\pi^{(k)}(a)^2 q(a)^2}{\mu(a)} - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 \quad (\forall a \notin \mathcal{A}_+, \forall k, \pi^{(k)}(a)q(a) = 0)\\
&= \sum_{a\in\mathcal{A}_+} \frac{\sum_{k\in[K]} \pi^{(k)}(a)^2 q(a)^2}{\mu(a)} - \sum_{k\in[K]} \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2. && (\mu \in \Lambda)
\end{aligned}$$

The second term is a constant and is unrelated to μ. Solving the optimization problem (5.4) is, therefore, equivalent to solving

$$\min_{\mu\in\Lambda} \quad \sum_{a\in\mathcal{A}_+} \frac{\sum_{k\in[K]} \pi^{(k)}(a)^2 q(a)^2}{\mu(a)}. \tag{C.3}$$

Case 1: $|\mathcal{A}_+| = 0$

In this case, optimization target (C.2) is always 0. Any $\mu \in \Lambda$ is optimal. In particular, $\mu^*(a) = \frac{1}{\mathcal{A}}$ is optimal.

Case 2: $|\mathcal{A}_+| > 0$

The definition of Λ in (5.3) can be equivalently expressed, using contraposition, as

$$\Lambda = \{\mu \in \Delta(\mathcal{A}) \mid \forall a, a \in \mathcal{A}_+ \implies \mu(a) > 0\}.$$

The optimization problem (C.3) can then be equivalently written as

$$\begin{aligned}
\min_{\mu\in\Delta(\mathcal{A})} \quad & \sum_{a\in\mathcal{A}_+} \frac{\sum_{k\in[K]} \pi^{(k)}(a)^2 q(a)^2}{\mu(a)} \\
\text{s.t.} \quad & \mu(a) > 0 \quad \forall a \in \mathcal{A}_+.
\end{aligned} \tag{C.4}$$

If for some μ we have $\sum_{a\in\mathcal{A}_+} \mu(a) < 1$, then there must exist some $a_0 \notin \mathcal{A}_+$ such that $\mu(a_0) > 0$. By the definition of $\mathcal{A}^+$ (C.1), $\forall a_0 \notin \mathcal{A}_+$,

$$\sum_{k\in[K]} {\pi^{(k)}}^2(a_0) q^2(a_0) = 0.$$

This means a_0 does not contribute to the summation in the objective function of (C.4), we can move the probability mass on a_0 to some other $a_1 \in \mathcal{A}_+$ to increase $\mu(a_1)$ to further decrease the objective. In other words, any optimal solution μ to (C.4) must put all its probability mass on $\mathcal{A}_+$. This motivates the following problem

$$\begin{aligned} \min_{z \in \Delta(\mathcal{A}_+)} \quad & \sum_{a \in \mathcal{A}_+} \frac{\sum_{k \in [K]} \pi^{(k)}(a)^2 q(a)^2}{z(a)} \\ \text{s.t.} \quad & z(a) > 0 \quad \forall a \in \mathcal{A}_+. \end{aligned} \tag{C.5}$$

In particular, if z^* is an optimal solution to (C.5), then an optimal solution to (C.4) can be constructed as

$$\mu^*(a) = \begin{cases} z^*(a) & a \in \mathcal{A}_+, \\ 0 & \text{otherwise.} \end{cases} \tag{C.6}$$

Let $\mathbb{R}_{++} \doteq (0, +\infty)$. According to the Cauchy-Schwarz inequality, for any $z \in \mathbb{R}_{++}^{|\mathcal{A}_+|}$, we have

$$\begin{aligned} \left(\sum_{a \in \mathcal{A}_+} \frac{\sum_{k \in [K]} \pi^{(k)}(a)^2 q(a)^2}{z(a)} \right) \left(\sum_{a \in \mathcal{A}_+} z(a) \right) &\geq \left(\sum_{a \in \mathcal{A}_+} \frac{\sqrt{\sum_{k \in [K]} \pi^{(k)}(a)^2 q(a)^2}}{\sqrt{z(a)}} \sqrt{z(a)} \right)^2 \\ &= \left(\sum_{a \in \mathcal{A}_+} \sqrt{\sum_{k \in [K]} \pi^{(k)}(a)^2 q(a)^2} \right)^2. \end{aligned}$$

It can be easily verified that the equality holds for

$$z^*(a) \doteq \frac{\sqrt{\sum_{k \in [K]} \pi^{(k)}(a)^2 q(a)^2}}{\sum_b \sqrt{\sum_{k \in [K]} \pi^{(k)}(b)^2 q(b)^2}} > 0.$$

Since $\sum_{a \in \mathcal{A}_+} z^*(a) = 1$, we conclude that z^* is an optimal solution to (C.5). An optimal solution μ^* to (5.4) can then be constructed according to (C.6). Making use of the fact that $\forall a \notin \mathcal{A}, \forall k, \pi^{(k)}(a) q(a) = 0$, this μ^* can be equivalently expressed as

$$\mu^*(a) = \frac{\sqrt{\sum_{k \in [K]} \pi^{(k)}(a)^2 q(a)^2}}{\sum_{b \in \mathcal{A}} \sqrt{\sum_{k \in [K]} \pi^{(k)}(b)^2 q(b)^2}},$$

which completes the proof. □

C.1.3 Proof of Lemma 6

Proof. $\forall k$, we first derive an upper-bound on $\mathbb{V}_{A\sim\mu^*}\left(\rho^{\pi^{(k)},\mu^*}(A)q(A)\right)$,

$$
\begin{aligned}
&\mathbb{V}_{A\sim\mu^*}(\rho^{\pi^{(k)},\mu^*}(A)q(A)) \\
=&\mathbb{E}_{A\sim\mu^*}\left[(\rho^{\pi^{(k)},\mu^*}(A)q(A))^2\right] - \mathbb{E}_{A\sim\mu^*}\left[\rho^{\pi^{(k)},\mu^*}(A)q(A)\right]^2 \\
=&\mathbb{E}_{A\sim\mu^*}\left[(\rho^{\pi^{(k)},\mu^*}(A)q(A))^2\right] - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 && \text{(Lemma 4)} \\
=&\mathbb{E}_{A\sim\mu^*}\left[\frac{w^{(k)}(a)}{\mu^*(a)^2}\right] - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 && \text{(By (5.5))} \\
=&\sum_a w^{(k)}(a)\frac{1}{\mu^*(a)} - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 \\
=&\sum_a w^{(k)}(a)\left(\frac{\sum_b\sqrt{\sum_{j\in[K]}w^{(j)}(b)}}{\sqrt{\sum_{j\in[K]}w^{(j)}(a))}}\right) - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 \\
& && \text{(By (5.5) and definition of } \mu^*) \\
=&\sum_a w^{(k)}(a)\left(\frac{\sum_b\sqrt{K\bar{w}(b)}}{\sqrt{K\bar{w}(a)}}\right) - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 && \text{(By (5.6))} \\
=&\sum_a w^{(k)}(a)\left(\frac{\sum_b\sqrt{\bar{w}(b)}}{\sqrt{\bar{w}(a)}}\right) - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 \\
=&\sum_a w^{(k)}(a)\left(\frac{\sum_b\sqrt{\frac{w^{(k)}(b)}{\eta^{(k)}(b)}}}{\sqrt{\frac{w^{(k)}(a)}{\eta^{(k)}(a)}}}\right) - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 && \text{(By (5.7))} \\
\leq&\sum_a w^{(k)}(a)\left(\frac{\sum_b\sqrt{\frac{w^{(k)}(b)}{\underline{\eta}}}}{\sqrt{\frac{w^{(k)}(a)}{\overline{\eta}}}}\right) - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 && \text{(By (5.8))} \\
=&\sum_a w^{(k)}(a)\left(\frac{\sqrt{\frac{1}{\underline{\eta}}}\sum_b\sqrt{w^{(k)}(b)}}{\sqrt{\frac{1}{\overline{\eta}}}\sqrt{w^{(k)}(a)}}\right) - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 \\
=&\sum_a w^{(k)}(a)\left(\frac{\sum_b\sqrt{\overline{\eta}w^{(k)}(b)}}{\sqrt{\underline{\eta}w^{(k)}(a)}}\right) - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{\frac{\overline{\eta}}{\underline{\eta}}} \sum_a w^{(k)}(a) \left(\frac{\sum_b \sqrt{w^{(k)}(b)}}{\sqrt{w^{(k)}(a)}} \right) - \mathbb{E}_{A \sim \pi^{(k)}}[q(A)]^2 \\
&= \sqrt{\frac{\overline{\eta}}{\underline{\eta}}} \left(\sum_a \sqrt{w^{(k)}(a)} \right) \left(\sum_b \sqrt{w^{(k)}(b)} \right) - \mathbb{E}_{A \sim \pi^{(k)}}[q(A)]^2 \\
&= \sqrt{\frac{\overline{\eta}}{\underline{\eta}}} \left(\sum_a \sqrt{w^{(k)}(a)} \right)^2 - \mathbb{E}_{A \sim \pi^{(k)}}[q(A)]^2 \\
&= \sqrt{\frac{\overline{\eta}}{\underline{\eta}}} \left(\sum_a \pi^{(k)}(a) q(a) \right)^2 - \mathbb{E}_{A \sim \pi^{(k)}}[q(A)]^2. \qquad \text{(By (5.5))} \quad \text{(C.7)}
\end{aligned}
$$

Then, $\forall k \in [K]$, observe the following inequality,

$$
\begin{aligned}
&\frac{1}{n} \left[\sqrt{\frac{\overline{\eta}}{\underline{\eta}}} \left(\sum_a \pi^{(k)}(a) q(a) \right)^2 - \mathbb{E}_{A \sim \pi^{(k)}}[q(A)]^2 \right] \\
=&\frac{1}{n} \left[\sqrt{\frac{\overline{\eta}}{\underline{\eta}}} \left(\sum_a \pi^{(k)}(a) q(a) \right)^2 - \left(\frac{n}{n_k} - 1 \right) \left[\sum_a \pi^{(k)}(a) q(a)^2 - \left(\sum_a \pi^{(k)}(a) q(a) \right)^2 \right] \right. \\
&\left. + \left(\frac{n}{n_k} - 1 \right) \left[\sum_a \pi^{(k)}(a) q(a)^2 - \left(\sum_a \pi^{(k)}(a) q(a) \right)^2 \right] - \mathbb{E}_{A \sim \pi^{(k)}}[q(A)]^2 \right] \\
\leq&\frac{1}{n} \left[\sum_a \pi^{(k)}(a) q(a)^2 + \left(\frac{n}{n_k} - 1 \right) \left[\sum_a \pi^{(k)}(a) q(a)^2 - \left(\sum_a \pi^{(k)}(a) q(a) \right)^2 \right] - \mathbb{E}_{A \sim \pi^{(k)}}[q(a)]^2 \right] \\
& \qquad \text{(By (5.11))} \\
=&\frac{1}{n} \left[\frac{n}{n_k} \sum_a \pi^{(k)}(a) q(a)^2 - \left(\frac{n}{n_k} - 1 \right) \mathbb{E}_{A \sim \pi^{(k)}}[q(a)]^2 - \mathbb{E}_{A \sim \pi^{(k)}}[q(a)]^2 \right] \\
=&\frac{1}{n} \left[\frac{n}{n_k} \sum_a \pi^{(k)}(a) q(a)^2 - \frac{n}{n_k} \mathbb{E}_{A \sim \pi^{(k)}}[q(a)]^2 \right] \\
=&\frac{1}{n_k} \left[\sum_a \pi^{(k)}(a) q(a)^2 - \mathbb{E}_{A \sim \pi^{(k)}}[q(a)]^2 \right]. \qquad \text{(C.8)}
\end{aligned}
$$

Now, we have $\forall k \in [K]$,

$$
\begin{aligned}
&\mathbb{V}_{A\sim\mu^*}\left(E^{\mathrm{off},\pi^{(k)}}\right)\\
=&\mathbb{V}_{A\sim\mu^*}\left(\frac{\sum_{i=1}^{n}\rho^{\pi^{(k)},\mu^*}(A^{[\mu^*,i]})q(A^{[\mu^*,i]})}{n}\right) && \text{(By (5.10))}\\
=&\frac{1}{n^2}\mathbb{V}_{A\sim\mu^*}\left(\sum_{i=1}^{n}\rho^{\pi^{(k)},\mu^*}(A^{[\mu^*,i]})q(A^{[\mu^*,i]})\right)\\
=&\frac{1}{n}\mathbb{V}_{A\sim\mu^*}\left(\rho^{\pi^{(k)},\mu^*}(A)q(A)\right) && \text{(Independence of } A^{[\mu^*,i]}\text{)}\\
\leq&\frac{1}{n}\left[\sqrt{\frac{\overline{\eta}}{\underline{\eta}}}\left(\sum_a \pi^{(k)}(a)q(a)\right)^2-\mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2\right] && \text{(by (C.7))}\\
\leq&\frac{1}{n_k}\left[\sum_a \pi^{(k)}(a)q(a)^2-\mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2\right] && \text{(by (C.8))}\\
=&\frac{1}{n_k}\mathbb{V}_{A\sim\pi^{(k)}}(q(A))\\
=&\frac{1}{n_k^2}\mathbb{V}_{A\sim\pi^{(k)}}\left(\sum_{i=1}^{n_k}q(A^{[\pi^{(k)},i]})\right) && \text{(Independence of } A^{[\pi^{(k)},i]}\text{)}\\
=&\mathbb{V}_{A\sim\pi^{(k)}}\left(\frac{\sum_{i=1}^{n_k}q(A^{[\pi^{(k)},i]})}{n_k}\right)\\
=&\mathbb{V}_{A\sim\pi^{(k)}}\left(E^{\mathrm{on},\pi^{(k)}}\right). && \text{(By (5.9))}
\end{aligned}
$$

□

C.1.4 Proof of Lemma 7

Proof. When sampling from the target policy $\pi^{(k)}$, we have $\forall k$,

$$
\begin{aligned}
&\mathbb{V}_{A\sim\pi^{(k)}}(q(A))\\
=&\mathbb{E}_{A\sim\pi^{(k)}}\left[q(A)^2\right]-\mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2\\
=&\sum_a \pi^{(k)}(a)q(a)^2-\mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2. && \text{(C.9)}
\end{aligned}
$$

With the sufficient condition (5.13), we show the variance reduction. $\forall k$,

$$\begin{aligned}
&\mathbb{V}_{A\sim\mu^*}\left(\rho^{\pi^{(k)},\mu^*}(A)q(A)\right)\\
=&\sqrt{\frac{\overline{\eta}}{\underline{\eta}}}\left(\sum_a \pi^{(k)}(a)q(a)\right)^2 - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 && \text{(by (C.7))}\\
\leq&\sum_a \pi^{(k)}(A)q(a)^2 - \mathbb{E}_{A\sim\pi^{(k)}}[q(A)]^2 && \text{(by (5.13))}\\
=&\mathbb{V}_{A\sim\pi^{(k)}}(q(A)). && \text{(By (C.9))}
\end{aligned}$$

□

C.1.5 Proof of Theorem 6

Before proving Theorem 6, we first present an auxiliary lemma that is a stronger version of Lemma 4.

Lemma 39. $\forall \mu \in \Lambda$, $\forall k, \forall t, \forall s$,

$$\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^{\pi^{(k)},\mu} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t = s\right] = \mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[q_{\pi^{(k)},t}(S_t, A_t) \mid S_t = s\right].$$

Proof. $\forall \mu \in \Lambda$, $\forall k, \forall t, \forall s$,

$$\begin{aligned}
&\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^{\pi^{(k)},\mu} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t = s\right]\\
=&\sum_{a\in\{a|\mu_t(a|s)>0\}} \mu_t(a|s)\frac{\pi_t^{(k)}(a|s)}{\mu_t(a|s)} q_{\pi^{(k)},t}(s,a)\\
=&\sum_{a\in\{a|\mu_t(a|s)>0\}} \pi_t^{(k)}(a|s) q_{\pi^{(k)},t}(s,a)\\
=&\sum_{a\in\{a|\mu_t(a|s)>0\}} \pi_t^{(k)}(a|s) q_{\pi^{(k)},t}(s,a) + \sum_{a\in\{a|\mu_t(a|s)=0\}} \pi_t^{(k)}(a|s) q_{\pi^{(k)},t}(s,a) && (\mu\in\Lambda)\\
=&\sum_a \pi_t^{(k)}(a|s) q_{\pi^{(k)},t}(s,a)\\
=&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[q_{\pi^{(k)},t}(S_t, A_t) \mid S_t = s\right].
\end{aligned}$$

□

Now, we are ready to prove Theorem 6.

Proof. We proceed via induction. $\forall k$, for $t = T-1$, we have

$$\begin{aligned}
&\mathbb{E}\left[G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} r(S_t, A_t) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] \\
=&\mathbb{E}_{A_t \sim \pi_t^{(k)}(\cdot|S_t)}\left[q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] && \text{(Lemma 39)} \\
=&v_{\pi^{(k)},t}(S_t).
\end{aligned}$$

For $t \in [T-2]$, we have

$$\begin{aligned}
&\mathbb{E}\left[G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} R_{t+1} + \rho_t^{\pi^{(k)},\mu} G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} R_{t+1} \mid S_t\right] + \mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t \sim \mu_t(\cdot|S_t), S_{t+1} \sim p(\cdot|S_t, A_t)}\left[\mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t, A_t, S_{t+1}\right] \mid S_t\right] \\
& && \text{(Law of total expectation)} \\
=&\mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t \sim \mu_t(\cdot|S_t), S_{t+1} \sim p(\cdot|S_t, A_t)}\left[\rho_t^{\pi^{(k)},\mu} \mathbb{E}\left[G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_{t+1}\right] \mid S_t\right] \\
& && \text{(Conditional independence and Markov property)} \\
=&\mathbb{E}\left[\rho_t^{\pi^{(k)},\mu} R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t \sim \mu_t(\cdot|S_t), S_{t+1} \sim p(\cdot|S_t, A_t)}\left[\rho_t^{\pi^{(k)},\mu} v_{\pi^{(k)},t+1}(S_{t+1}) \mid S_t\right] \\
& && \text{(Inductive hypothesis)} \\
=&\mathbb{E}_{A_t \sim \mu_t(\cdot|S_t)}\left[\rho_t^{\pi^{(k)},\mu} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] && \text{(Definition of } q_{\pi^{(k)},t}\text{)} \\
=&\mathbb{E}_{A_t \sim \pi_t^{(k)}(\cdot|S_t)}\left[q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] && \text{(Lemma 39)} \\
=&v_{\pi^{(k)},t}(S_t).
\end{aligned}$$

This completes the proof. □

C.1.6 Proof of Theorem 7

To prove Theorem 7, we rely on a recursive expression of the PDIS Monte Carlo estimator, which is restated from Liu and Zhang (2024), as summarized by the following lemma.

Lemma 40 (Recursive Expression of Variance)**.** *For any $\mu \in \Lambda$, $\forall k$, we have for $t = T-1$,*

$$\mathbb{V}\left(G_k^{PDIS}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t\right) = \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^{\pi^{(k)},\mu^2} q_{\pi^{(k)},t}(S_t, A_t)^2 \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2;$$

For $t \in [T-2]$,

$$\begin{aligned}
&\mathbb{V}\left(G_k^{PDIS}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t\right) \\
=&\mathbb{E}_{A_t\sim\mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{PDIS}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t\right) \mid S_t, A_t\right] + \nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] \\
&- v_{\pi^{(k)},t}(S_t)^2.
\end{aligned}$$

Proof. When $t \in [T-2]$, we have

$$\begin{aligned}
&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t\right) && \text{(C.10)}\\
=&\mathbb{E}_{A_t}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\mathbb{E}\left[G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t, A_t\right] \mid S_t\right) \\
& && \text{(Law of total variance)}\\
=&\mathbb{E}_{A_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\mathbb{V}\left(r(S_t,A_t) + G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t, A_t\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\left(\rho_t^{\pi^{(k)},\mu}\mathbb{E}\left[r(S_t,A_t) + G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t, A_t\right] \mid S_t\right) && \text{(Using (5.1))}\\
=&\mathbb{E}_{A_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\rho_t^{\pi^{(k)},\mu}q_{\pi^{(k)},t}(S_t,A_t) \mid S_t\right). \\
& && \text{(Deterministic reward } r\text{)}
\end{aligned}$$

Further decomposing the first term, we have

$$\begin{aligned}
&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t, A_t\right) && \text{(C.11)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t, A_t, S_{t+1}\right) \mid S_t, A_t\right] \\
&+ \mathbb{V}_{S_{t+1}}\left(\mathbb{E}\left[G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_t, A_t, S_{t+1}\right] \mid S_t, A_t\right) && \text{(Law of total variance)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}\left(\mathbb{E}\left[G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_{t+1}\right] \mid S_t, A_t\right) \\
& && \text{(Markov property)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}\left(v_{\pi^{(k)},t+1}(S_{t+1}) \mid S_t, A_t\right). \\
& && \text{(Theorem 6)}
\end{aligned}$$

With $\nu_{\pi^{(k)},t}$ defined in (5.16), plugging (C.11) back to (C.10) yields

$$\begin{aligned}
&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{V}_{A_t}\left(\rho_t^{\pi^{(k)},\mu} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{E}_{A_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 q_{\pi^{(k)},t}(S_t, A_t)^2 \mid S_t\right] - \left(\mathbb{E}_{A_t}\left[\rho_t^{\pi^{(k)},\mu} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right]\right)^2 \\
=&\mathbb{E}_{A_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}\right) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_t(S_t, A_t)\right) \mid S_t\right] \\
&+ \mathbb{E}_{A_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 q_{\pi^{(k)},t}(S_t, A_t)^2 \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2. \qquad \text{(Lemma 39)}
\end{aligned}$$

When $t = T-1$, we have

$$\begin{aligned}
\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\mu_{t:T-1}}\right) \mid S_t\right) =&\mathbb{V}\left(\rho_t^{\pi^{(k)},\mu} r(S_t, A_t) \mid S_t\right) \\
=&\mathbb{V}\left(\rho_t^{\pi^{(k)},\mu} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right) \\
=&\mathbb{E}_{A_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 q_{\pi^{(k)},t}(S_t, A_t)^2 \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2,
\end{aligned}$$

which completes the proof. □

Then, to solve the variance minimization problem, we manipulate the variance expression in (5.17). For any policy k, for any $\mu \in \hat{\Lambda}$, when $t = T-1$,

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\left\{\mu_t, \pi_{t+1}^{(k)}, \ldots, \pi_{T-1}^{(k)}\right\}}\right) \mid S_t = s\right) \qquad \text{(C.12)} \\
=&\sum_{k=1}^K \mathbb{E}_{A_t \sim \mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 q_{\pi^{(k)},t}(S_t, A_t)^2 \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 \qquad \text{(Lemma 40)} \\
=&\sum_{k=1}^K \mathbb{E}_{A_t \sim \mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 \hat{q}_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 \qquad \text{(By (5.18))} \\
=&\sum_{k=1}^K \mathbb{V}_{A_t \sim \mu_t}\left(\rho_t^{\pi^{(k)},\mu}\sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right) - \sum_{k=1}^K \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^{\pi^{(k)},\mu}\sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right]^2 \\
&- v_{\pi^{(k)},t}(S_t)^2 \\
=&\sum_{k=1}^K \mathbb{V}_{A_t \sim \mu_t}\left(\rho_t^{\pi^{(k)},\mu}\sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right) - \sum_{k=1}^K \mathbb{E}_{A_t \sim \pi_t^{(k)}}\left[\sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right]^2 \\
&- v_{\pi^{(k)},t}(S_t)^2. \qquad \text{(Lemma 39 and } \mu_t \in \hat{\Lambda} \subseteq \Lambda)
\end{aligned}$$

For $t \in [T-2]$,

$$\sum_{k=1}^{K} \mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\{\mu_t, \pi_{t+1}^{(k)}, \ldots, \pi_{T-1}^{(k)}\}}\right) \mid S_t = s\right) \tag{C.13}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{A_t \sim \mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 \left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\pi^{(k)}_{t+1:T-1}}\right) \mid S_{t+1}\right) \mid S_t, A_t\right]\right.\right.$$

$$\left.\left. + \nu_{\pi^{(k)},t}(S_t, A_t) + q_{\pi^{(k)},t}(S_t, A_t)^2\right) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 \qquad \text{(Lemma 40)}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{A_t \sim \mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 \hat{q}_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 \qquad \text{(By (5.18))}$$

$$= \sum_{k=1}^{K} \mathbb{V}_{A_t \sim \mu_t}\left(\rho_t^{\pi^{(k)},\mu} \sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right) - \sum_{k=1}^{K} \mathbb{E}_{A_t \sim \mu_t}\left[\rho_t^{\pi^{(k)},\mu} \sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right]^2$$

$$- v_{\pi^{(k)},t}(S_t)^2$$

$$= \sum_{k=1}^{K} \mathbb{V}_{A_t \sim \mu_t}\left(\rho_t^{\pi^{(k)},\mu} \sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right) - \sum_{k=1}^{K} \mathbb{E}_{A_t \sim \pi_t^{(k)}}\left[\sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right]^2$$

$$- v_{\pi^{(k)},t}(S_t)^2. \qquad \text{(Lemma 39 and } \mu_t \in \hat{\Lambda} \subseteq \Lambda\text{)}$$

Since for both (C.12) and (C.13), the second and third terms are unrelated to $\hat{\mu}$, solving (5.17) is equivalent to solve

$$\min_{\mu_t \in \hat{\Lambda}} \quad \sum_{k=1}^{K} \mathbb{V}_{A_t \sim \mu_t}\left(\rho_t^{\pi^{(k)},\mu} \sqrt{\hat{q}_{\pi^{(k)},t}(S_t, A_t)} \mid S_t\right). \tag{C.14}$$

Then, with Lemma 5, we conclude that $\hat{\mu}_t$ as defined in (5.19) is an optimal solution to (C.14), which completes the proof.

C.1.7 Proof of Theorem 8

Before proving Theorem 8, given the sufficient condition in (5.26), we first observe the following equation.

$$\mathbb{E}_{A_t \sim \hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2 \nu_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] + \mathbb{V}_{A_t \sim \hat{\mu}_t}\left(\rho^{\pi^{(k)},\hat{\mu}} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right) \tag{C.15}$$

$$= \mathbb{E}_{A_t \sim \hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2 \nu_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] + \mathbb{E}_{A_t \sim \hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2 q_{\pi^{(k)},t}(S_t, A_t)^2 \mid S_t\right]$$

$$- \mathbb{E}_{A_t \sim \hat{\mu}_t}\left[\rho^{\pi^{(k)},\hat{\mu}} q_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right]^2$$

$$= \mathbb{E}_{A_t \sim \hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2 \nu_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] + \mathbb{E}_{A_t \sim \hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2 q_{\pi^{(k)},t}(S_t, A_t)^2 \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2.$$

(Theorem 6)

Thus, we obtain $\forall t, s$,

$$
\begin{aligned}
&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(a|s)}\right)^2 && \text{(C.16)}\\
&-\frac{n-n_k}{n}\left(\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho^{\pi^{(k)},\hat{\mu}^2}\left(\nu_{\pi^{(k)},t}(s,A_t)+q_{\pi^{(k)},t}(s,A_t)^2\right)\mid s\right]-v_{\pi^{(k)},t}(s)^2\right)\\
=&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(a|s)}\right)^2\\
&-\frac{n-n_k}{n}\left(\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\rho^{\pi^{(k)},\hat{\mu}^2}\nu_{\pi^{(k)},t}(S_t,A_t)\mid S_t=s\right]+\mathbb{V}_{A_t\sim\hat{\mu}_t}\left(\rho^{\pi^{(k)},\hat{\mu}}q_{\pi^{(k)},t}(S_t,A_t)\right)\mid S_t=s\right)\\
& && \text{(By (C.15))}\\
\leq&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t,A_t)\mid S_t=s\right]. && \text{(By (5.26))}
\end{aligned}
$$

We also discover the following inequality. $\forall k$, $\forall s$, $\forall t$,

$$
\begin{aligned}
&\frac{1}{n}\left[\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(a|s)}\right)^2-v_{\pi^{(k)},t}(s)^2\right]\\
=&\frac{1}{n}\left[\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(a|s)}\right)^2-\left(\frac{n}{n_k}-1\right)\left(\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)-v_{\pi^{(k)},t}(s)^2\right)\right.\\
&\left.+\left(\frac{n}{n_k}-1\right)\left(\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)-v_{\pi^{(k)},t}(s)\right)-v_{\pi^{(k)},t}(s)^2\right]\\
\leq&\frac{1}{n}\left[\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)+\left(\frac{n}{n_k}-1\right)\left(\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)-v_{\pi^{(k)},t}(s)^2\right)-v_{\pi^{(k)},t}(s)^2\right]\\
& && \text{(By (C.16))}\\
=&\frac{1}{n}\left[\frac{n}{n_k}\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)-\left(\frac{n}{n_k}-1\right)v_{\pi^{(k)},t}(s)^2-v_{\pi^{(k)},t}(s)^2\right]\\
=&\frac{1}{n}\left[\frac{n}{n_k}\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)-\frac{n}{n_k}v_{\pi^{(k)},t}(s)^2\right]\\
=&\frac{1}{n_k}\left[\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)-v_{\pi^{(k)},t}(s)^2\right]\\
=&\frac{1}{n_k}\left(\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t,A_t)\mid S_t=s\right]-v_{\pi^{(k)},t}(s)^2\right). && \text{(C.17)}
\end{aligned}
$$

Next, to prove Theorem 9, we present a closed-form representation of the variance of the on-policy estimator.

Lemma 41. *For any k, t and s,*

$$\mathbb{V}\left(G_k^{PDIS}\left(\tau_{t:T-1}^{\pi_{t:T-1}^{(k)}}\right) \mid S_t = s\right) = \mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t, A_t) \mid S_t = s\right] - v_{\pi^{(k)},t}(s)^2.$$

Proof. We proceed via induction. When $t = T - 1$,

$$\begin{aligned}
&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\pi_{t:T-1}^{(k)}}\right) \mid S_t\right) \\
=&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[q_{\pi^{(k)},t}(S_t, A_t)^2 \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 && \text{(Lemma 40 and on-policy)} \\
=&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2. && \text{(By (5.18))}
\end{aligned}$$

When $t \in [T-2]$,

$$\begin{aligned}
&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\pi_{t:T-1}}\right) \mid S_t\right) \\
=&\mathbb{E}_{A_t\sim\pi_t}\left[\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\pi_{t+1:T-1}}\right) \mid S_t\right) \mid S_t, A_t\right] + \nu_{\pi^{(k)},t}(S_t, A_t) + q_{\pi^{(k)},t}(S_t, A_t)^2\right) \mid S_t\right] \\
&- v_{\pi^{(k)},t}(S_t)^2 && \text{(Lemma 40 and on-policy)} \\
=&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t, A_t) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2, && \text{(By (5.18))}
\end{aligned}$$

which completes the induction. □

Additionally, we discover the following inequality. $\forall k$, $\forall s$, $\forall t$,

$$
\begin{aligned}
&\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^{\pi^{(k)},\mu^2}\hat{q}_{\pi^{(k)},t}(S_t,A_t)\mid S_t=s\right]\\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[\frac{w_t^{(k)}(S_t,A_t)}{\hat{\mu}_t(A_t|S_t)^2}\mid S_t=s\right] && \text{(By (5.20))}\\
=&\sum_a w_t^{(k)}(s,a)\frac{1}{\hat{\mu}_t(a|s)}\\
=&\sum_a w_t^{(k)}(s,a)\left(\frac{\sum_b\sqrt{\sum_{j\in[K]}w_t^{(j)}(s,b)}}{\sqrt{\sum_{j\in[K]}w_t^{(j)}(s,a)}}\right) && \text{(By (5.19))}\\
=&\sum_a w_t^{(k)}(s,a)\left(\frac{\sum_b\sqrt{K\bar{w}_t(s,b)}}{\sqrt{K\bar{w}_t(s,a)}}\right) && \text{(by (5.21))}\\
=&\sum_a w_t^{(k)}(s,a)\left(\frac{\sum_b\sqrt{\bar{w}_t(s,b)}}{\sqrt{\bar{w}_t(s,a)}}\right)\\
=&\sum_a w_t^{(k)}(s,a)\left(\frac{\sum_b\sqrt{\frac{w_t^{(k)}(s,b)}{\eta_t^{(k)}(s,b)}}}{\sqrt{\frac{w_t^{(k)}(s,a)}{\eta_t^{(k)}(s,a)}}}\right) && \text{(By (5.22))}\\
\le&\sum_a w_t^{(k)}(s,a)\left(\frac{\sum_b\sqrt{\frac{w_t^{(k)}(s,b)}{\underline{\eta}_t}}}{\sqrt{\frac{w_t^{(k)}(s,a)}{\overline{\eta}_t}}}\right) && \text{(By (5.23))}\\
=&\sum_a w_t^{(k)}(s,a)\left(\frac{\sqrt{\frac{1}{\underline{\eta}_t}}\sum_b\sqrt{w_t^{(k)}(s,b)}}{\sqrt{\frac{1}{\overline{\eta}_t}}\sqrt{w_t^{(k)}(s,a)}}\right)\\
=&\sum_a w_t^{(k)}(s,a)\left(\frac{\sum_b\sqrt{\overline{\eta}_t w_t^{(k)}(s,b)}}{\sqrt{\underline{\eta}_t w_t^{(k)}(s,a)}}\right)\\
=&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\sum_a w_t^{(k)}(s,a)\left(\frac{\sum_b\sqrt{w_t^{(k)}(s,b)}}{\sqrt{w_t^{(k)}(s,a)}}\right)\\
=&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a\sqrt{w_t^{(k)}(s,a)}\right)\left(\sum_b\sqrt{w_t^{(k)}(s,b)}\right)\\
=&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a\sqrt{w_t^{(k)}(s,a)}\right)^2\\
=&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a\pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(s,a)}\right)^2. && \text{(By (5.20)(C.18)}
\end{aligned}
$$

From here, we restate Theorem 8 and give its proof.

Theorem 8 (Variance Reduction with Same Sample Sizes)**.** $\forall k$, $\forall t$, $\forall s$,

$$\mathbb{V}\left(E_{t:T-1}^{off,\pi^{(k)}} \mid S_t = s\right) \leq \mathbb{V}\left(E_{t:T-1}^{on,\pi^{(k)}} \mid S_t = s\right).$$

if the similarity η *has* $\forall k, \forall t, \forall s$,

$$\begin{aligned}
&\sqrt{\frac{\overline{\eta_t}}{\underline{\eta_t}}}\left(\textstyle\sum_a \pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(a|s)}\right)^2 - \left(1 - \tfrac{n_k}{n}\right)\Delta_t^{(k)}(s) \\
&\leq \textstyle\sum_a \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a),
\end{aligned} \tag{5.26}$$

where

$$\begin{aligned}
\Delta_t^{(k)}(s) \doteq \mathbb{E}_{A_t\sim\hat{\mu}_t}&\left[\rho^{\pi^{(k)},\hat{\mu}^2}\nu_{\pi^{(k)},t}(S_t,A_t) \mid S_t = s\right] \\
&+\mathbb{V}_{A_t\sim\hat{\mu}_t}\left(\rho^{\pi^{(k)},\hat{\mu}}q_{\pi^{(k)},t}(S_t,A_t) \mid S_t = s\right).
\end{aligned}$$

First, we manipulate the variance of both $E_{0:T-1}^{\text{off},\pi^{(k)}}$ and $E_{0:T-1}^{\text{on},\pi^{(k)}}$.

$$\begin{aligned}
&\mathbb{V}\left(E_{0:T-1}^{\text{off},\pi^{(k)}}\right) && \text{(C.19)}\\
=&\mathbb{V}\left(\frac{\sum_{i=1}^n G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{[\hat{\mu}_{0:T-1},i]}\right)}{n}\right) && \text{(By (5.25))}\\
=&\tfrac{1}{n^2}\mathbb{V}\left(\textstyle\sum_{i=1}^n G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{[\hat{\mu}_{0:T-1},i]}\right)\right)\\
=&\frac{1}{n}\mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\hat{\mu}_{0:T-1}}\right)\right) && \text{(Independence of } \tau_{0:T-1}^{[\hat{\mu}_{0:T-1},i]}\text{)}\\
=&\frac{1}{n}\mathbb{E}_{S_0}\left[\mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\hat{\mu}_{0:T-1}}\right) \mid S_0 = s\right)\right] + \frac{1}{n}\mathbb{V}_{S_0}\left(\mathbb{E}\left[G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\hat{\mu}_{0:T-1}}\right) \mid S_0 = s\right]\right)\\
& && \text{(Law of total variance)}\\
=&\frac{1}{n}\mathbb{E}_{S_0}\left[\mathbb{V}\left(G_k^{\text{PDIS}}\left(\tau_{0:T-1}^{\hat{\mu}_{0:T-1}}\right) \mid S_0 = s\right)\right] + \frac{1}{n}\mathbb{V}_{S_0}(v_{\pi,0}(S_0)). && \text{(Theorem 6)}
\end{aligned}$$

Similarly, we have

$$
\begin{aligned}
&\mathbb{V}\left(E_{0:T-1}^{\mathrm{on},\pi^{(k)}}\right) && \text{(C.20)}\\
=&\mathbb{V}\left(\frac{\sum_{i=1}^{n_k} G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{[\pi_{0:T-1}^{(k)},i]}\right)}{n_k}\right) && \text{(By (5.24))}\\
=&\frac{1}{{n_k}^2}\mathbb{V}\left(\sum_{i=1}^{n_k} G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{[\pi_{0:T-1}^{(k)},i]}\right)\right)\\
=&\frac{1}{n_k}\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\pi_{0:T-1}^{(k)}}\right)\right) && \text{(Independence of } \tau_{0:T-1}^{[\pi_{0:T-1}^{(k)},i]}\text{)}\\
=&\frac{1}{n_k}\mathbb{E}_{S_0}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\pi_{0:T-1}^{(k)}}\right)\mid S_0=s\right)\right]+\frac{1}{n_k}\mathbb{V}_{S_0}\left(\mathbb{E}\left[G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\pi_{0:T-1}^{(k)}}\right)\mid S_0=s\right]\right)\\
& && \text{(Law of total variance)}\\
=&\frac{1}{n_k}\mathbb{E}_{S_0}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\pi_{0:T-1}^{(k)}}\right)\mid S_0=s\right)\right]+\frac{1}{n_k}\mathbb{V}_{S_0}(v_{\pi,0}(S_0)).\\
& && \text{(Theorem 6 and on-policy)}
\end{aligned}
$$

With the manipulated sufficient condition in (C.16), we present the following lemma.

Lemma 42. *Under the condition in* (5.26)*,* $\forall k,t,s,$

$$
\begin{aligned}
&\frac{n_k}{n}\mathbb{V}\left(G_k^{PDIS}\left(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}\right)\mid S_t=s\right)\\
\leq&\mathbb{V}\left(G_k^{PDIS}\left(\tau_{t:T-1}^{\pi_{t:T-1}^{(k)}}\right)\mid S_t=s\right).
\end{aligned}
$$

Proof. We proceed via induction. $\forall k$, $\forall s$, when $t=T-1$,

$$
\begin{aligned}
&\frac{n_k}{n}\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}\right)\mid S_t=s\right)\\
=&\frac{n_k}{n}\left[\mathbb{E}_{A_t\sim\mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 q_{\pi^{(k)},t}(S_t,A_t)^2\mid S_t=s\right]-v_{\pi^{(k)},t}(s)^2\right] && \text{(Lemma 40)}\\
=&\frac{n_k}{n}\left[\mathbb{E}_{A_t\sim\mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 \hat{q}_{\pi^{(k)},t}(S_t,A_t)\mid S_t=s\right]-v_{\pi^{(k)},t}(s)^2\right] && \text{(By (5.18))}\\
\leq&\frac{n_k}{n}\left[\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|s)\sqrt{\hat{q}_{\pi^{(k)},t}(s,a)}\right)^2-v_{\pi^{(k)},t}(s)^2\right] && \text{(By (C.18))}\\
\leq&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t,A_t)\mid S_t=s\right]-v_{\pi^{(k)},t}(s)^2 && \text{(By (C.17))}\\
=&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\pi_{t:T-1}^{(k)}}\right)\mid S_t=s\right). && \text{(Lemma 41)}
\end{aligned}
$$

For $t \in [T-2]$,

$$
\begin{aligned}
&\frac{n_k}{n}\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}\right) \mid S_t = s\right) \\
=&\frac{n_k}{n}\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\hat{\mu}_{t+1:T-1}}\right) \mid S_t\right) \mid S_t, A_t\right]\right.\right. \\
&\left.\left.+\nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - \frac{n_k}{n}v_{\pi^{(k)},t}(S_t)^2 && \text{(Lemma 40)} \\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\mathbb{E}_{S_{t+1}}\left[\frac{n_k}{n}\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\hat{\mu}_{t+1:T-1}}\right) \mid S_t\right) \mid S_t, A_t\right] \mid S_t\right] \\
&+\frac{n_k}{n}\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - \frac{n_k}{n}v_{\pi^{(k)},t}(S_t)^2 \\
&&\text{(Linearity of Expectation)} \\
\leq&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\pi^{(k)}_{t+1:T-1}}\right) \mid S_t\right) \mid S_t, A_t\right]\right) \mid S_t\right] \\
&+\frac{n_k}{n}\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - \frac{n_k}{n}v_{\pi^{(k)},t}(S_t)^2 \\
&&\text{(Indutive Hypothesis)} \\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\pi^{(k)}_{t+1:T-1}}\right) \mid S_t\right) \mid S_t, A_t\right]\right.\right. \\
&\left.\left.+\nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] \\
&+\left(\frac{n_k}{n}-1\right)\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - \frac{n_k}{n}v_{\pi^{(k)},t}(S_t)^2 \\
=&\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\hat{q}_{\pi^{(k)},t}(S_t,A_t) \mid S_t\right] \\
&-\frac{n-n_k}{n}\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - \frac{n_k}{n}v_{\pi^{(k)},t}(S_t)^2 \\
&&\text{(By (5.18))} \\
\leq&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|S_t)\sqrt{\hat{q}_{\pi^{(k)},t}(a|S_t)}\right)^2 \\
&-\frac{n-n_k}{n}\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - \frac{n_k}{n}v_{\pi^{(k)},t}(S_t)^2 \\
&&\text{(by (C.18))} \\
=&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|S_t)\sqrt{\hat{q}_{\pi^{(k)},t}(a|S_t)}\right)^2 \\
&-\frac{n-n_k}{n}\left(\mathbb{E}_{A_t\sim\hat{\mu}_t}\left[{\rho^{\pi^{(k)},\hat{\mu}}}^2\left(\nu_{\pi^{(k)},t}(S_t,A_t) + q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2\right) - v_{\pi^{(k)},t}(S_t)^2 \\
\leq&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t,A_t) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 && \text{(By (C.16))} \\
=&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\pi^{(k)}_{t:T-1}}\right) \mid S_t\right). && \text{(Lemma 41)}
\end{aligned}
$$

□

Now, we are ready to present the proof of Theorem 8.

$$
\begin{aligned}
&\mathbb{V}\left(E_{0:T-1}^{\mathrm{off},\pi^{(k)}}\right) \\
=&\frac{1}{n}\mathbb{E}_{S_0}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\hat{\mu}_{0:T-1}}\right) \mid S_0=s\right)\right]+\frac{1}{n}\mathbb{V}_{S_0}(v_{\pi,0}(S_0)) && \text{(By (C.19))}\\
\leq&\frac{1}{n}\mathbb{E}_{S_0}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\hat{\mu}_{0:T-1}}\right) \mid S_0=s\right)\right]+\frac{1}{n_k}\mathbb{V}_{S_0}(v_{\pi,0}(S_0)) && (n_k\leq n)\\
=&\frac{1}{n_k}\cdot\frac{n_k}{n}\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\hat{\mu}_{0:T-1}}\right) \mid S_0=s\right)+\frac{1}{n_k}\mathbb{V}_{S_0}(v_{\pi,0}(S_0))\\
\leq&\frac{1}{n_k}\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{0:T-1}^{\pi_{0:T-1}^{(k)}}\right) \mid S_0=s\right)+\frac{1}{n_k}\mathbb{V}_{S_0}(v_{\pi,0}(S_0)) && \text{(Lemma 42)}\\
=&\mathbb{V}\left(E_{0:T-1}^{\mathrm{on},\pi^{(k)}}\right). && \text{(By (C.20))}
\end{aligned}
$$

C.1.8 Proof of Theorem 9

Proof. We prove it using induction. When $t=T-1$, $\forall k$, $\forall s$,

$$
\begin{aligned}
&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}\right) \mid S_t=s\right)\\
=&\mathbb{E}_{A_t\sim\mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 q_{\pi^{(k)},t}(S_t,A_t)^2 \mid S_t\right]-v_{\pi^{(k)},t}(S_t)^2 && \text{(Lemma 40)}\\
=&\mathbb{E}_{A_t\sim\mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2 \hat{q}_{\pi^{(k)},t}(S_t,A_t) \mid S_t\right]-v_{\pi^{(k)},t}(S_t)^2 && \text{(By (5.18))}\\
\leq&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|S_t)\sqrt{\hat{q}_{\pi^{(k)},t}(S_t,a)}\right)^2-v_{\pi^{(k)},t}(S_t)^2 && \text{(By (C.18))}\\
\leq&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t,A_t) \mid S_t\right]-v_{\pi^{(k)},t}(S_t)^2 && \text{(By (5.27))}\\
=&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\pi_{t:T-1}^{(k)}}\right) \mid S_t\right). && \text{(Lemma 41)}
\end{aligned}
$$

When $t \in [T-2]$,

$$
\begin{aligned}
&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\hat{\mu}_{t:T-1}}\right) \mid S_t = s\right) \\
=&\mathbb{E}_{A_t\sim\mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\hat{\mu}_{t+1:T-1}}\right) \mid S_{t+1}\right) \mid S_t, A_t\right]\right.\right. \\
&\left.\left.+\nu_{\pi^{(k)},t}(S_t,A_t)+q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 && \text{(Lemma 40)} \\
\leq&\mathbb{E}_{A_t\sim\mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\pi_{t+1:T-1}^{(k)}}\right) \mid S_{t+1}\right) \mid S_t, A_t\right]\right.\right. \\
&\left.\left.+\nu_{\pi^{(k)},t}(S_t,A_t)+q_{\pi^{(k)},t}(S_t,A_t)^2\right) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 && \text{(Inductive Hypothesis)} \\
=&\mathbb{E}_{A_t\sim\mu_t}\left[{\rho_t^{\pi^{(k)},\mu}}^2\hat{q}_{\pi^{(k)},t}(S_t,A_t) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 && \text{(By (5.18))} \\
\leq&\sqrt{\frac{\overline{\eta}_t}{\underline{\eta}_t}}\left(\sum_a \pi_t^{(k)}(a|S_t)\sqrt{\hat{q}_{\pi^{(k)},t}(S_t,a)v}\right)^2 - v_{\pi^{(k)},t}(S_t)^2 && \text{(By (C.18))} \\
\leq&\mathbb{E}_{A_t\sim\pi_t^{(k)}}\left[\hat{q}_{\pi^{(k)},t}(S_t,A_t) \mid S_t\right] - v_{\pi^{(k)},t}(S_t)^2 && \text{(By (5.27))} \\
=&\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t:T-1}^{\pi_{t:T-1}^{(k)}}\right) \mid S_t\right). && \text{(Lemma 41)}
\end{aligned}
$$

□

C.2 Experiment Details

C.2.1 Learning Closed-Form Behavior Policy

In this section, we present an efficient algorithms to learn the closed-form optimal behavior policy $\hat{\mu}$ with previously logged offline data. By (5.19), $\hat{\mu}$ is defined as $\hat{\mu}_t(a|s) \propto \sqrt{\sum_{k=1}^K \pi_t^{(k)}(a|s)\hat{q}_{\pi^{(k)},t}(s,a)^2}$, where for each target policy k, $\hat{q}_{\pi^{(k)}}$ is defined in (5.18) as

$$
\hat{q}_{\pi^{(k)},t}(s,a) \doteq q_{\pi^{(k)},t}(s,a)^2 + \nu_{\pi^{(k)},t}(s,a) + \textstyle\sum_{s'} p(s'|s,a)\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\pi_{t+1:T-1}^{(k)}}\right) \mid S_{t+1}\right).
$$

Learning $\hat{\mu}$ from this perspective is very inefficient because it requires approximations of the complex variance term $\mathbb{V}\left(G_k^{\mathrm{PDIS}}\left(\tau_{t+1:T-1}^{\pi_{t+1:T-1}^{(k)}}\right) \mid S_{t+1}\right)$ regarding future trajectory. To solve this problem, we restate the recursive expression of $\hat{q}$ in the form of a Bellman equation (Tamar et al., 2016; O'Donoghue et al., 2018; Sherstan et al., 2018) from Liu and Zhang (2024) and give its proof.

Theorem 22. *For any target policy $\pi^{(k)}$, define*

$$
\hat{r}_{\pi^{(k)},t}(s,a) \doteq 2r(s,a)q_{\pi^{(k)},t}(s,a) - r^2(s,a). \tag{C.21}
$$

Then $\hat{q}_{\pi^{(k)},t}(s,a) = \hat{r}_{\pi^{(k)},t}(s,a)$ *for* $t = T-1$ *and otherwise*

$$\hat{q}_{\pi^{(k)},t}(s,a) = \hat{r}_{\pi^{(k)},t}(s,a) + \textstyle\sum_{s',a'} p(s'|s,a)\pi^{(k)}_{t+1}(a'|s')\hat{q}_{\pi^{(k)},t+1}(s',a'). \tag{C.22}$$

Proof. For any k, for $t = T-1$, we have

$$\begin{aligned}
\hat{q}_{\pi^{(k)},t}(s,a) &= q_{\pi^{(k)},t}(s,a)^2 && \text{(Definition of } \hat{q}_{\pi^{(k)},t} \text{ (5.18))} \\
&= \hat{r}_{\pi^{(k)},t}(s,a). && \text{(By } q_{\pi^{(k)},T-1}(s,a) = r(s,a) \text{ and (C.21))}
\end{aligned}$$

For $t \in [T-2]$, we have

$$\begin{aligned}
&\hat{q}_{\pi^{(k)},t}(s,a) \\
=&q_{\pi^{(k)},t}(s,a)^2 + \nu_{\pi^{(k)},t}(s,a) + \sum_{s'} p(s'|s,a)\mathbb{V}\left(G^{\text{PDIS}}_k\left(\tau^{\pi^{(k)}_{t+1:T-1}}_{t+1:T-1}\right) \mid S_{t+1} = s'\right) \\
& && \text{(Definition of } \hat{q} \text{ (5.18))} \\
=&q_{\pi^{(k)},t}(s,a)^2 + \nu_{\pi^{(k)},t}(s,a) + \sum_{s'} p(s'|s,a)\left(\mathbb{E}_{A_{t+1}\sim\pi^{(k)}_{t+1}}\left[\hat{q}_{\pi^{(k)},t+1}(S_{t+1},A_{t+1}) \mid S_{t+1} = s'\right]\right. \\
&\left. -v_{\pi^{(k)},t+1}(S_{t+1})^2\right) && \text{(Lemma 41)} \\
=&\nu_{\pi^{(k)},t}(s,a) + q_{\pi^{(k)},t}(s,a)^2 - \sum_{s'} p(s'|s,a)v_{\pi^{(k)},t+1}(s')^2 + \sum_{s',a'} p(s'|s,a)\pi^{(k)}_{t+1}(a'|s')\hat{q}_{\pi^{(k)},t+1}(s',a') \\
=&-\left(\mathbb{E}[v_{\pi^{(k)},t+1}(S_{t+1}) \mid S_t = s, A_t = a]\right)^2 + q_{\pi^{(k)},t}(s,a)^2 + \sum_{s',a'} p(s'|s,a)\pi^{(k)}_{t+1}(a'|s')\hat{q}_{\pi^{(k)},t+1}(s',a') \\
& && \text{(Definition of } \nu \text{ (5.16))} \\
=&-\left(q_{\pi^{(k)},t}(s,a) - r(s,a)\right)^2 + q_{\pi^{(k)},t}(s,a)^2 + \sum_{s',a'} p(s'|s,a)\pi^{(k)}_{t+1}(a'|s')\hat{q}_{\pi^{(k)},t+1}(s',a') \\
=&2r(s,a)q_{\pi^{(k)},t}(s,a) - r(s,a)^2 + \sum_{s',a'} p(s'|s,a)\pi^{(k)}_{t+1}(a'|s')\hat{q}_{\pi^{(k)},t+1}(s',a') \\
=&\hat{r}_{\pi^{(k)},t}(s,a) + \sum_{s',a'} p(s'|s,a)\pi^{(k)}_{t+1}(a'|s')\hat{q}_{\pi^{(k)},t+1}(s',a'),
\end{aligned}$$

which completes the proof. □

This derivation enables the implementation of any off-the-shelf offline policy evaluation methods to learn $\hat{q}_{\pi^{(k)}}$, after which the behavior policy $\hat{\mu}$ can be computed easily with (5.19). For generality, we consider the behavior policy agnostic offline learning setting (Nachum et al., 2019), where the offline data in the form of $\{(t_i, s_i, a_i, r_i, s'_i)\}_{i=1}^m$ consists of m previously logged data tuples. In the i-th data tuple, t_i is the time step, s_i is the state at time step t_i, a_i is the action executed on state s_i, r_i is the sampled

reward, and s_i' is the successor state. Those tuples can be generated by one or more, known or unknown behavior policies. Those tuples do not need to form a complete trajectory.

In this work, we use Fitted Q-Evaluation (FQE, Le et al. (2019)) as a demonstration, but our algorithm can incorporate any state-of-the-art offline policy evaluation methods to approximate $\hat{q}_\pi^{(k)}$. To learn $\hat{r}_\pi^{(k)}$, it is sufficient to learn q, in which FQE can be applied. Then, FQE is invoked to learn an approximation of $\hat{q}_\pi^{(k)}$. We refer the reader to Algorithm 2 for a detailed exposition of our algorithm. In practice, we split the offline data into training sets and test sets to tune all the hyperparameters offline in Algorithm 2.

C.2.2 GridWorld

For a Gridworld with size m^3, we set its width, height, and time horizon T all to be m. We test Gridworlds with $m^3 = 1,000$ and $m^3 = 27,000$ states. The action space contains four possible actions: up, down, left, and right. After taking an action, the agent has a probability of 0.9 to move accordingly and a probability of 0.1 to move uniformly at random. If running into a boundary, the agent stays in the current position. The reward function $r(s, a)$ is randomly generated.

We generate target policies using the proximal policy optimization (PPO) algorithm (Schulman et al., 2017) with the default parameters in CleanRL (Huang et al., 2022). We choose PPO just for a demonstration. Our method copes with any other deep RL algorithm. We randomly draw 10 policies in a randomly chosen time step interval. We obtain the ground truth policy performance for each target policy by executing on-policy Monte Carlo evaluation for 10^6 total episodes. Our offline dataset includes 10^4 episodes from various policies with a wide range of performances. We execute Algorithm 1 to learn our tailored behavior policy. When approximating q and $\hat{q}$, we use Fitted Q-Evaluation (Le et al., 2019). We use a one-hidden-layer neural network for Fitted Q-Evaluation. We test the neural network size for Fitted Q-Evaluation with $[64, 128, 256]$ and choose 64 as the final size. We test the learning rate for Adam optimizer with $[1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}]$ and choose to use the default learning rate $1e^{-3}$ as learning rate for Adam optimizer (Kingma and Ba, 2015).

The results for each target policy are shown in Figure C.1 and Figure C.2 in terms of the *relative error* against total *samples*, as described in the main text. Notably, for the On-policy Monte Carlo estimator and the ODI estimator (Liu and Zhang, 2024), samples for each single target policy $\pi^{(k)}$ are collected once in every $K = 10$ total sample steps. Smooth lines are plotted through interpolation. Each line in Figure C.1

and Figure C.2 is averaged over 900 different runs (30 groups of target policies, each having 30 independent runs), indicating strong statistical significance. Our method (MPE) *consistently outperforms* all other estimators for the evaluation of *every* single target policy, demonstrating *state-of-the-art performance*.

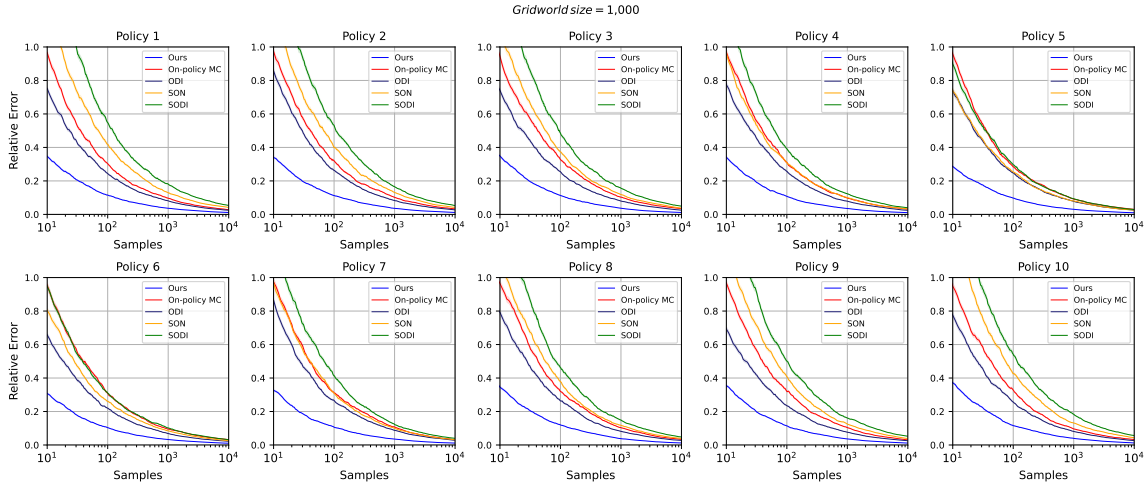

Figure C.1: Results on Gridworld. Each curve is averaged over 900 runs (the corresponding target policies from 30 groups, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves because they are too small.

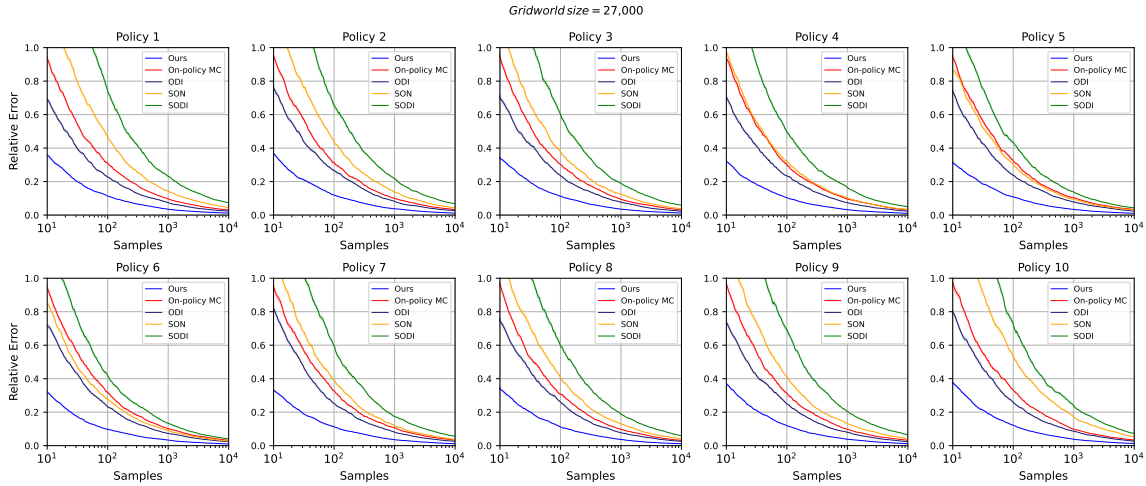

Figure C.2: Results on Gridworld. Each curve is averaged over 900 runs (the corresponding target policies from 30 groups, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves because they are too small.

C.2.3 MuJoCo

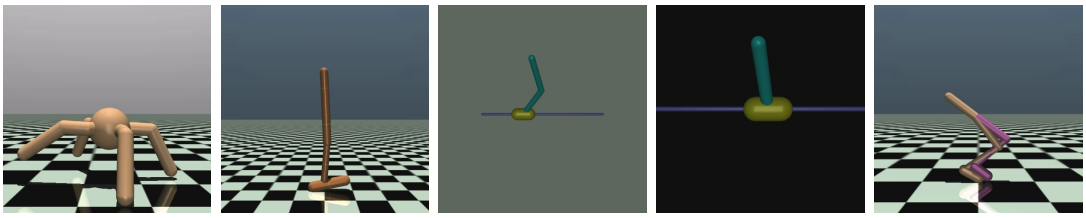

Figure C.3: MuJoCo robot simulation tasks (Todorov et al., 2012). Pictures are adapted from (Liu and Zhang, 2024). Environments from the left to the right are Ant, Hopper, InvertedDoublePendulum, InvertedPendulum, and Walker.

In experiments of MuJoCo robot simulation tasks (Todorov et al., 2012), we use the same method for obtaining offline data, randomly generating target policies, training behavior policies, and applying the same hyperparameters as in the Gridworld experiment. We discretize the first dimension of MuJoCo action space in our experiment. The policies of remaining dimensions are obtained during PPO (Schulman et al., 2017) training process, and deemed as part of the environment. The following table offers an additional interpretation to Figure 5.2.

Env ID	**Ours**	On-policy MC	ODI	SON	SODI
Ant	**0.115**	1.000	0.606	1.144	1.548
Hopper	**0.114**	1.000	0.580	1.287	1.413
InvertedDoublePendulum	**0.111**	1.000	0.494	0.882	1.582
InvertedPendulum	**0.124**	1.000	0.565	0.889	1.250
Walker	**0.094**	1.000	0.590	0.759	1.056

Table C.1: Relative variance of estimators on MuJoCo environments. The relative variance is defined as the variance of each estimator divided by the variance of the on-policy Monte Carlo estimator. Numbers are averaged over 900 independent runs (30 groups of target policies, each having 30 independent runs).

Env ID	**Ours**	On-policy MC	ODI	SON	SODI
Ant	**125**	1000	626	1171	1701
Hopper	**115**	1000	583	1253	1413
InvertedDoublePendulum	**103**	1000	464	835	1547
InvertedPendulum	**111**	1000	530	916	1166
Walker	**97**	1000	542	749	1033

Table C.2: Episodes needed to achieve the same estimation accuracy that on-policy Monte Carlo achieves with 1000 episodes on MuJoCo environments. Numbers are averaged over 900 independent runs (30 groups of target policies, each having 30 independent runs) and their standard errors are shown in Figure 6.2.

Appendix D

Appendix for Chapter 7

D.1 Proofs

D.1.1 Proof of Lemma 11

Proof. $\forall s$, $\forall \mu \in \Lambda$,

$$\begin{aligned}
\mathbb{E}_{a\sim\mu}[\rho(a|s)r(s,a)] =& \sum_{a\in\{a|\mu(a|s)>0\}} \mu(a|s)\frac{\pi(a|s)}{\mu(a|s)}r(s,a) \\
=& \sum_{a\in\{a|\mu(a|s)>0\}} \pi(a|s)r(s,a) \\
=& \sum_{a\in\{a|\mu(a|s)>0\}} \pi(a|s)r(s,a) + \sum_{a\in\{a|\mu(a|s)=0\}} \pi(a|s)r(s,a) \quad (\mu\in\Lambda) \\
=& \sum_{a} \pi(a|s)r(s,a) \\
=& \mathbb{E}_{a\sim\pi}[r(s,a)].
\end{aligned}$$

□

D.1.2 Proof of Lemma 12

Proof. To prove Lemma 12, we express the objective function as

$$\mathbb{E}_{a\sim\mu}[\rho(a|s)^2 r(s,a)^2] = \sum_{a\in\{a|\mu(a|s)>0\}} \frac{\pi(a|s)^2 r(s,a)^2}{\mu(a|s)}.$$

To prove the problem is convex, we begin by examining the feasible set of each constraint separately.

In the first constraint of Λ (7.2),

$$\forall s, a, \mu(a|s) = 0 \implies \pi(a|s)r(s,a) = 0. \tag{D.1}$$

The feasible set of (D.1) is a linear subspace of $\mathbb{R}^{|\mathcal{A}|}$ defined by a set of linear equations. Thus, this feasible set is convex.

Next, we decompose the other constraint of Λ (7.2), $\mu(\cdot|s) \in \Delta^{|\mathcal{A}|-1}\ \forall s$, into two subconstraints:

$$\sum_a \mu(a|s) = 1, \tag{D.2}$$

$$\forall a, \mu(a|s) \geq 0. \tag{D.3}$$

For all s, the feasible set in (D.2) can be written in the vector form as

$$\mathbf{1}^T \overrightarrow{\mu_s} = 1, \tag{D.4}$$

where $\mathbf{1} \in \mathbb{R}^{|\mathcal{A}|}$ is the vector of ones defined as

$$\mathbf{1} \doteq \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

and $\overrightarrow{\mu_s} \in \mathbb{R}^{|\mathcal{A}|}$ is defined as

$$\overrightarrow{\mu_s} \doteq \begin{bmatrix} \mu(a_1|s) \\ \vdots \\ \mu(a_{|\mathcal{A}|}|s) \end{bmatrix}.$$

Since (D.4) is linear, the constraint (D.2) is affine and thus convex (Boyd et al., 2004).

For all s, the feasible set of (D.3) is the non-negative orthant, defined as

$$\mathbb{R}_+^{|\mathcal{A}|} \doteq \{\mu(\cdot|s) \in \mathbb{R}^{|\mathcal{A}|} \mid \mu(a|s) \geq 0, \forall a\}.$$

Since the non-negative orthant forms a convex cone and is known to be a convex set (Boyd et al., 2004), we conclude that this constraint's feasible set is convex.

Next, we define the vector of costs for all s as

$$\mathbf{c}_s \doteq \begin{bmatrix} c(s, a_1) \\ \vdots \\ c(s, a_{|\mathcal{A}|}) \end{bmatrix}.$$

Then, for all ϵ and s, the safety constraint (7.8) can be rewritten as

$$\mathbf{c}_s^\top \overrightarrow{\mu_s} \leq \delta_\epsilon(s),$$

which is a linear inequality in μ. Thus, its feasible set is in a convex half-space. Because all the constraints are convex, we conclude that the feasible set $\mathcal{F}$ in (7.9) is convex.

Finally, we examine the minimization objective (7.7), where π and r are fixed and independent of the behavior policy μ. For all s, we express the objective function as

$$\mathbb{E}_{a\sim\mu}[\rho(a|s)^2 r(s,a)^2] = \sum_{a\in\{a|\mu(a|s)>0\}} \frac{\pi(a|s)^2 r(s,a)^2}{\mu(a|s)}.$$

Then, for each a, we decompose the objective function as

$$f_a(\mu(a|s)) \doteq \frac{\pi(a|s)^2 r(s,a)^2}{\mu(a|s)}. \tag{D.5}$$

Taking the first and second derivatives of f_a, we get

$$f_a'(\mu(a|s)) = -\frac{\pi(a|s)^2 r(s,a)^2}{\mu(a|s)^2},$$
$$f_a''(\mu(a|s)) = \frac{2\pi(a|s)^2 r(s,a)^2}{\mu(a|s)^3}.$$

Since $\forall s, a,\ f_a''(\mu(a|s)) \geq 0$, we know that (D.5) is convex for all a. Then, as a summation of convex functions, (7.7) is also convex. In conclusion, by the convexity of the feasible set $\mathcal{F}$ and the objective function (7.7), we obtain the convexity of the constrained optimization problem in Lemma 12.

For feasibility, note that by Lemma 13, $\pi \in \mathcal{F}$, which is the feasible set. Thus, we confirm the feasibility in Lemma 12.

□

D.1.3 Proof of Lemma 1

Proof. We proceed via induction. For $t = T-1$, we have

$$\begin{aligned}
\mathbb{E}\left[G^{\text{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right] &= \mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] = \mathbb{E}\left[\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right] \\
&= \mathbb{E}_{A_t\sim\pi_t(\cdot|S_t)}\left[q_{\pi,t}(S_t, A_t) \mid S_t\right] && \text{(Lemma 11)} \\
&= v_{\pi,t}(S_t).
\end{aligned}$$

For $t \in [T-2]$, we have

$$
\begin{aligned}
&\mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t R_{t+1} + \rho_t G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] + \mathbb{E}\left[\rho_t G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t\right] \\
=&\mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t\sim\mu_t(\cdot|S_t), S_{t+1}\sim p(\cdot|S_t,A_t)}\left[\mathbb{E}\left[\rho_t G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right] \mid S_t\right] \\
& \qquad \text{(Law of total expectation)} \\
=&\mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t\sim\mu_t(\cdot|S_t), S_{t+1}\sim p(\cdot|S_t,A_t)}\left[\rho_t \mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right] \mid S_t\right] \\
& \qquad \text{(Conditional independence and Markov property)} \\
=&\mathbb{E}\left[\rho_t R_{t+1} \mid S_t\right] + \mathbb{E}_{A_t\sim\mu_t(\cdot|S_t), S_{t+1}\sim p(\cdot|S_t,A_t)}\left[\rho_t v_{\pi,t+1}(S_{t+1}) \mid S_t\right] \\
& \qquad \text{(Inductive hypothesis)} \\
=&\mathbb{E}_{A_t\sim\mu_t(\cdot|S_t)}\left[\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right] \qquad \text{(Definition of } q_{\pi,t}\text{)} \\
=&\mathbb{E}_{A_t\sim\pi_t(\cdot|S_t)}\left[q_{\pi,t}(S_t, A_t) \mid S_t\right] \qquad \text{(Lemma 11)} \\
=&v_{\pi,t}(S_t),
\end{aligned}
$$

which completes the proof. □

D.1.4 Proof of Theorem 16

Proof. We first define the set of feasible policies as

$$\mathcal{F} \doteq \{\mu \in \Lambda \mid \forall \epsilon, t, s, \mathbb{E}_{a\sim\mu_t}[v_{\mu,t}^c(s)] \le \delta_{\epsilon,t}(s)\}. \tag{D.6}$$

We begin by examining each constraint. In the first constraint of Λ (7.11),

$$\forall t, s, a, \mu_t(a|s) = 0 \implies \pi_t(a|s) q_{\pi,t}(s,a) = 0. \tag{D.7}$$

The feasible set of (D.7) is a linear subspace of $\mathbb{R}^{|\mathcal{A}|}$ defined by a set of linear equations. Thus, this feasible set is convex.

Next, we decompose the other constraint of Λ (7.11), $\mu_t(\cdot|s) \in \Delta^{|\mathcal{A}|-1}$, into two constraints:

$$\sum_a \mu_t(a|s) = 1, \tag{D.8}$$

$$\forall a, \mu_t(a|s) \ge 0. \tag{D.9}$$

For all t and s, in (D.8), the feasible set can be written as

$$\mathbf{1}^\top \overrightarrow{\mu_{s,t}} = 1, \tag{D.10}$$

where $\mathbf{1} \in \mathbb{R}^{|\mathcal{A}|}$ is the vector of ones and $\overrightarrow{\mu_{s,t}} \in \mathbb{R}^{|\mathcal{A}|}$ is defined as

$$\overrightarrow{\mu_{s,t}} \doteq \begin{bmatrix} \mu_t(a_1|s) \\ \vdots \\ \mu_t(a_{|\mathcal{A}|}|s) \end{bmatrix}.$$

Since (D.10) is linear, the feasible set of constraint (D.8) is affine and thus convex (Boyd et al., 2004).

For all t and s, the feasible set for the constraint in (D.9) is the non-negative orthant, defined as

$$\mathbb{R}_+^{|\mathcal{A}|} \doteq \{\mu_t(\cdot|s) \in \mathbb{R}^{|\mathcal{A}|} \mid \mu_t(a|s) \geq 0, \forall a\}.$$

Since the non-negative orthant forms a convex cone and is known to be a convex set (Boyd et al., 2004), we conclude that this constraint is convex.

Next, we define the vector of the state-action value function for the cost c for each s as

$$\mathbf{q}_{\mu,t} \doteq \begin{bmatrix} q^c_{\mu,t}(s, a_1) \\ \vdots \\ q^c_{\mu,t}(s, a_{|\mathcal{A}|}) \end{bmatrix}.$$

Then, for all ϵ, t and s, the safety constraint (7.16) can be rewritten as

$$\mathbf{q}_{\mu,t}^\top \overrightarrow{\mu_{s,t}} \leq \delta_{\epsilon,t}(s),$$

which is a linear inequality in μ_t. Thus, its feasible set is a convex half-space. Because all the constraints' feasible sets are convex, we conclude that the feasible set $\mathcal{F}$ in (D.6) is convex.

To prove Theorem 16, we express the objective function as

$$\mathbb{E}_{a\sim\mu_t}[\rho_t^2 \tilde{r}_t(s,a)] = \sum_{a\in\{a|\mu_t(a|s)>0\}} \frac{\pi_t(a|s)^2 \tilde{r}_t(s,a)}{\mu_t(a|s)},$$

where $\tilde{r}$ in (7.14) is defined as

$$\tilde{r}_t(s,a) \doteq \begin{cases} r_{\pi,t}(s,a)^2 & t = T-1, \\ \nu_{\pi,t}(s,a) + q_{\pi,t}(s,a)^2 + \mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid s,a\right] & t \in [T-2]. \end{cases}$$

Here, $\tilde{r}_t$ can be learned with logged offline data, as detailed in Algorithm 4, and it is unrelated to μ_t. Then, for each a, we decompose the objective function as

$$f_a(\mu_t(a|s)) \doteq \frac{\pi_t(a|s)^2 \tilde{r}_t(s,a)}{\mu_t(a|s)}. \tag{D.11}$$

Taking the first and second derivatives of f_a, we get

$$f'_a(\mu_t(a|s)) = -\frac{\pi_t(a|s)^2 \tilde{r}(s,a)}{\mu_t(a|s)^2},$$
$$f''_a(\mu_t(a|s)) = \frac{2\pi_t(a|s)^2 \tilde{r}(s,a)}{\mu_t(a|s)^3}.$$

Notice that the extended reward $\tilde{r}$ defined in (7.14) is non-negative, since all the summands are non-negative. Thus, $\forall t, s, a$, $f''_a(\mu_t(a|s)) \geq 0$, and we know that (D.11) is convex for all a. Then, as a summation of convex functions, (7.15) is also convex. In conclusion, by the convexity of the feasible set $\mathcal{F}$ and the objective function (7.15), we obtain the convexity of the constrained optimization problem in Theorem 16.

For feasibility, we show that the set of feasible policies (D.6) is non-empty. Because $\epsilon \in [0, \infty)$, for the safety constraint, we have

$$\mathbb{E}_{a\sim\pi_t}[v^c_{\mu,t}(s)] \leq (1+\epsilon)\mathbb{E}_{a\sim\pi_t}[v^c_{\mu,t}(s)] = \delta_{\epsilon,t}(s).$$

By the definition of Λ (7.11), $\forall t$, $\pi_t \in \Lambda$. Therefore, the set of feasible policies (D.6) is non-empty. Thus, the constrained optimization problem in Theorem 16 is feasible. □

D.1.5 Proof of Theorem 17

To prove Theorem 17, we first restate a recursive expression of the variance $\mathbb{V}\left(G^{\text{PDIS}}(\tau^{\mu_{t:T-1}}_{t:T-1}) \mid S_t\right)$ for all $\mu \in \Lambda$ from Liu and Zhang (2024), and present its proof for completeness.

Lemma 43. *For any $\mu \in \Lambda$, for $t = T-1$,*

$$\mathbb{V}\left(G^{PDIS}(\tau^{\mu_{t:T-1}}_{t:T-1}) \mid S_t\right) = \mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^2 q^2_{\pi,t}(S_t, A_t) \mid S_t\right] - v^2_{\pi,t}(S_t),$$

for $t \in [T-2]$,

$$\begin{aligned}&\mathbb{V}\left(G^{PDIS}(\tau^{\mu_{t:T-1}}_{t:T-1}) \mid S_t\right)\\ =&\mathbb{E}_{A_t\sim\mu_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{PDIS}(\tau^{\mu_{t+1:T-1}}_{t+1:T-1}) \mid S_t\right) \mid S_t, A_t\right] + \nu_{\pi,t}(S_t, A_t) + q^2_{\pi,t}(S_t, A_t)\right) \mid S_t\right]\\ &- v^2_{\pi,t}(S_t).\end{aligned}$$

Proof. For completeness, we provide the proof from Liu and Zhang (2024). We proceed via induction. When $t = T-1$, we have

$$\begin{aligned}\mathbb{V}\left(G^{\text{PDIS}}(\tau^{\mu_{t:T-1}}_{t:T-1}) \mid S_t\right) =&\mathbb{V}\left(\rho_t r(S_t, A_t) \mid S_t\right)\\ =&\mathbb{V}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right)\\ =&\mathbb{E}_{A_t}\left[\rho_t^2 q_{\pi,t}(S_t, A_t)^2 \mid S_t\right] - v_{\pi,t}(S_t)^2,\end{aligned}$$

When $t \in [T-2]$, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right) && \text{(D.12)}\\
=&\mathbb{E}_{A_t}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t, A_t\right] \mid S_t\right)\\
& && \text{(Law of total variance)}\\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(r(S_t, A_t) + G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right]\\
&+ \mathbb{V}_{A_t}\left(\rho_t \mathbb{E}\left[r(S_t, A_t) + G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right] \mid S_t\right) && \text{(By (7.1))}\\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\rho_t \mathbb{E}\left[r(S_t, A_t) + G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right] \mid S_t\right)\\
& && \text{(Deterministic reward } r\text{)}\\
=&\mathbb{E}_{A_t}\left[\rho_t^2 \mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) \mid S_t\right] + \mathbb{V}_{A_t}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right).
\end{aligned}
$$

Further decomposing the first term, we have

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t\right) && \text{(D.13)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right) \mid S_t, A_t\right]\\
&+ \mathbb{V}_{S_{t+1}}\left(\mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_t, A_t, S_{t+1}\right] \mid S_t, A_t\right) && \text{(Law of total variance)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}\left(\mathbb{E}\left[G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right] \mid S_t, A_t\right)\\
& && \text{(Markov property)}\\
=&\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}\left(v_{\pi,t+1}(S_{t+1}) \mid S_t, A_t\right).\\
& && \text{(Lemma 1)}
\end{aligned}
$$

Then, plugging (D.13) back to (D.12) yields

$$
\begin{aligned}
&\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t:T-1}^{\mu_{t:T-1}}) \mid S_t\right)\\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}(v_{\pi,t}(S_{t+1}) \mid S_t = s, A_t = a)\right) \mid S_t\right]\\
&+ \mathbb{V}_{A_t}\left(\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right)\\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}(v_{\pi,t}(S_{t+1}) \mid S_t = s, A_t = a)\right) \mid S_t\right]\\
&+ \mathbb{E}_{A_t}\left[\rho_t^2 q_{\pi,t}(S_t, A_t)^2 \mid S_t\right] - \left(\mathbb{E}_{A_t}\left[\rho_t q_{\pi,t}(S_t, A_t) \mid S_t\right]\right)^2\\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \mathbb{V}_{S_{t+1}}(v_{\pi,t}(S_{t+1}) \mid S_t = s, A_t = a)\right) \mid S_t\right]\\
&+ \mathbb{E}_{A_t}\left[\rho_t^2 q_{\pi,t}(S_t, A_t)^2 \mid S_t\right] - v_{\pi,t}(S_t)^2, && \text{(Lemma 11)}\\
=&\mathbb{E}_{A_t}\left[\rho_t^2\left(\mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu_{t+1:T-1}}) \mid S_{t+1}\right) \mid S_t, A_t\right] + \nu_{\pi,t}(S_t, A_t) + q_{\pi,t}(S_t, A_t)^2\right) \mid S_t\right]\\
&- v_{\pi,t}(S_t)^2, && \text{(Definition of } \nu\text{)}
\end{aligned}
$$

which completes the proof. □

Then, with the extended reward $\tilde{r}$ in (7.14) defined as

$$\tilde{r}_t(s,a) \doteq \begin{cases} r_{\pi,t}(s,a)^2 & t = T-1, \\ \nu_{\pi,t}(s,a) + q_{\pi,t}(s,a)^2 + \mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid s,a\right] & t \in [T-2], \end{cases}$$

we can express the variance in a succinct form

$$\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) = \mathbb{E}_{a\sim\mu^*}[\rho_t^2 \tilde{r}_t(s,a)] - v_{\pi,t}(s)^2, \quad \forall s,t. \tag{D.14}$$

Now, we restate Theorem 17 and present its proof.

Theorem 17. *The behavior policy μ^* reduces variance compared with the on-policy evaluation method.*

$$\forall t,s, \mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) \leq \mathbb{V}\left(G^{PDIS}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right).$$

In Appendix D.1.4, we show that $\forall t$, $\pi_t \in \mathcal{F}$, where $\mathcal{F}$ in (D.6) is the set of feasible policies for the constrained optimization problem in Theorem 16. Recall that μ_t^* is defined as the optimal solution to the problem (7.15), i.e.,

$$\mu_t^* \doteq \operatorname*{argmin}_{\mu_t \in \mathcal{F}} \mathbb{E}_{a\sim\mu_t}[\rho_t^2 \tilde{r}(s,a)]. \tag{D.15}$$

Thus, $\forall t,s$,

$$\begin{aligned}
&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\mu^*_{t:T-1}}) \mid S_t = s\right) \\
=&\mathbb{E}_{a\sim\mu_t^*}[\rho_t^2 \tilde{r}_t(s,a)] - v_{\pi,t}(s)^2 && \text{(By (D.14))} \\
\leq&\mathbb{E}_{a\sim\pi_t}[\rho_t^2 \tilde{r}_t(s,a)] - v_{\pi,t}(s)^2 && \text{(By (D.15) and } \pi_t \in \mathcal{F}) \\
=&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{t:T-1}^{\pi_{t:T-1}}) \mid S_t = s\right), && \text{(By (D.14))}
\end{aligned}$$

which completes the proof.

D.1.6 Proof of Theorem 18

Proof. We first prove the variance reduction property.

$$\begin{aligned}
&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu^*_{0:T-1}})\right) \\
=&\mathbb{E}_{S_0}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu^*_{0:T-1}}) \mid S_0\right)\right] + \mathbb{V}_{S_0}\left(\mathbb{E}\left[G^{\text{PDIS}}(\tau_{0:T-1}^{\mu^*_{0:T-1}}) \mid S_0\right]\right) \\
& && \text{(Law of Total Variance)} \\
=&\mathbb{E}_{S_0}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\mu^*_{0:T-1}}) \mid S_0\right)\right] + \mathbb{V}_{S_0}\left(v_{\pi,0}(S_0)\right) && \text{(By Lemma 1 and } \mu^* \in \Lambda) \\
\leq&\mathbb{E}_{S_0}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\pi_{0:T-1}}) \mid S_0\right)\right] + \mathbb{V}_{S_0}\left(v_{\pi,0}(S_0)\right) && \text{(Theorem 17)} \\
=&\mathbb{E}_{S_0}\left[\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\pi_{0:T-1}}) \mid S_0\right)\right] + \mathbb{V}_{S_0}\left(\mathbb{E}\left[G^{\text{PDIS}}(\tau_{0:T-1}^{\pi_{0:T-1}}) \mid S_0\right]\right) \\
& && \text{(By Lemma 1 and } \pi \in \Lambda) \\
=&\mathbb{V}\left(G^{\text{PDIS}}(\tau_{0:T-1}^{\pi_{0:T-1}})\right). && \text{(Law of Total Variance)}
\end{aligned}$$

Next, we prove the safety constraint satisfaction.

$$
\begin{aligned}
& J^c(\mu^*) \\
=& \sum_s p_0(s) v^c_{\mu^*,0}(s) \\
=& \sum_s p_0(s) \mathbb{E}_{a\sim\mu^*_0}[q^c_{\mu^*,0}(s,a)] \\
\leq& \sum_s p_0(s) \delta_{\epsilon,0}(s) && \text{(Theorem 16)} \\
=& \sum_s p_0(s)(1+\epsilon) v^c_{\pi,0}(s) && \text{(By (7.13))} \\
=& (1+\epsilon) \sum_s p_0(s) v^c_{\pi,0}(s) \\
=& (1+\epsilon) J^c(\pi),
\end{aligned}
$$

which completes the proof. □

D.1.7 Proof of Lemma 15

Proof. $\forall s, a$, when $t = T-1$, $\tilde{r}_t(s,a) = r_{\pi,t}(s,a)^2$, as defined in (7.14). For $t \in [T-2]$,

$$
\begin{aligned}
&\tilde{r}_t(s,a) \\
=&\nu_{\pi,t}(s,a) + q_{\pi,t}(s,a)^2 + \mathbb{E}_{S_{t+1}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+1:T-1}^{\mu^*_{t+1:T-1}}) \mid S_{t+1}\right) \mid s,a\right] && \text{(By (7.14))} \\
=&\nu_{\pi,t}(s,a) + q_{\pi,t}(s,a)^2 \\
&+ \sum_{s'} p(s'|s,a)\left[\mathbb{E}_{A_{t+1}\sim\mu^*_{t+1}}\left[\rho_{t+1}^2\left(\mathbb{E}_{S_{t+2}}\left[\mathbb{V}\left(G^{\mathrm{PDIS}}(\tau_{t+2:T-1}^{\mu^*_{t+2:T-1}}) \mid S_{t+2}\right)\right] \mid S_{t+1}, A_{t+1}\right]\right.\right. \\
&\left.\left.+\nu_{\pi,t+1}(S_{t+1},A_{t+1}) + q_{\pi,t+1}(S_{t+1},A_{t+1})^2\right) \mid S_{t+1} = s'\right] - v_{\pi,t+1}(s')^2\Big] \\
& && \text{(By Lemma 43)} \\
=&\nu_{\pi,t}(s,a) + q_{\pi,t}(s,a)^2 + \sum_{s'} p(s'|s,a)\left[\mathbb{E}_{A_{t+1}\sim\mu^*_{t+1}}\left[\rho_{t+1}^2\tilde{r}_{\pi,t+1}(S_{t+1},A_{t+1}) \mid S_{t+1} = s'\right]\right. \\
&\left.- v_{\pi,t+1}(s')^2\right] && \text{(By (7.14))} \\
=&\nu_{\pi,t}(s,a) + q_{\pi,t}(s,a)^2 + \sum_{s',a'} p(s'|s,a)\left[\rho_{t+1}\pi_{t+1}(a'|s')\tilde{r}_{\pi,t+1}(s',a') - v_{\pi,t+1}(s')^2\right]. \\
=&\mathbb{V}_{S_{t+1}}\left(v_{\pi,t+1}(S_{t+1}) \mid S_t = s, A_t = a\right) + q_{\pi,t}(s,a)^2 \\
&+ \sum_{s',a'} p(s'|s,a)\left[\rho_{t+1}\pi_{t+1}(a'|s')\tilde{r}_{\pi,t+1}(s',a') - v_{\pi,t+1}(s')^2\right] && \text{(Definition of } \nu) \\
=&\mathbb{E}_{S_{t+1}}\left[v_{\pi,t+1}(S_{t+1})^2 \mid S_t = s, A_t = a\right] - \mathbb{E}_{S_{t+1}}\left[v_{\pi,t+1}(S_{t+1}) \mid S_t = s, A_t = a\right]^2 \\
&+ q_{\pi,t}(s,a)^2 + \sum_{s',a'} p(s'|s,a)\left[\rho_{t+1}\pi_{t+1}(a'|s')\tilde{r}_{\pi,t+1}(s',a') - v_{\pi,t+1}(s')^2\right] \\
=&\sum_{s'} p(s'|s,a)v_{\pi,t+1}(s')^2 - \left(q_{\pi,t}(s,a) - r(s,a)\right)^2 + q_{\pi,t}(s,a)^2 \\
&+ \sum_{s',a'} p(s'|s,a)\rho_{t+1}\pi_{t+1}(a'|s')\tilde{r}_{\pi,t+1}(s',a') - \sum_{s'} p(s'|s,a)v_{\pi,t+1}(s')^2 \\
=&2q_{\pi,t}(s,a)r(s,a) - r(s,a)^2 + \sum_{s',a'} p(s'|s,a)\rho_{t+1}\pi_{t+1}(a'|s')\tilde{r}_{\pi,t+1}(s',a') \\
=&2q_{\pi,t}(s,a)r(s,a) - r(s,a)^2 + \mathbb{E}_{s'\sim p, a'\sim\pi}\left[\frac{\pi_{t+1}(a'|s')}{\mu^*_{t+1}(a'|s')}\tilde{r}_{\pi,t+1}(s',a')\right].
\end{aligned}
$$

□

Environment Size	On-policy MC	**Ours**	ODI	ROS
1,000	1.000	**0.547**	0.460	0.953
27,000	1.000	**0.575**	0.484	0.987

Table D.1: Relative variance for estimators on Gridworld. The relative variance is defined as the variance of each estimator divided by the variance of the on-policy Monte Carlo estimator. Numbers are averaged over 900 independent runs (30 target policies, each having 30 independent runs). Standard errors are plotted in Figure 7.1.

Env Size	On-policy MC	**Ours**	ODI	ROS	Saved Cost Percentage
10	1000	**472**	738	1035	(1000 - 472)/1000 = **52.8%**
30	1000	**487**	765	1049	(1000 - 487)/1000 = **51.3%**

Table D.2: Cost needed to achieve the same estimation accuracy that on-policy Monte Carlo achieves with 1000 episodes on Gridworld. Each number is averaged over 900 runs. Standard errors are plotted in Figure 7.2.

D.2 Experiment Details

D.2.1 GridWorld

We conduct experiments on Gridworlds with $n^3 = 1,000$ and $n^3 = 27,000$ states, where for a Gridworld with size n^3, we set the width, height, and time horizon T all to be n. The action space contains four different possible actions: up, down, left, and right. After taking an action, the agent has a probability of 0.9 to move accordingly and a probability of 0.1 to move uniformly at random. When the agent runs into a boundary, it stays in its current position. We randomly generate the reward function $r(s, a)$ and cost function $c(s, a)$. We consider 30 randomly generated target policies with various performances. The ground truth policy performance is estimated by the on-policy Monte Carlo method, running each target policy for 10^6 episodes. We experiment two different sizes of the Gridworld with a number of $1,000$ and $27,000$ states.

The offline dataset of each environment contains a total of $1,000$ episodes generated by 30 policies with various performances. The performance of those policies ranges from completely random initialized policies to well-trained policies in each environment. For example, in Hopper, the performance of those 30 policies ranges from around 18 to around 2800. We let offline data be generated by various policies to simulate the fact that offline data are from different past collections.

We learn functions $q_{\pi,t}, q^c_{\pi,t}$, and $\hat{r}_{\pi,t}$ using Fitted Q-Evaluation algorithms (FQE,

Le et al. (2019)) by passing data tuples in $\mathcal{D}_\nu$ from $t = T - 1$ to 0. It is worth noticing that Fitted Q-Evaluation (FQE, Le et al. (2019)) is a different algorithm from Fitted Q-Improvement (FQI). Importantly, Fitted Q-Evaluation is not prone to overestimate the action-value function $q_{\pi,t}$ because it does not have any max operator and does not change the policy. All hyperparameters are tuned offline based on Fitted Q-learning loss. We leverage a one-hidden-layer neural network and test the neural network size with $[64, 128, 256]$. We then choose 64 as the final size. We also test the learning rate for Adam optimizer with $[1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}]$ and finally choose to use the default learning rate $1e^{-3}$ as learning rate for Adam optimizer (Kingma and Ba, 2015). For the benchmark algorithms, we use their reported hyperparameters (Zhong et al., 2022; Liu and Zhang, 2024). Each policy has 30 independent runs, resulting in a total of $30 \times 30 = 900$ runs. Thus, each curve in Figure 7.1, Figure 7.2 and each number in Table 7.1, Table D.1 and Table D.2 are averaged from 900 different runs over a wide range of policies, demonstrating a strong statistical significance.

D.2.2 MuJoCo

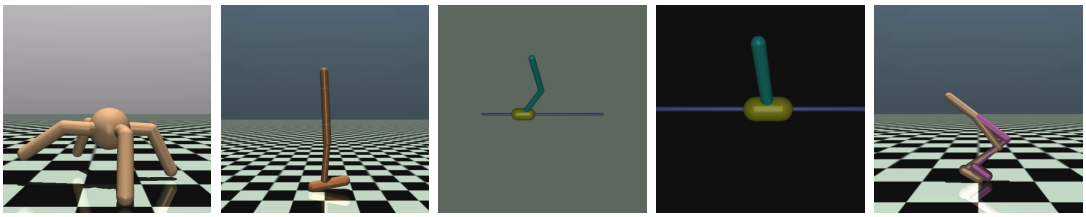

Figure D.1: MuJoCo robot simulation tasks (Todorov et al., 2012). Pictures are adapted from (Liu and Zhang, 2024). Environments from the left to the right are Ant, Hopper, InvertedDoublePendulum, InvertedPendulum, and Walker.

	On-policy MC	Ours	ODI	ROS
Ant	1.000	**0.835**	0.811	1.032
Hopper	1.000	**0.596**	0.542	1.005
I. D. Pendulum	1.000	**0.778**	0.724	0.992
I. Pendulum	1.000	**0.439**	0.351	0.900
Walker	1.000	**0.728**	0.696	0.908

Table D.3: Relative variance of estimators on MuJoCo. The relative variance is defined as the variance of each estimator divided by the variance of the on-policy Monte Carlo estimator. All numbers are averaged over 900 independent runs (30 target policies, each having 30 independent runs).

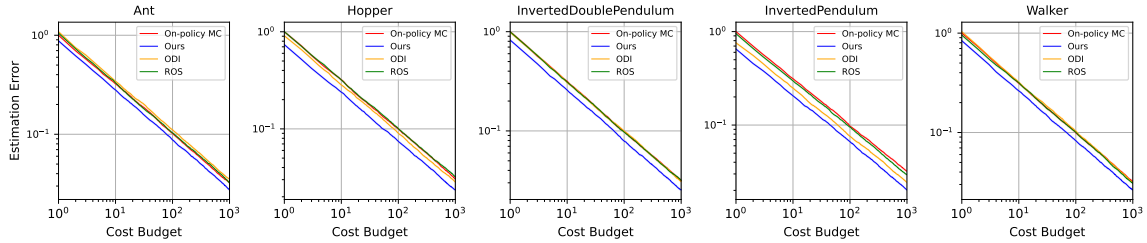

Figure D.2: Results on MuJoCo with log-scale y-axis to show the error does not converge. Each curve is averaged over 900 runs (30 target policies, each having 30 independent runs). Shaded regions denote standard errors and are invisible for some curves because they are too small.

	On-policy MC	Ours	ODI	ROS
Ant	1.000	**0.897**	1.397	1.033
Hopper	1.000	**0.930**	1.523	1.021
I. D. Pendulum	1.000	**0.876**	1.399	1.012
I. Pendulum	1.000	**0.961**	1.743	0.990
Walker	1.000	**0.953**	1.485	1.061

Table D.4: Average trajectory cost on MuJoCo. Numbers are normalized by the cost of the on-policy estimator. ODI and ROS have much larger costs because they both ignore safety constraints. **Our method is the only method consistently achieving both variance reduction and safety constraint satisfaction.**

MuJoCo is a physics engine with various stochastic environments, in which the goal is to control a robot to achieve different behaviors such as walking, jumping, and balancing. We construct 30 policies in each environment, resulting a total of 150 policies. The policies demonstrate a wide range of performance generated by the proximal policy optimization (PPO) algorithm (Schulman et al., 2017) using the default PPO implementation in Huang et al. (2022). Original MuJoCo environments are Markov decision processes (MDP) and do not have cost functions. We enhance it with cost functions to make it constrained Markov decision processes (CMDP). Specifically, the cost of the MuJoCo environments is built on the control cost of the robot. The control cost is the L2 norm of the action and is proposed by OpenAI Gymnasium (Brockman et al., 2016). This control cost is motivated by the fact that large actions in robots induce sudden changes in the robot's state and may cause safety issues.

We set each environment in MuJuCo to have a fixed time horizon 100 in OpenAI Gymnasium (Towers et al., 2024). Because our methods are designed for discrete

action space, we discretize the first dimension of the MuJoCo action space. The remaining dimensions are then controlled by the PPO policies and are deemed as part of the environment. The offline dataset for each environment contains $1,000$ episodes generated by 30 policies with various performances, following the same method as in the Gridworld environments. Functions $q_{\pi,t}, q^c_{\pi,t}$, and $\hat{r}_{\pi,t}$ are learned using the same way as in Gridworld environments. Notably, our algorithm is robust on hyperparameters, as all hyperparameters in Algorithm 4 are tuned offline and are the same across all MuJoCo and Gridworld experiments. Each policy in MuJoCo has 30 independent runs, resulting in a total of $30 \times 30 = 900$ runs. As a result, curves in all figures are averaged from 900 different runs with a wide range of policies, showing a strong statistical significance.

Appendix E

Appendix for Chapter 8

E.1 Proof

E.1.1 Proof of Lemma 16

Lemma 16 (Transition Gradient of the Variance)**.** *For a fixed behavior policy* π_θ*,*

$$\begin{aligned}
&\frac{\partial}{\partial \omega}\mathbb{V}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)] \\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,H)\sum_{t=0}^{T-1}\frac{\partial}{\partial \omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right] \\
&-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)]\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,H)\sum_{t=0}^{T-1}\frac{\partial}{\partial \omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right].
\end{aligned}$$

Proof. To prove Lemma 16, we aim at decomposing the term $\Pr(H=h\mid p_\omega)$ into two parts: one that depends on p_ω and one that does not. Let

$$m_{p_\omega}(h) \doteq \prod_{t=0}^{T-1} p_\omega(S_{t+1}|S_t,A_t) \tag{E.1}$$

and

$$p(h) \doteq \frac{\Pr(H=h\mid p_\omega)}{m_{p_\omega}(h)}, \tag{E.2}$$

then we have

$$\Pr(H=h\mid p_\omega) = p(h)m_{p_\omega}(h). \tag{E.3}$$

Next, we manipulate the term $\frac{\partial}{\partial\theta} m_{p_\omega}(h)$.

$$
\begin{aligned}
\frac{\partial}{\partial\theta} m_{p_\omega}(h) =& \frac{\partial}{\partial\omega} \prod_{t=0}^{T-1} p_\omega(S_{t+1}|S_t, A_t) && \text{(By (5.5))} \\
=& \sum_{t=0}^{T-1} \left(\prod_{i\neq t} p_\omega(S_{i+1}|S_i, A_i) \frac{\partial p_\omega(S_{t+1}|S_t, A_t)}{\partial\omega} \right) \\
=& \sum_{t=0}^{T-1} \left(\frac{\prod_{i=0}^{L-1} p_\omega(S_{i+1}|S_i, A_i)}{p_\omega(S_{t+1}|S_t, A_t)} . \frac{\partial p_\omega(S_{t+1}|S_t, A_t)}{\partial\omega} \right) \\
=& \prod_{i=0}^{L-1} p_\omega(S_{i+1}|S_i, A_i) \cdot \sum_{t=0}^{T-1} \left(\frac{1}{p_\omega(S_{t+1}|S_t, A_t)} \frac{\partial p_\omega(S_{t+1}|S_t, A_t)}{\partial\omega} \right) \\
\overset{\text{(a)}}{=}& \prod_{i=0}^{L-1} p_\omega(S_{i+1}|S_i, A_i) \cdot \sum_{t=0}^{T-1} \left(\frac{1}{p_\omega(S_{t+1}|S_t, A_t)} p_\omega(S_{t+1}|S_t, A_t) \frac{\partial \log p_\omega(S_{t+1}|S_t, A_t)}{\partial\omega} \right) \\
=& \prod_{i=0}^{L-1} p_\omega(S_{i+1}|S_i, A_i) \sum_{t=0}^{T-1} \left(\frac{\partial}{\partial\omega} \log p_\omega(S_{t+1}|S_t, A_t) \right) \\
=& m_{p_\omega}(h) \sum_{t=0}^{T-1} \left(\frac{\partial}{\partial\omega} \log p_\omega(S_{t+1}|S_t, A_t) \right) \\
=& m_{p_\omega}(h) \sum_{t=0}^{T-1} \frac{\partial}{\partial\omega} \log(p_\omega(S_{t+1}|S_t, A_t)) && \text{(E.4)}
\end{aligned}
$$

Here, (a) follows from the fact that

$$
\begin{aligned}
\frac{\partial}{\partial x} \log f(x) &= \frac{1}{f(x)} \frac{\partial f(x)}{\partial x} \\
\implies \frac{\partial f(x)}{\partial x} &= f(x) \cdot \frac{\partial \log f(x)}{\partial x}.
\end{aligned}
$$

Then, we decompose the variance objective

$$
\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)]\\
=&\frac{\partial}{\partial\omega}\left(\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}^2(\pi_e,\pi_\theta,h)]-\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,h)]^2\right)\\
=&\frac{\partial}{\partial\omega}\sum_h \Pr(H=h|p_\omega)\mathrm{OPE}^2(\pi_e,\pi_\theta,h)\\
&-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,h)]\frac{\partial}{\partial\omega}\sum_h \Pr(H=h|p_\omega)\mathrm{OPE}(\pi_e,\pi_\theta,h)\\
=&\sum_h p(h)\mathrm{OPE}^2(\pi_e,\pi_\theta,h)\frac{\partial}{\partial\omega}m_{p_\omega}(h)\\
&-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,h)]\sum_h p(h)\mathrm{OPE}(\pi_e,\pi_\theta,h)\frac{\partial}{\partial\omega}m_{p_\omega}(h) \qquad \text{(By (E.3))}\\
=&\sum_h p(h)\mathrm{OPE}^2(\pi_e,\pi_\theta,h)m_{p_\omega}(h)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\\
&-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,h)]\sum_h p(h)\mathrm{OPE}(\pi_e,\pi_\theta,h)m_{p_\omega}(h)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By (E.4))}\\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}^2(\pi_\theta,H)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))]\\
&-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,h)]\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,H)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right].
\end{aligned}
$$

□

E.1.2 Proof of Lemma 17

Lemma 17 (Transition Gradient Convergence)**.** *For a fixed behavior policy π_θ, Algorithm 5 converges. That is, $\mathbb{V}_{H_i\sim p_{\omega_i},\pi_\theta}[\mathrm{IS}(\pi_e,\pi_\theta,H_i)]$ convereges to a finite value and $\lim_{i\to\infty}\frac{\partial}{\partial\omega}\mathbb{V}_{H_i\sim p_{\omega_i},\pi_\theta}[\mathrm{IS}(\pi_e,\pi_\theta,H_i)]=0$.*

Proof. The proof leverages Proposition 3 in Bertsekas and Tsitsiklis (2000), for which we have to show that Algorithm 5 satisfies the following conditions:

1. $\mathbb{V}[\mathrm{IS}(\pi_\theta,p_{\omega_i},H_i)]$ is continuously differentiable w.r.t. ω.
2. The gradient of the variance objectives, $\frac{\partial}{\partial\omega}\mathbb{V}[\mathrm{IS}(\pi_\theta,p_{\omega_i},H_i)]$, is Lipschitz continuous w.r.t. ω.

3. The variance of the gradient estimate used by Algorithm 5 is bounded.

The other conditions of Proposition 3 in Bertsekas and Tsitsiklis (2000) are satisfied because of the unbiasedness of the gradient estimates in Algorithm 5. Additionally, since the gradient objective, as a variance, is bounded below by zero, we can avoid the case of converging to $-\infty$ according to Proposition 3 (Bertsekas and Tsitsiklis, 2000).

For the continuous differentiability, we have p_ω is continuously differentiable because it is obtained through a soft-max layer. In addition, since π_θ is attained through the neural network with soft-max layer, it is always non-zero. Thus, the quotient $\frac{w_{\pi_e}}{w_{\pi_\theta}}$ always exists and so is the estimator $\text{IS}(\pi_\theta, p_\omega, H)$. Therefore, by the gradient expression in Lemma 16, we conclude that $\frac{\partial}{\partial\omega}V_{H\sim p_\omega,\pi_\theta}[\text{IS}(\pi_\theta, p_\omega, H)]$ is continuously differentiable, verifying condition 1.

Next, we show the Lipschitz continuity of $\frac{\partial}{\partial\omega}V_{H\sim p_\omega,\pi_\theta}[\text{IS}(\pi_\theta, p_\omega, H)]$ by verifying the boundedness of its second derivative.

$$
\begin{aligned}
&\frac{\partial^2}{\partial^2\omega}\mathbb{V}_{H\sim p_\omega,\pi_\theta}[\text{IS}(\pi_\theta, p_\omega, H)]\\
=&\frac{\partial}{\partial\omega}\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\text{IS}^2(\pi_e,\pi_\theta,p_\omega,H)\textstyle\sum_{t=0}^{T-1}\log(p_\omega(S_{t+1}|S_t,A_t))\right]\\
&-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\text{IS}(\pi_e,\pi_\theta,p_\omega,H)]\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\text{IS}(\pi_e,\pi_\theta,p_\omega,H)\textstyle\sum_{t=0}^{T-1}\log(p_\omega(S_{t+1}|S_t,A_t))\right]\\
=&\tfrac{\partial}{\partial\omega}\Big(\textstyle\sum_h\Big(p(h)m_{p_\omega}(h)\text{IS}^2(\pi_e,\pi_\theta,p_\omega,H)\sum_{t=0}^{T-1}\tfrac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\Big)\\
&-2\textstyle\sum_h(p(h)m_{p_\omega}(h)\text{IS}(\pi_e,\pi_\theta,p_\omega,H))\\
&\cdot\textstyle\sum_h\Big(p(h)m_{p_\omega}(h)\text{IS}(\pi_e,\pi_\theta,p_\omega,H)\sum_{t=0}^{T-1}\tfrac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\Big)\Big)\\
&\hfill\text{(By Lemma 16 and (E.3))}\\
=&\tfrac{\partial}{\partial\omega}\left(\textstyle\sum_h\left(p(h)m_{p_\omega}(h)\text{IS}^2(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\log m_{p_\omega}(h)\right)\right.\\
&\left.-2\textstyle\sum_h(p(h)m_{p_\omega}(h)\text{IS}(\pi_e,\pi_\theta,p_\omega,H))\cdot\sum_h\left(p(h)m_{p_\omega}(h)\text{IS}(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\log m_{p_\omega}(h)\right)\right)\\
=&\tfrac{\partial}{\partial\omega}\Big(\textstyle\sum_h\Big(p(h)m_{p_\omega}(h)\text{IS}^2(\pi_e,\pi_\theta,p_\omega,H)\tfrac{1}{m_{p_\omega}(h)}\tfrac{\partial}{\partial\omega}m_{p_\omega}(h)\Big)\\
&-2\textstyle\sum_h(p(h)m_{p_\omega}(h)\text{IS}(\pi_e,\pi_\theta,p_\omega,H))\cdot\sum_h\Big(p(h)m_{p_\omega}(h)\text{IS}(\pi_e,\pi_\theta,p_\omega,H)\tfrac{1}{m_{p_\omega}(h)}\tfrac{\partial}{\partial\omega}m_{p_\omega}(h)\Big)\Big)\\
=&\tfrac{\partial}{\partial\omega}\left(\textstyle\sum_h\left(p(h)\text{IS}^2(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}m_{p_\omega}(h)\right)\right.\\
&\left.-2\textstyle\sum_h(p(h)m_{p_\omega}(h)\text{IS}(\pi_e,\pi_\theta,p_\omega,H))\cdot\sum_h\left(p(h)\text{IS}(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}m_{p_\omega}(h)\right)\right)
\end{aligned}
$$

$$= \sum_h p(h) \left(\underbrace{\mathrm{IS}^2(\pi_e, \pi_\theta, p_\omega, H)}_{(1)} \underbrace{\frac{\partial^2}{\partial^2 \omega} m_{p_\omega}(h)}_{(2)} \right)$$
$$- 2\frac{\partial}{\partial \omega} \left[\sum_h \left(p(h) m_{p_\omega}(h) \mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \right) \cdot \sum_h \left(p(h) \mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial}{\partial \omega} m_{p_\omega}(h) \right) \right].$$

We further decompose the term in the square brackets.

$$\begin{aligned}
&\frac{\partial}{\partial \omega} \left[\sum_h \left(p(h) m_{p_\omega}(h) \mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \right) \cdot \sum_h \left(p(h) \mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial}{\partial \omega} m_{p_\omega}(h) \right) \right] \\
=& \sum_h p(h) \frac{\partial}{\partial \omega} \left(m_{p_\omega}(h) \mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \right) \cdot \sum_h \left(p(h) \mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial}{\partial \omega} m_{p_\omega}(h) \right) \\
&+ \sum_h \left(p(h) m_{p_\omega}(h) \mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \right) \cdot \sum_h p(h) \frac{\partial}{\partial \omega} \left(\mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial}{\partial \omega} m_{p_\omega}(h) \right) \\
=& \sum_h p(h) \left(\underbrace{\mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H)}_{(3)} \underbrace{\frac{\partial}{\partial \omega} m_{p_\omega}(h)}_{(4)} \right) \cdot \sum_h p(h) \left(\mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial}{\partial \omega} m_{p_\omega}(h) \right) \\
&+ \sum_h p(h) \left(\underbrace{m_{p_\omega}(h)}_{(5)} \mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \right) \cdot \sum_h p(h) \left(\mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial^2}{\partial^2 \omega} m_{p_\omega}(h) \right).
\end{aligned}$$

Notice that since $p(h)$ is defined as $\frac{Pr(H=h|\pi)}{w_\pi(h)}$ in (E.2), where $\Pr(H = h)$ is trivially bounded and $w_\pi(h)$ is always positive. Thus, $p(h)$ is bounded. We then analyze the boundedness of $\frac{\partial^2}{\partial^2 \omega} V_{H \sim p_\omega, \pi_\theta}[\mathrm{IS}(\pi_\theta, p_\omega, H)]$ through the above 5 terms.

For (1) and (3), the quotient $\frac{\pi_e(a|s)}{\pi_\theta(a|s)}$ is bounded above by assumption. Besides, since the reward is bounded, so is $g(h)$. Therefore, both (1), $\mathrm{IS}^2(\pi_e, \pi_\theta, p_\omega, H)$ and (3) $\mathrm{IS}(\pi_e, \pi_\theta, p_\omega, H)$ are bounded.

For (5), it is bounded because $m_{p_\omega}(h) = \prod_{t=0}^{T-1} p_\omega(S_{t+1}|S_t, A_t) \leq 1$. Then, for (4),

$$\begin{aligned}
\frac{\partial}{\partial \omega} m_{p_\omega}(h) =& \frac{\partial}{\partial \omega} \prod_{t=0}^{T-1} p_\omega(S_{t+1}|S_t, A_t) \\
=& \sum_{t=0}^{T-1} \frac{\partial}{\partial \omega} p_\omega(S_{t+1}|S_t, A_t) \frac{\prod_{i=0}^{L-1} p_\omega(S_{i+1}|S_i, A_i)}{p_\omega(S_{t+1}|S_t, A_t)}.
\end{aligned}$$

Here, $\frac{\partial}{\partial\omega}p_\omega(S_{t+1}|S_t,A_t)$ is bounded by construction and $\frac{\prod_{i=0}^{L-1}p_\omega(S_{i+1}|S_i,A_i)}{p_\omega(S_{t+1}|S_t,A_t)} \leq 1$. Thus, (4) is bounded. Lastly, for (2)

$$
\begin{aligned}
&\frac{\partial^2}{\partial^2\omega}m_{p_\omega}(h)\\
=&\frac{\partial}{\partial\omega}\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}p_\omega(S_{t+1}|S_t,A_t)\frac{\prod_{i=0}^{L-1}p_\omega(S_{i+1}|S_i,A_i)}{p_\omega(S_{t+1}|S_t,A_t)}\\
=&\frac{\partial}{\partial\omega}\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}p_\omega(S_{t+1}|S_t,A_t)\prod_{i\neq t}p_\omega(S_{i+1}|S_i,A_i)\\
=&\sum_{t=0}^{T-1}\frac{\partial^2}{\partial^2\omega}p_\omega(S_{t+1}|S_t,A_t)\prod_{i\neq t}p_\omega(S_{i+1}|S_i,A_i)+\frac{\partial}{\partial\omega}p_\omega(S_{t+1}|S_t,A_t)\\
&\cdot\sum_{i\neq t}\frac{\partial}{\partial\omega}p_\omega(S_{i+1}|S_i,A_i)\prod_{j\neq t,i}p_\omega(S_{j+1}|S_j,A_j),
\end{aligned}
$$

which is bounded because p_ω is constructed to be twice differentiable with bounded first and second derivatives.

Therefore, we conclude that the gradient objective $\frac{\partial}{\partial\omega}V_{H\sim p_\omega,\pi_\theta}[\mathrm{IS}(\pi_\theta,p_\omega,H)]$ is Lipschitz continuous w.r.t. ω, verifying condition 1.

Finally, we show that the variance of the gradient estimate used by Algorithm 5 is bounded. According to Algorithm 5, we use the unbiased estimate as

$$
\begin{aligned}
\frac{\partial}{\partial\omega}V_{H\sim p_\omega,\pi_\theta}[\mathrm{IS}(\pi_\theta,p_\omega,H)] \approx &\underbrace{\mathrm{IS}^2(\pi_\theta,p_\omega,H)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))}_{A}\\
&\underbrace{-2\mathrm{IS}(\pi_\theta,p_\omega,H)\mathrm{IS}(\pi_\theta,p_\omega,H)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))}_{B}.
\end{aligned}
$$

Then, the variance of the estimate is decomposed into

$$\mathbb{V}[A]+\mathbb{V}[B]+2\mathrm{Cov}[A,B],$$

where $\mathrm{Cov}[A,B] \leq \sqrt{\mathbb{V}[A]}\cdot\sqrt{\mathbb{V}[B]}$ by the Cauchy-Schwarz inequality. Thus, it is sufficient to show the boundedness of $\mathbb{V}[A]$ and $\mathbb{V}[B]$. For $\mathbb{V}[A]$, since the variance of a bounded random variable is bounded, we aim to demonstrate that for any trajectory

h, the term $\mathrm{IS}^2(\pi_\theta, p_\omega, h) \sum_{t=0}^{T-1} \log(p_\omega(S_{t+1}|S_t, A_t))$ is bounded.

$$\begin{aligned}
&\mathrm{IS}^2(\pi_\theta, p_\omega, h) \sum_{t=0}^{T-1} \frac{\partial}{\partial \omega} \log(p_\omega(S_{t+1}|S_t, A_t)) \\
=&\mathrm{IS}^2(\pi_\theta, p_\omega, h) \sum_{t=0}^{T-1} \frac{\partial}{\partial \omega} \log(p_\omega(S_{t+1}|S_t, A_t)) \\
=&\mathrm{IS}^2(\pi_\theta, p_\omega, h) \frac{\partial}{\partial \omega} \log m_{p_\omega}(h) \\
=&\mathrm{IS}^2(\pi_\theta, p_\omega, h) \frac{\frac{\partial}{\partial \omega} m_{p_\omega}(h)}{m_{p_\omega}(h)}.
\end{aligned} \tag{E.5}$$

The boundedness of $\mathrm{IS}^2(\pi_\theta, p_\omega, h)$ and $\frac{\partial}{\partial \omega} m_{p_\omega}(h)$ is shown by the argument above for term (3) and (4). And the boundedness of $\frac{1}{m_{p_\omega}(h)} = \frac{1}{\prod_{t=0}^{T-1} p_\omega(S_{t+1}|S_t, A_t)}$ comes from the fact that p_ω is always nonzero by construction. Thus, we conclude that $\mathbb{V}[A]$ is bounded.

Next, we decompose term B into two parts because of the different samples used to estimate them:

$$\underbrace{\mathrm{IS}(\pi_\theta, p_\omega, H)}_{C} \underbrace{\mathrm{IS}(\pi_\theta, p_\omega, H) \sum_{t=0}^{T-1} \frac{\partial}{\partial \omega} \log(p_\omega(S_{t+1}|S_t, A_t))}_{D}.$$

We then have

$$\mathbb{V}[B] = \mathbb{V}[CD] = \mathbb{E}[C^2]\mathbb{V}[D] + \mathbb{E}[D^2]\mathbb{V}[C].$$

We show their boundedness term by term.

$$\mathbb{E}[C^2] = \mathbb{E}_{H \sim p_\omega, \pi_\theta}[\mathrm{IS}^2(\pi_\theta, p_\omega, H)] = \sum_h p(h) m_{p_\omega}(h) \mathrm{IS}^2(\pi_e, \pi_\theta, p_\omega, H), \tag{E.6}$$

where each term is shown to be bounded above. Next, by the derivation from (E.5),

$$\mathbb{E}[D^2] = \sum_h p(h) m_{p_\omega}(h) \mathrm{IS}^2(\pi_e, \pi_\theta, p_\omega, H) \left(\frac{\frac{\partial}{\partial \omega} m_{p_\omega}(h)}{m_{p_\omega}(h)} \right)^2,$$

where the boundedness follows from the analysis of (E.6) and (E.5).

As for the two variance terms, $\mathbb{V}[C]$ and $\mathbb{V}[D]$, we show the boundedness of the random variable C and D for each trajectory h, where $\mathrm{IS}(\pi_\theta, p_\omega, H)$ is shown to be bounded in term (3) above, and the boundedness of $\sum_{t=0}^{T-1} \frac{\partial}{\partial \omega} \log(p_\omega(S_{t+1}|S_t, A_t))$ is incorporated in (E.5).

Therefore, we conclude that the variance of our estimate is bounded. By far, we show that the three conditions of Proposition 3 in Bertsekas and Tsitsiklis (2000) are satisfied, demonstrating the convergence of Algorithm 5.

□

E.1.3 Proof of Lemma 18

Proof.

Lemma 18 (Transition Gradient of Variance with KL)**.** *For a fixed behavior policy π_θ and a regularization coefficient $\eta > 0$,*

$$\begin{aligned}
&\frac{\partial}{\partial \omega}\mathbb{V}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)] - D_{\mathrm{KL}}(\Pr(H|p_\omega)\|\Pr(H|p_{\omega_0}))\\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right]\\
&-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)]\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right]\\
&-\eta\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\left(\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right)\left(1+\ell_{p_\omega}-\ell_{p_{\omega_0}}\right)\right].
\end{aligned}$$

We begin by manipulating the KL-divergence term.

$$\begin{aligned}
D_{\mathrm{KL}}(\Pr(H|p_\omega)\|\Pr(H|p_{\omega_0})) =&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\log\frac{\Pr(H|p_\omega)}{\Pr(H|p_{\omega_0})}\right]\\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\log\frac{m_{p_\omega}(H)}{m_{p_{\omega_0}}(H)}\right] && \text{(By (E.3))}\\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\log m_{p_\omega}(H)-\log m_{p_{\omega_0}}(H)\right].
\end{aligned}$$

Next, we decompose the following gradient:

$$\begin{aligned}
&\frac{\partial}{\partial\omega}\log m_{p_\omega}(H) && \text{(E.7)}\\
=&\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log p_\omega(S_{t+1}|S_t,A_t)\\
=&\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t)). && \text{(By definition)}
\end{aligned}$$

Then, we take the gradient of the KL-divergence with respect to ω:

$$
\begin{aligned}
&\frac{\partial}{\partial\omega} D_{\mathrm{KL}}(\Pr(H|p_\omega) \| \Pr(H|p_{\omega_0})) && \text{(E.8)}\\
=&\frac{\partial}{\partial\omega} \mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\log m_{p_\omega}(H) - \log m_{p_{\omega_0}}(H)\right]\\
=&\frac{\partial}{\partial\omega} \sum_h \Pr(H=h|p_\omega)\left[\log m_{p_\omega}(h) - \log m_{p_{\omega_0}}(h)\right]\\
=&\frac{\partial}{\partial\omega} \sum_h p(h) m_{p_\omega}(h)\left[\log m_{p_\omega}(h) - \log m_{p_{\omega_0}}(h)\right] && \text{(By (E.3))}\\
=&\sum_h p(h)\left[\frac{\partial}{\partial\omega} m_{p_\omega}(h) \log m_{p_\omega}(h) - \log m_{p_{\omega_0}}(h)\frac{\partial}{\partial\omega} m_{p_\omega}(h)\right]\\
=&\sum_h p(h)\left[\log m_{p_\omega}(h)\frac{\partial}{\partial\omega} m_{p_\omega}(h) + m_{p_\omega}(h)\frac{\partial}{\partial\omega}\log m_{p_\omega}(h)\right.\\
&\left. - \log m_{p_{\omega_0}}(h) m_{p_\omega}(h) \sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right] && \text{(By (E.4))}\\
=&\sum_h p(h)\left[\log m_{p_\omega}(h) m_{p_\omega}(h)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t)) + m_{p_\omega}(h)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right.\\
&\left. - \log m_{p_{\omega_0}}(h) m_{p_\omega}(h) \sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right] && \text{(By (E.4) (E.7))}\\
=&\sum_h p(h) m_{p_\omega}(h)\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\left[\log m_{p_\omega}(h) + 1 - \log m_{p_{\omega_0}}(h)\right]\\
=&\sum_h \Pr(H=h|p_\omega)\sum_{t=0}^{T-1}\left[\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right]\left[\log m_{p_\omega}(h) + 1 - \log m_{p_{\omega_0}}(h)\right]\\
& && \text{(By (E.3))}\\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\left(\frac{\partial}{\partial\omega}\ell_{p_\omega}\right)\left(1 + \ell_{p_\omega} - \ell_{p_{\omega_0}}\right)\right]. && \text{(By (E.1))}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,H)\right]-D_{\mathrm{KL}}(\Pr(H|p_\omega)\|\Pr(H|p_{\omega_0}))\\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,H)\textstyle\sum_{t=0}^{T-1}\log(p_\omega(S_{t+1}|S_t,A_t))\right]-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)]\\
&\cdot\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,H)\textstyle\sum_{t=0}^{T-1}\log(p_\omega(S_{t+1}|S_t,A_t))\right]-\eta\frac{\partial}{\partial\omega}D_{\mathrm{KL}}(\Pr(H|p_\omega)\|\Pr(H|p_{\omega_0}))\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(By Lemma 16)}\\
=&\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right]-2\mathbb{E}_{H\sim p_\omega,\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,H)]\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right]\\
&-\mathbb{E}_{H\sim p_\omega,\pi_\theta}\left[\left(\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right)\left(1+\ell_{p_\omega}-\ell_{p_{\omega_0}}\right)\right].\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(By (E.8))}
\end{aligned}
$$

□

E.1.4 Proof of Lemma 19

Lemma 19 (Off-Transition Gradient of Variance)**.** *When $p_\omega \neq p_{\omega'}$, for a fixed behavior policy π_θ,*

$$
\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)]\\
=&2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right]\\
&-2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)]\cdot\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right].
\end{aligned}
$$

Proof. For simplification, we define $w_\pi(h)\doteq\prod_{t=0}^{T-1}\pi(A_t|S_t)$ under trajectory h. Then,

$$
\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)]\\
=&\frac{\partial}{\partial\omega}\left(\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}^2(\pi_e,\pi_\theta,p_\omega,H)]-\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)]^2\right)\\
=&\frac{\partial}{\partial\omega}\left(\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\frac{m^2_{p_\omega}(H)}{m^2_{p_{\omega'}}(H)}\mathrm{OPE}^2(\pi_e,\pi_\theta,H)\right]-\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\frac{m_{p_\omega}(H)}{m_{p_{\omega'}}(H)}\mathrm{OPE}(\pi_e,\pi_\theta,H)\right]^2\right)\\
=&\frac{\partial}{\partial\omega}\sum_h\left(\Pr(H=h|p_{\omega'})\frac{m^2_{p_\omega}(H)}{m^2_{p_{\omega'}}(H)}\mathrm{OPE}^2(\pi_e,\pi_\theta,H)\right)-2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)]\\
&\frac{\partial}{\partial\omega}\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\frac{m_{p_\omega}(H)}{m_{p_{\omega'}}(H)}\mathrm{OPE}(\pi_e,\pi_\theta,H)\right]
\end{aligned}
$$

$$
\begin{aligned}
=&\sum_h \left(p(h) \frac{1}{m_{p_{\omega'}}(H)} \mathrm{OPE}^2(\pi_e, \pi_\theta, H) \frac{\partial}{\partial \omega} m^2_{p_\omega}(h) \right) \\
&- 2\mathbb{E}_{H\sim p_{\omega'}, \pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)] \frac{\partial}{\partial \omega} \sum_h \left(p(h) m_{p_{\omega'}(h)} \frac{m_{p_\omega}(h)}{m_{p_{\omega'}}(h)} \mathrm{OPE}(\pi_e, \pi_\theta, H) \right) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By (E.3))} \\
=&2\sum_h \left(p(h) \frac{m_{p_\omega}(h)}{m_{p_{\omega'}}(h)} \mathrm{OPE}^2(\pi_e, \pi_\theta, H) \frac{\partial}{\partial \omega} m_{p_\omega}(h) \right) \\
&- 2\mathbb{E}_{H\sim p_{\omega'}, \pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)] \sum_h \left(p(h) \mathrm{OPE}(\pi_e, \pi_\theta, H) \frac{\partial}{\partial \omega} m_{p_\omega}(h) \right) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By (E.3))} \\
=&2\sum_h \left(p(h) \mathrm{OPE}^2(\pi_e, \pi_\theta, H) \frac{m_{p_\omega}(h)}{m_{p_{\omega'}}(h)} m_{p_\omega}(h) \frac{\partial}{\partial \omega} \ell_{p_\omega} \right) \\
&- 2\mathbb{E}_{H\sim p_{\omega'}, \pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)] \sum_h \left(p(h) \mathrm{OPE}(\pi_e, \pi_\theta, H) m_{p_\omega}(h) \frac{\partial}{\partial \omega} \ell_{p_\omega} \right) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By (E.4))} \\
=&2\sum_h \left(p(h) m_{p_{\omega'}}(h) \frac{m^2_{p_\omega}(h)}{m^2_{p_{\omega'}}(h)} \mathrm{OPE}^2(\pi_e, \pi_\theta, H) \frac{\partial}{\partial \omega} \ell_{p_\omega} \right) \\
&- 2\mathbb{E}_{H\sim p_{\omega'}, \pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)] \\
&\cdot \sum_h \left(p(h) m_{p_{\omega'}}(h) \frac{m_{p_\omega}(h)}{m_{p_{\omega'}}(h)} \mathrm{OPE}(\pi_e, \pi_\theta, H) \sum_{t=0}^{T-1} \log(p_\omega(S_{t+1}|S_t, A_t)) \right) \qquad \text{(By (E.4))} \\
=&2\sum_h \left(\Pr(H = h|p_{\omega'}) \mathrm{OPE}^2(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial}{\partial \omega} \ell_{p_\omega} \right) \\
&- 2\mathbb{E}_{H\sim p_{\omega'}, \pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)] \\
&\cdot \sum_h \left(\Pr(H = h|p'_\omega) \mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H) \sum_{t=0}^{T-1} \frac{\partial}{\partial \omega} \log(p_\omega(S_{t+1}|S_t, A_t)) \right) \qquad \text{(By (E.3))} \\
=&2\mathbb{E}_{H\sim p_{\omega'}, \pi_\theta} \left[\mathrm{OPE}^2(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial}{\partial \omega} \ell_{p_\omega} \right] \\
&- 2\mathbb{E}_{H\sim p_{\omega'}, \pi_\theta}[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H)] \mathbb{E}_{H\sim p_{\omega'}, \pi_\theta} \left[\mathrm{OPE}(\pi_e, \pi_\theta, p_\omega, H) \frac{\partial}{\partial \omega} \ell_{p_\omega} \right].
\end{aligned}
$$

□

E.1.5 Proof of Lemma 20

Lemma 20 (Off-transition Gradient of Variance with KL). *For a fixed behavior policy π_θ and a regularization coefficient $\eta > 0$,*

$$\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)] - \eta D_{\mathrm{KL}}(\Pr(H|p_{\omega'})\|\Pr(H|p_\omega))\\
=&2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right]\\
&-2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)]\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right]\\
&-\eta\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[-\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right].
\end{aligned}$$

The KL-divergence between two probability distribution p and q is defined as $D_{\mathrm{KL}}(p\|q) \doteq \mathbb{E}_{X\sim p}\left[\log\frac{p(X)}{q(X)}\right]$. Therefore, the KL-divergence between the trajectory distribution of the target transition p_ω and the simulator's transition $p_{\omega'}$ is given by

$$\begin{aligned}
D_{\mathrm{KL}}(\Pr(H|p_{\omega'})\|\Pr(H|p_\omega)) =&\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\log\frac{\Pr(H|p_{\omega'})}{\Pr(H|p_\omega)}\right]\\
=&\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\log\frac{m_{p_{\omega'}}(H)}{m_{p_\omega}(H)}\right] \qquad \text{(By (E.3))}\\
=&\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\log m_{p_{\omega'}}(H) - \log m_{p_\omega}(H)\right].
\end{aligned}$$

We take the gradient of the KL-divergence with respect to ω:

$$\begin{aligned}
\frac{\partial}{\partial\omega}D_{\mathrm{KL}}(\Pr(H|p_{\omega'})\|\Pr(H|p_\omega)) =&\frac{\partial}{\partial\omega}\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\log m_{p_{\omega'}}(H) - \log m_{p_\omega}(H)\right]\\
=&\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[-\frac{\partial}{\partial\omega}\log m_{p_\omega}(H)\right]\\
=&\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[-\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log p_\omega(S_{t+1}|S_t,A_t)\right]\\
=&\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[-\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right]. \qquad \text{(E.9)}
\end{aligned}$$

Thus,

$$
\begin{aligned}
&\frac{\partial}{\partial\omega}\mathbb{V}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)] - \eta D_{\mathrm{KL}}(\Pr(H|p_{\omega'})\|\Pr(H|p_\omega)) \\
=&2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right] \\
&- 2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)]\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right] \\
&- \tfrac{\partial}{\partial\omega}\eta D_{\mathrm{KL}}(\Pr(H|p_{\omega'})\|\Pr(H|p_\omega)) && \text{(By Lemma 20)} \\
=&2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}^2(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right] \\
&- 2\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)]\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[\mathrm{OPE}(\pi_e,\pi_\theta,p_\omega,H)\tfrac{\partial}{\partial\omega}\ell_{p_\omega}\right] \\
&- \eta\mathbb{E}_{H\sim p_{\omega'},\pi_\theta}\left[-\textstyle\sum_{t=0}^{T-1}\frac{\partial}{\partial\omega}\log(p_\omega(S_{t+1}|S_t,A_t))\right]. && \text{(By (E.9))}
\end{aligned}
$$

Appendix F

Appendix for Chapter 9

F.1 Mathematical Background

Theorem F.1.1 (**Gronwall Inequality**). *(Lemma 6 in Section 11.2 in Borkar (2009)) For a continuous function $u(\cdot) \geq 0$ and scalars $C, K, T \geq 0$,*

$$u(t) \leq C + K \int_0^t u(s)ds \quad \forall t \in [0, T]$$

implies

$$u(t) \leq Ce^{tK}, \forall t \in [0, T].$$

Theorem F.1.2 (**Gronwall Inequality in the Reverse Time**). *For a continuous function $u(\cdot) \geq 0$ and scalars $C, K, T \geq 0$,*

$$u(t) \leq C + K \int_t^0 u(s)ds \quad \forall t \in [-T, 0] \tag{F.1}$$

implies

$$u(t) \leq Ce^{-tK}, \forall t \in [-T, 0].$$

Proof. $\forall s \in [0, T]$, define

$$v(s) \doteq e^{sK} K \int_s^0 u(r)dr. \tag{F.2}$$

Taking the derivative of $v(s)$,

$$\begin{aligned} v'(s) &= -e^{sK} K u(s) + e^{sK} K^2 \int_s^0 u(r)dr \\ &= e^{sK} K \left[-u(s) + K \int_s^0 u(r)dr \right] && \text{(by (F.1))} \\ &\geq -Ce^{sK} K. \end{aligned}$$

Thus,

$$v(t) = v(0) - \int_t^0 v'(s)ds \leq v(0) + \int_t^0 Ce^{sK}Kds = KC\int_t^0 e^{sK}ds.$$

By (F.2),

$$\begin{aligned} K\int_t^0 u(s)ds &= v(t)e^{-tK} \\ &\leq KC\int_t^0 e^{sK}dse^{-tK} \\ &\leq KC\int_t^0 e^{(s-t)K}ds \\ &= KC[\frac{1}{k}e^{(0-t)K} - \frac{1}{k}e^{(t-t)K}] \\ &= -C + Ce^{-tK}. \end{aligned}$$

Thus,

$$u(t) \leq C + K\int_t^0 u(s)ds \leq Ce^{-tK}.$$

□

Theorem F.1.3 (**Discrete Gronwall Inequality**). *(Lemma 8 in Section 11.2 in Borkar (2009)) For non-negative sequences* $\{x_n, n \geq 0\}$ *and* $\{a_n, n \geq 0\}$ *and scalars* $C, L \geq 0$,

$$x_{n+1} \leq C + L\sum_{i=0}^{n} a_i x_i \quad \forall n$$

implies

$$x_{n+1} \leq Ce^{L\sum_{i=0}^{n} a_i} \quad \forall n.$$

Theorem F.1.4 (The Arzela-Ascoli Theorem in the Extended Sense on $[0, T)$). *Let* $\{t \in [0, T) \mapsto g_n(t)\}$ *be equicontinuous in the extended sense. Then, there exists a subsequence* $\{g_{n_k}(t)\}$ *that converges to some continuous limit* $g^{\lim}(t)$, *uniformly in* t *on* $[0, T)$.

The proof of the Arzela-Ascoli Theorem can be found in any standard analysis textbook, see, e.g., Royden and Fitzpatrick (1968); Dunford and Schwartz (1988). The proof of the Arzela-Ascoli Theorem in the extended sense is virtually the same. The difference is that in the standard Arzela-Ascoli Theorem, one uses the compactness to find a finite subcover. But in the extended one, $[0, T)$ is not compact. However, finding a finite cover for this specific set $[0, T)$ is indeed trivial. We anyway still include the full proof below for completeness.

Proof. Fix an arbitrary $\epsilon > 0$, by Definition 4, $\exists \delta > 0$ such that

$$\limsup_n \sup_{0 \le |t_1 - t_2| \le \delta,\, 0 \le t_1 \le t_2 < T} \|g_n(t_1) - g_n(t_2)\| \le \epsilon. \tag{F.3}$$

This means by the definition of equicontinuity in the extended sense, when n is large enough, for any $0 \le |t_1 - t_2| \le \delta$, the function values $g_n(t_1)$ and $g_n(t_2)$ are also close. To conveniently utilize this property, we divide $[0, T)$ into a set of disjoint intervals and each interval has a length δ such that the t in each interval is close. In particular, we define

$$\begin{aligned} N &\doteq \max\{i \mid i\delta < T,\, i \in \mathbb{Z}\}, \\ I_i &\doteq [i\delta, (i+1)\delta), \quad i = 0, 1, \ldots, N. \end{aligned}$$

The set of intervals $\{I_i\}_{i=0}^N$ covers the domain $[0, T)$,

$$[0, T) \subseteq \bigcup_{i=0}^{N} I_i.$$

We now show $g_n(t)$ is uniformly bounded uniformly on the set of dividing points $\{i\delta\}_{i=0}^N$. In particular, we have for any $i \in \{0, 1, \ldots, N\}$,

$$\begin{aligned} &\limsup_n \|g_n(i\delta)\| \\ \le &\limsup_n \|g_n(i\delta) - g_n((i-1)\delta)\| \\ &+ \limsup_n \|g_n((i-1)\delta) - g_n((i-2)\delta)\| \\ &+ \ldots \\ &+ \limsup_n \|g_n(\delta) - g_n(0)\| \\ &+ \limsup_n \|g_n(0)\| \\ \le &(N+1)\epsilon + \limsup_n \|g_n(0)\| && \text{(by (F.3))} \\ \le &(N+1)\epsilon + \sup_n \|g_n(0)\| \\ < &\infty. && (\sup_n \|g_n(0)\| < \infty \text{ in Definition 4}) \end{aligned}$$

This implies

$$\sup_{i \in \{0,1,\ldots,N\}, n \ge 0} \|g_n(i\delta)\| < \infty.$$

By the Bolzano-Weierstrass theorem, there exists a subsequence of functions $\{g_{n_{0,k}}\}$ in $\{g_n\}$ such that $\{g_{n_{0,k}}(0 \cdot \delta)\}$ converges. Repeating the same argument for the

sequence of points $\left\{g_{n_{0,k}}(1\cdot\delta)\right\}$, there exists a subsequence $\left\{g_{n_{1,k}}\right\}$ of $\left\{g_{n_{0,k}}\right\}$ such that $\left\{g_{n_{1,k}}(1\cdot\delta)\right\}$ converges. Repeating this process, because N is finite, there exists a subsequence $\{g_{n_k}\}$ that converges at all dividing points $t\in\{i\delta\}_{i=0}^{N}$. Due to the finiteness of N, $\exists k_0$, such that $\forall i\in\{0,1,\ldots,N\}$, $\forall k_1\geq k_0, \forall k_2\geq k_0$, we have

$$\left\|g_{n_{k_1}}(i\delta)-g_{n_{k_2}}(i\delta)\right\|\leq\epsilon. \tag{F.4}$$

By (F.3), $\exists k_1$ such that $\forall k\geq k_1$,

$$\sup_{0\leq|t_1-t_2|\leq\delta,\,0\leq t_1\leq t_2<T}\|g_{n_k}(t_1)-g_{n_k}(t_2)\|\leq 2\epsilon. \tag{F.5}$$

Thus, $\forall t\in[0,T)$, $\forall k\geq\max\{k_0,k_1\}, \forall k'\geq\max\{k_0,k_1\}$,

$$\begin{aligned}
&\left\|g_{n_k}(t)-g_{n_{k'}}(t)\right\|\\
\leq&\|g_{n_k}(t)-g_{n_k}(\lfloor t/\delta\rfloor\cdot\delta)\|+\left\|g_{n_k}(\lfloor t/\delta\rfloor\cdot\delta)-g_{n_{k'}}(\lfloor t/\delta\rfloor\cdot\delta)\right\|\\
&+\left\|g_{n_{k'}}(\lfloor t/\delta\rfloor\cdot\delta)-g_{n_{k'}}(t)\right\|\\
\leq&2\epsilon+\left\|g_{n_k}(\lfloor t/\delta\rfloor\cdot\delta)-g_{n_{k'}}(\lfloor t/\delta\rfloor\cdot\delta)\right\|+2\epsilon && \text{(by (F.5))}\\
\leq&2\epsilon+\epsilon+2\epsilon && \text{(by (F.4))}\\
=&5\epsilon.
\end{aligned}$$

This shows that the sequence $\{g_{n_k}\}$ is uniformly Cauchy and therefore uniformly converges to a continuous function.

□

Theorem F.1.5 (Moore-Osgood Theorem for Interchanging Limits)**.** *If* $\lim_{n\to\infty}a_{n,m}=b_m$ *uniformly in* m *and* $\lim_{m\to\infty}a_{n,m}=c_n$ *for each large* n*, then both* $\lim_{m\to\infty}b_m$ *and* $\lim_{n\to\infty}c_n$ *exists and are equal to the double limit, i.e.,*

$$\lim_{m\to\infty}\lim_{n\to\infty}a_{n,m}=\lim_{n\to\infty}\lim_{m\to\infty}a_{n,m}=\lim_{\substack{n\to\infty\\m\to\infty}}a_{n,m}.$$

F.2 Technical Proofs

F.2.1 Proof of Lemma 23

Proof. Let Assumptions 1, 2, 3, and 4 hold. Fix an arbitrary sample path $\{x_0, \{Y_i\}_{i=1}^{\infty}\}$. Use $\mathcal{B}$ to denote an arbitrary compact set of x.

$$\begin{aligned}
&\lim_{c\to\infty} \sup_{x\in\mathcal{B}} \sup_{n} \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) \left[H_c(x, Y_{i+1}) - H_\infty(x, Y_{i+1}) \right] \right\| \\
=&\lim_{c\to\infty} \sup_{x\in\mathcal{B}} \sup_{n} \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i)\kappa(c) b(x, Y_{i+1}) \right\| && \text{(by (9.5))} \\
=&\lim_{c\to\infty} \kappa(c) \sup_{x\in\mathcal{B}} \sup_{n} \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x, Y_{i+1}) \right\| \\
=&0 \sup_{x\in\mathcal{B}} \sup_{n} \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x, Y_{i+1}) \right\| && \text{(F.6)}
\end{aligned}$$

We now show that the function

$$x \mapsto \sup_{n} \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x, Y_{i+1}) \right\| \tag{F.7}$$

is Lipschitz continuous. $\forall x, x'$,

$$\begin{aligned}
&\left| \sup_{n} \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x, Y_{i+1}) \right\| - \sup_{n} \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x', Y_{i+1}) \right\| \right| \\
\le& \sup_{n} \sup_{t\in[0,T]} \left| \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x, Y_{i+1}) \right\| - \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x', Y_{i+1}) \right\| \right| \\
& && \text{(by } |\sup_x f(x) - \sup_x g(x)| \le \sup_x |f(x) - g(x)|\text{)} \\
\le& \sup_{n} \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x, Y_{i+1}) - \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) b(x', Y_{i+1}) \right\| \\
\le& \sup_{n} \sup_{t\in[0,T]} \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) \| b(x, Y_{i+1}) - b(x', Y_{i+1}) \| \\
\le& \sup_{n} \sup_{t\in[0,T]} \left(\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L_b(Y_{i+1}) \right) \| x - x' \| && \text{(by (9.6))}
\end{aligned}$$

Additionally, let Assumption 6 or 6′ hold. By Lemma 22 and (F.49),

$$\sup_n \sup_{t\in[0,T]} \left(\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L_b(Y_{i+1}) \right) < \infty$$

can be viewed as the Lipschitz constant. Thus, (F.7) is a continuous function. Since $\mathcal{B}$ is compact, the extreme value theorems asserts that the supremum of (F.7) in $\mathcal{B}$ is attainable at some $x_{\mathcal{B}}$ and is finite. This means the RHS of (F.6) is 0,

$$\lim_{c\to\infty} \sup_{x\in\mathcal{B}} \sup_n \sup_{t\in[0,T]} \left\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) \left[H_c(x, Y_{i+1}) - H_\infty(x, Y_{i+1}) \right] \right\| = 0.$$

□

F.2.2 Proof of Lemma 24

Proof. By (9.24),

$$\sup_n \|\hat{x}(T_n + 0)\| \le 1.$$

$\forall \xi > 0$, by (F.44), $\exists \delta_0$, such that $\forall 0 < \delta \le \delta_0$,

$$\sup_{c\ge 1} \limsup_n \sup_{0\le t_2 - t_1 \le \delta} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_c(0, Y_{i+1}) \right\| \le \xi. \tag{F.8}$$

By (F.48), $\exists \delta_1$, such that $\forall 0 < \delta \le \delta_1$,

$$\limsup_n \sup_{0\le t_2 - t_1 \le \delta} \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) L(Y_{i+1}) \le \xi. \tag{F.9}$$

Without loss of generality, let $t_1 \le t_2$. Then$\forall \delta \le \min\{\delta_0, \delta_1\}$, we have

$$
\begin{aligned}
&\limsup_n \sup_{0\le t_2-t_1\le\delta} \|\hat{x}(T_n+t_1) - \hat{x}(T_n+t_2)\| \\
=&\limsup_n \sup_{0\le t_2-t_1\le\delta} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(\hat{x}(t(i)), Y_{i+1}) \right\| \\
\le&\limsup_n \sup_{0\le t_2-t_1\le\delta} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(\hat{x}(t(i)), Y_{i+1}) \right\| - \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(0, Y_{i+1}) \right\| \\
&+\limsup_n \sup_{0\le t_2-t_1\le\delta} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(0, Y_{i+1}) \right\| \\
\le&\limsup_n \sup_{0\le t_2-t_1\le\delta} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(\hat{x}(t(i)), Y_{i+1}) \right\| - \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(0, Y_{i+1}) \right\| \\
&+\sup_{c\ge 1}\limsup_n \sup_{0\le t_2-t_1\le\delta} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_c(0, Y_{i+1}) \right\| \\
\le&\limsup_n \sup_{0\le t_2-t_1\le\delta} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(\hat{x}(t(i)), Y_{i+1}) \right\| - \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(0, Y_{i+1}) \right\| \\
&+\xi && \text{(by (F.8))} \\
\le&\limsup_n \sup_{0\le t_2-t_1\le\delta} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(\hat{x}(t(i)), Y_{i+1}) - \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_{r_n}(0, Y_{i+1}) \right\| \\
&+\xi \\
\le&\limsup_n \sup_{0\le t_2-t_1\le\delta} \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) \|H_{r_n}(\hat{x}(t(i)), Y_{i+1}) - H_{r_n}(0, Y_{i+1})\| + \xi \\
\le&\limsup_n \sup_{0\le t_2-t_1\le\delta} \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) L(Y_{i+1}) \|\hat{x}(t(i))\| + \xi \\
\le&C_{\hat{x}} \limsup_n \sup_{0\le t_2-t_1\le\delta} \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) L(Y_{i+1}) + \xi && \text{(by Lemma 54)} \\
\le&C_{\hat{x}}\xi + \xi, && \text{(by (F.9))}
\end{aligned}
$$

which implies that $\{\hat{x}(T_n + t)\}$ is equicontinuous in the extended sense.
For $\{z_n(t)\}$, by (9.24) and (9.26), we have

$$\sup_n \|z_n(0)\| \leq 1.$$

Without loss of generality, let $t_1 \leq t_2$. Then $\forall \delta > 0$, we have

$$\begin{aligned}
&\sup_n \sup_{0\leq|t_1-t_2|\leq\delta, 0\leq t_1\leq t_2<T} \|z_n(t_1) - z_n(t_2)\| \\
=&\sup_n \sup_{0\leq|t_1-t_2|\leq\delta, 0\leq t_1\leq t_2<T} \left\| \int_{t_1}^{t_2} h_{r_n}(z_n(s))ds \right\| \\
=&\sup_n \sup_{0\leq|t_1-t_2|\leq\delta, 0\leq t_1\leq t_2<T} \left\| \int_{t_1}^{t_2} \left[h_{r_n}(z_n(s)) - h_{r_n}(0)\right] ds + \int_{t_1}^{t_2} h_{r_n}(0)ds \right\| \\
\leq&\sup_n \sup_{0\leq|t_1-t_2|\leq\delta, 0\leq t_1\leq t_2<T} \int_{t_1}^{t_2} \|h_{r_n}(z_n(s)) - h_{r_n}(0)\|ds + \sup_n \sup_{0\leq|t_1-t_2|\leq\delta, 0\leq t_1\leq t_2<T} \int_{t_1}^{t_2} \|h_{r_n}(0)\|ds \\
\leq&\sup_n \sup_{0\leq|t_1-t_2|\leq\delta, 0\leq t_1\leq t_2<T} \int_{t_1}^{t_2} L\|z_n(s)\|ds + \sup_n \sup_{0\leq|t_1-t_2|\leq\delta, 0\leq t_1\leq t_2<T} \int_{t_1}^{t_2} \|h_{r_n}(0)\|ds \\
& \qquad \text{(by Lemma 50)} \\
\leq&\delta L C_{\hat{x}} + \sup_n \sup_{0\leq|t_1-t_2|\leq\delta, 0\leq t_1\leq t_2<T} \int_{t_1}^{t_2} \|h_{r_n}(0)\|ds \qquad \text{(by Lemma 55)} \\
\leq&\delta(L C_{\hat{x}} + C_H), \qquad \text{(by (F.51))}
\end{aligned}$$

which implies that $\{z_n(t)\}$ is equicontinuous.
For $\{f_n(t)\}$, we have

$$\sup_n f_n(0) = \sup_n \hat{x}(T_n) - z_n(0) = \sup_n \hat{x}(T_n) - \hat{x}(T_n) = 0 < \infty.$$

Because $\{\hat{x}(T_n + t)\}$ and $\{z_n(t)\}$ are equicontinuous, $\forall \epsilon > 0$, $\exists \delta$ such that

$$\begin{aligned}
\limsup_n \sup_{0\leq t_2-t_1\leq\delta} \|\hat{x}(T_n + t_1) - \hat{x}(T_n + t_2)\| \leq \frac{\epsilon}{2}, \\
\sup_n \sup_{0\leq t_2-t_1\leq\delta} \|z_n(t_1) - z_n(t_2)\| \leq \frac{\epsilon}{2}.
\end{aligned}$$

Without loss of generality let $t_1 \leq t_2$. Then $\forall \epsilon$, $\exists \delta$ such that

$$\begin{aligned}
&\limsup_n \sup_{0\leq t_2-t_1\leq\delta} \|f_n(t_1) - f_n(t_2)\| \\
=&\limsup_n \sup_{0\leq t_2-t_1\leq\delta} \|\hat{x}(T_n + t_1) - \hat{x}(T_n + t_2) - (z_n(t_1) - z_n(t_2))\| \\
\leq&\limsup_n \sup_{0\leq t_2-t_1\leq\delta} \|\hat{x}(T_n + t_1) - \hat{x}(T_n + t_2)\| + \limsup_n \sup_{0\leq t_2-t_1\leq\delta} \|z_n(t_1) - z_n(t_2)\| \\
\leq&\limsup_n \sup_{0\leq t_2-t_1\leq\delta} \|\hat{x}(T_n + t_1) - \hat{x}(T_n + t_2)\| + \sup_n \sup_{0\leq t_2-t_1\leq\delta} \|z_n(t_1) - z_n(t_2)\| \\
\leq&\epsilon,
\end{aligned}$$

which implies that $\{f_n\}$ is equicontinuous in the extended sense. □

F.2.3 Proof of Lemma 25

Proof. We can construct a subsequence $\left\{r_{n_{1,k}}\right\}$ that diverges to infinity and satisfies $\forall k,\ \forall n < n_{1,k}$,

$$r_n < r_{n_{1,k}}. \tag{F.10}$$

For example, we can define

$$\begin{aligned} n_{1,0} &\doteq 1 \\ n_{1,k} &\doteq \min\left\{n \mid n > n_{1,k-1}, r_n > r_{n_{1,k-1}} + 1\right\}. \end{aligned} \tag{F.11}$$

Because $\limsup_n r_n = \infty$, we know $\forall k > 0, \left\{n \mid n > n_{1,k-1}, r_n > r_{n_{1,k-1}} + 1\right\} \neq \emptyset$. Because $\forall k > 0$, $r_{n_{1,k}} - r_{n_{1,k-1}} > 1$,

$$\lim_{k\to\infty} r_{n_{1,k}} = \infty. \tag{F.12}$$

Because (F.11) defines $n_{1,k}$ to be the first index that is large enough after $n_{1,k-1}$, (F.10) holds. Otherwise $n_{1,k}$ would not be the first. Define a sequence $\{n_{2,k}\}$ as

$$n_{2,k} \doteq n_{1,k} - 1 \quad \forall k. \tag{F.13}$$

We make two observations. First, $n_{2,k}$ and $n_{1,k}$ are neighbors so $r_{n_{2,k}}$ and $r_{n_{1,k}}$ correspond to $\bar{x}(T_n)$ and $\bar{x}(T_{n+1})$ for some n. Second, by Lemma 56, the increment of $\bar{x}(t)$ in $[T_n, T_{n+1})$ is bounded in the following sense $\forall n$,

$$\|\bar{x}(T_{n+1})\| \leq (\|\bar{x}(T_n)\|C_H + C_H)\, e^{C_H} + \|\bar{x}(T_n)\|$$

where C_H is a positive constant. This means that if $r_{n_{2,k}}$ is not large enough, $r_{n_{1,k}}$ will not be large enough either. We can then prove by contradiction in Lemma 57 that

$$\limsup_k r_{n_{2,k}} = \infty.$$

Thus, using the similar method as (F.11), we can construct a subsequence $\{n_{3,k}\}$ from $\{n_{2,k}\}$ such that

$$\lim_k r_{n_{3,k}} = \infty.$$

Moreover, since $\{n_{3,k} + 1\}$ is a subsequence of $\{n_{1,k}\}$, (F.10) implies that

$$r_{n_{3,k}} < r_{n_{3,k}+1}.$$

Since $\{f_n\}$ is equicontinuous in the extended sense, $\left\{f_{n_{3,k}}\right\}_{k=0,1,\ldots}$ is also equicontinuous in the extended sense. By the Arzela-Ascoli Theorem (Theorem F.1.4), it has a uniformly convergent subsequence, referred to as $\left\{f_{n_{4,k}}\right\}$. Because the sequence $\left\{\hat{x}(T_{n_{4,k}}+t)\right\}$ is also equicontinuous in the extended sense, it has a uniformly convergent subsequence $\{\hat{x}(T_{n_k}+t)\}$. To summarize,

$$\{n_k\} \subseteq \{n_{4,k}\} \subseteq \{n_{3,k}\} \subseteq \{n_{2,k}\} \subseteq \{n_{1,k}-1\} \subseteq \mathbb{N}. \tag{F.14}$$

We construct $\{n_k\}$ in this way because it then inherits all uniform convergence properties. Precisely speaking, by the Arzela-Ascoli theorem in Appendix F.1.4, we have the following corollary.

Corollary 2. *There exist some continuous functions $f^{\lim}(t)$ and $\hat{x}^{\lim}(t)$ such that $\forall t \in [0,T)$,*

$$\begin{aligned}\lim_{k\to\infty} f_{n_k}(t) =& f^{\lim}(t),\\ \lim_{k\to\infty} \hat{x}(T_{n_k}+t) =& \hat{x}^{\lim}(t).\end{aligned}$$

Moreover, the convergence is uniform in t on $[0,T)$.

In terms of the three sequences of functions in (9.29), Corollary 2 has identified that two of them converge along $\{n_k\}$. Lemma 61 further confirms that $z^{\lim}$ is the limit of $\{z_{n_k}\}$. That is $\forall t \in [0,T)$,

$$\lim_{k\to\infty} z_{n_k}(t) = z^{\lim}(t).$$

Moreover, the convergence is uniform in t on $[0,T)$. By (F.14), we have

$$\begin{aligned}\lim_{k\to\infty} r_{n_k} =& \infty,\\ \lim_{k\to\infty} r_{n_k+1} =& \infty,\end{aligned} \tag{F.15}$$

which completes the proof. □

F.2.4 Proof of Lemma 26

Proof. $\forall j$, $\forall k$, $\forall t \in [0, T)$,

$$\begin{aligned}
&\left\| \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s)) ds \right\| \right. \\
&\left. - \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{\infty}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{\infty}(\hat{x}^{\lim}(s)) ds \right\| \right\| \\
\leq & \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s)) ds \right. \\
&\left. - \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{\infty}(\hat{x}(t(i)), Y_{i+1}) + \int_0^t h_{\infty}(\hat{x}^{\lim}(s)) ds \right\| \\
& \qquad \text{(by } |\|a\| - \|b\|| \leq \|a - b\| \text{)} \\
\leq & \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) (H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - H_{\infty}(\hat{x}(t(i)), Y_{i+1})) \right\| + \left\| \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s)) - h_{\infty}(\hat{x}^{\lim}(s)) ds \right\| \\
\leq & \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) (H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - H_{\infty}(\hat{x}(t(i)), Y_{i+1})) \right\| + \int_0^t \left\| h_{r_{n_j}}(\hat{x}^{\lim}(s)) - h_{\infty}(\hat{x}^{\lim}(s)) \right\| ds
\end{aligned} \tag{F.16}$$

By Lemma 54, $\hat{x}(t(i))$ is in a compact set $\mathcal{B}_{\hat{x}}$. By Lemma 23, for the compact set $\mathcal{B}_{\hat{x}}$, $\forall \epsilon > 0$, $\exists j_1$ such that $\forall j \geq j_1$, $\forall k$, $\forall x \in \mathcal{B}$, $\forall t \in [0, T)$,

$$\left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) \left[H_{r_{n_j}}(x, Y_{i+1}) - H_{\infty}(x, Y_{i+1}) \right] \right\| \leq \epsilon. \tag{F.17}$$

Similar to the proof of Lemma 60, we have

$$\lim_{j \to \infty} h_{r_{n_j}}(\hat{x}(T_k + t)) = h_{\infty}(\hat{x}(T_k + t)) \tag{F.18}$$

uniformly in k and $t \in [0, T)$. By (F.18), $\forall \epsilon > 0$, $\exists j_2$ such that $\forall j > j_2$, $\forall k$, $\forall t \in [0, T)$,

$$\left\| h_{r_{n_j}}(\hat{x}(T_k + t)) - h_{\infty}(\hat{x}(T_k + t)) \right\| \leq \epsilon. \tag{F.19}$$

Define $j_0 \doteq \max\{j_1, j_2\}$. $\forall j \geq j_0$, $\forall k$, $\forall t \in [0, T)$,

$$
\begin{aligned}
&\left\| \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s)) ds \right\| \right. \\
&\left. - \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{\infty}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{\infty}(\hat{x}^{\lim}(s)) ds \right\| \right\| \\
\leq & \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) (H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - H_{\infty}(\hat{x}(t(i)), Y_{i+1})) \right\| + T\epsilon \\
& \qquad \text{(by (F.16), (F.19))} \\
\leq & \epsilon + T\epsilon \qquad \text{(by (F.16), (F.17))} \\
\leq & (T+1)\epsilon.
\end{aligned}
$$

This completes the proof of uniform convergence.

□

F.2.5 Proof of Lemma 28

Proof.

$$
\begin{aligned}
&\lim_{\substack{j\to\infty \\ k\to\infty}} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s)) ds \right\| \\
= & \lim_{j\to\infty} \lim_{k\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s)) ds \right\| \\
& \text{(by Lemma 26, 27, and Moore-Osgood Theorem for interchanging limits in Theorem F.1.5)} \\
= & \lim_{j\to\infty} 0 \qquad \text{(by Lemma 27)} \\
= & 0. \qquad \text{(F.20)}
\end{aligned}
$$

□

F.2.6 Proof of Lemma 29

Proof. We now proceed to investigate the property of $f_{n_k}(t)$. $\forall t \in [0, T)$,

$$
\begin{aligned}
&\lim_{k\to\infty} \|f_{n_k}(t)\| \\
\leq &\lim_{k\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_k}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds \right\| \\
&+ \lim_{k\to\infty} \left\| \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds - \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds \right\| && \text{(by (9.35))} \\
= &\lim_{k\to\infty} \left\| \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds - \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds \right\| && \text{(by (F.20))} \\
= &\left\| \int_0^t h_{\infty}(\hat{x}^{\lim}(s))ds - \int_0^t h_{\infty}(z^{\lim}(s))ds \right\|. && \text{(by Lemma 63 and Lemma 64)(F.21)}
\end{aligned}
$$

We now show the relationship between $\hat{x}^{\lim}(t)$ and $z^{\lim}(t)$.

$$
\begin{aligned}
&\left\|\hat{x}^{\lim}(t) - z^{\lim}(t)\right\| && \text{(F.22)} \\
= &\left\| \lim_{k\to\infty} \left[\hat{x}(T_{n_k}) + \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds\right] - \left[\hat{x}^{\lim}(0) + \int_0^t h_{\infty}(z^{\lim}(s))ds\right] \right\| \\
& && \text{(by (9.34) and (9.38))} \\
= &\left\| \hat{x}^{\lim}(0) + \int_0^t h_{\infty}(\hat{x}^{\lim}(s))ds - \left[\hat{x}^{\lim}(0) + \int_0^t h_{\infty}(z^{\lim}(s))ds\right] \right\| && \text{(by Lemma 63)} \\
= &\left\| \int_0^t h_{\infty}(\hat{x}^{\lim}(s))ds - \int_0^t h_{\infty}(z^{\lim}(s))ds \right\| && \text{(F.23)} \\
\leq &\int_0^t L\left\|\hat{x}^{\lim}(s) - z^{\lim}(s)\right\|ds && \text{(by Lemma 50)} \\
\leq &0. && \text{(by Gronwall inequality in Theorem F.1.1)}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\left\| \lim_{k\to\infty} f_{n_k}(t) \right\| \\
\leq &\left\| \int_0^t h_{\infty}(\hat{x}^{\lim}(s))ds - \int_0^t h_{\infty}(z^{\lim}(s))ds \right\| && \text{(by (F.21))} \\
= &\left\|\hat{x}^{\lim}(t) - z^{\lim}(t)\right\| && \text{(by (F.23))} \\
\leq &0. && \text{(by (F.22))}
\end{aligned}
$$

□

F.2.7 Proof of Lemma 30

Proof. According to (9.25), to study $\{z_{n_k}(t)\}$, it is instrumental to study the following ODE

$$\frac{\mathrm{d}\phi_c(t)}{\mathrm{d}t} = h_c(\phi_c(t))$$

for some $c \geq 1$. Let $\phi_{c,x}(t)$ denote the unique solution of the ODE above with the initial condition $\phi_{c,x}(0) = x$. Intuitively, as $c \to \infty$, the above ODE approaches the (ODE@∞). Since any trajectory of (ODE@∞) will diminish to 0 (Assumption 5), $\phi_{c,x}(t)$ should also diminish to some extent for sufficiently large c. Precisely speaking, we have the following lemma.

Lemma 44. *(Corollary 3.3 in Borkar (2009)) There exist $c_1 > 0$ and $\tau > 0$ such that for all initial conditions x with $\|x\| \leq 1$, we have*

$$\|\phi_{c,x}(t)\| \leq \frac{1}{4}$$

for $t \in [\tau, \tau + 1]$ and $c \geq c_1$.

Here the $\frac{1}{4}$ is entirely arbitrary. Now we fix any $c_0 \geq \max\{c_1, 1\}$ and set $T = \tau$. Then Lemma 44 confirms that $z_{n_k}(t)$ will diminish to some extent as t approaches T for sufficiently large k, so does $\hat{x}(T_{n_k} + t)$. We, however, recall that $\hat{x}(T_{n_k} + t)$ and $\bar{x}(T_{n_k} + t)$ are well defined on $[0, T_{n+1} - T_n)$ and we restrict them to $[0, T)$ for applying the Arzela-Ascoli theorem. Lemma 66 processes the excess part $[T, T_{n+1} - T_n)$, by showing that $\bar{x}(T_{n_k} + t)$ cannot grow too much in the excess part. By Lemma 66,

$$\lim_{k\to\infty} \frac{\|\bar{x}(T_{n_k+1})\| - \lim_{t\to T^-} \|\bar{x}(T_{n_k} + t)\|}{\|\bar{x}(T_{n_k})\|} = 0. \tag{F.24}$$

We are now in the position to identify the contradiction. By (F.15), $\exists k_1$ such that $\forall k \geq k_1$,

$$r_{n_k+1} > (c_0 C_H + C_H)\, e^{C_H} + c_0 > c_0 > 1. \tag{F.25}$$

By Lemma 29, $\exists k_2$ such that $\forall k \geq k_2$,

$$\lim_{t\to T^-} \|f_{n_k}(t)\| = \lim_{t\to T^-} \|\hat{x}(T_{n_k} + t) - z_{n_k}(t)\| \leq \frac{1}{4}. \tag{F.26}$$

By (F.24), $\exists k_3$ such that $\forall k \geq k_3$,

$$\frac{\|\bar{x}(T_{n_k+1})\| - \lim_{t\to T^-} \|\bar{x}(T_{n_k} + t)\|}{\|\bar{x}(T_{n_k})\|} \leq \frac{1}{4}. \tag{F.27}$$

By (F.15), $\exists k_4$ such that $\forall k \geq k_4$,

$$r_{n_k} > c_0.$$

Define $k_0 \doteq \max\{k_1, k_2, k_3, k_4\}$. Because $r_{n_{k_0}} > c_0$, by Lemma 44 and (9.25), we have

$$\lim_{t \to T^-} \left\| z_{n_{k_0}}(t) \right\| \leq \frac{1}{4}. \tag{F.28}$$

We have

$$\begin{aligned}
&\lim_{t \to T^-} \left\| \hat{x}(T_{n_{k_0}} + t) \right\| \\
\leq &\lim_{t \to T^-} \left\| \hat{x}(T_{n_{k_0}} + t) - z_{n_{k_0}}(t) \right\| + \left\| z_{n_{k_0}}(t) \right\| \\
\leq &\frac{1}{2}. \qquad \text{(by (F.26) and (F.28))}
\end{aligned} \tag{F.29}$$

This implies

$$\begin{aligned}
&\frac{\left\| \bar{x}(T_{n_{k_0}+1}) \right\|}{\left\| \bar{x}(T_{n_{k_0}}) \right\|} \\
= &\frac{\left\| \bar{x}(T_{n_{k_0}+1}) \right\| - \lim_{t \to T^-} \left\| \bar{x}(T_{n_{k_0}} + t) \right\|}{\left\| \bar{x}(T_{n_{k_0}}) \right\|} + \frac{\lim_{t \to T^-} \left\| \bar{x}(T_{n_{k_0}} + t) \right\|}{\left\| \bar{x}(T_{n_{k_0}}) \right\|} \\
\leq &\frac{1}{4} + \frac{\lim_{t \to T^-} \left\| \bar{x}(T_{n_{k_0}} + t) \right\|}{\left\| \bar{x}(T_{n_{k_0}}) \right\|} \qquad \text{(by (F.27))} \\
= &\frac{1}{4} + \frac{\lim_{t \to T^-} \left\| \hat{x}(T_{n_{k_0}} + t) \right\|}{\left\| \hat{x}(T_{n_{k_0}}) \right\|} \qquad \text{(by (9.22))} \\
= &\frac{1}{4} + \lim_{t \to T^-} \left\| \hat{x}(T_{n_{k_0}} + t) \right\| \qquad \left(\left\| \hat{x}(T_{n_{k_0}}) \right\| = 1 \text{ because of } r_{n_{k_0}} > c_0 > 1 \text{ and (9.22)} \right) \\
\leq &\frac{3}{4}. \qquad \text{(by (F.29))}
\end{aligned} \tag{F.30}$$

Now, we can derive the following inequality.

$$\begin{aligned}
r_{n_{k_0}+1} &= \left\| \bar{x}(T_{n_{k_0}+1}) \right\| && \text{(by (F.25))} \\
&\leq \frac{3}{4} \left\| \bar{x}(T_{n_{k_0}}) \right\| && \text{(by (F.30))} \\
&\leq \left\| \bar{x}(T_{n_{k_0}}) \right\| \\
&\leq r_{n_{k_0}}, && \text{(by } r_{n_{k_0}} > c_0 > 1 \text{ and (9.23))}
\end{aligned}$$

which completes the proof. □

F.2.8 Proof of Lemma 31

Proof.

$$
\begin{aligned}
&\sup_n \sup_{i\in\{i|m(T_n)\le m(T_n)+i<m(T_{n+1})\}} \left\|x_{m(T_n)+i}\right\| - \left\|x_{m(T_n)}\right\| \\
\le &\sup_n \sup_{i\in\{i|m(T_n)\le m(T_n)+i<m(T_{n+1})\}} \left\|x_{m(T_n)+i} - x_{m(T_n)}\right\| \\
= &\sup_n \sup_{i\in\{i|m(T_n)\le m(T_n)+i<m(T_{n+1})\}} \|\bar{x}(t(m(T_n)+i)) - \bar{x}(T_n)\| \\
= &\sup_n \sup_{t\in[T_n,T_{n+1})} \|\bar{x}(T_n+t) - \bar{x}(T_n)\| && \text{(by (9.21))} \\
\le &\sup_n \sup_{t\in[T_n,T_{n+1})} [\|\bar{x}(T_n)\|C_H + C_H]\, e^{C_H} && \text{(by (F.55))} \\
\le &\sup_n \sup_{t\in[T_n,T_{n+1})} [r_n C_H + C_H]\, e^{C_H} && \text{(by (9.23))} \\
= &\sup_n [r_n C_H + C_H] e^{C_H} \\
< &\infty. && \text{(by (9.39))}
\end{aligned}
$$

□

F.2.9 Proof of Corollary 1

This proof follows the idea of the proof of Theorem 2.1 in Chapter 5 of Kushner and Yin (2003).

Proof. Let Assumptions 1 - 5 hold. Let Assumption 6 or 6′ hold. To prove convergence results on $t \in (-\infty, \infty)$ in Corollary 1, we fix an arbitrary sample path $\{x_0, \{Y_i\}_{i=1}^{\infty}\}$. The stability results from Theorem 19 hold. To prove properties on $t \in (-\infty, \infty)$, we first fix an arbitrary $\tau > 0$ and show properties on $\forall t \in [-\tau, \tau]$.

Definition 5. *$\forall n \in \mathbb{N}$, define $\bar{z}_n(t)$ as the solution to the ODE (9.13) in $(-\infty, \infty)$ with an initial condition*

$$\bar{z}_n(0) = \bar{x}(t(n)).$$

Apparently, $\bar{z}_n(t)$ can also be written as

$$\bar{z}_n(t) = \bar{x}(t(n)) + \int_0^t h(\bar{z}_n(s))ds, \quad \forall t \in (-\infty, \infty). \tag{F.31}$$

The major difference between the $\{\bar{z}_n(t)\}$ here and the $\{z_n(t)\}$ in (9.25) is that all $\{\bar{z}_n(t)\}$ here are solutions to one same ODE (9.13), just with different initial conditions,

but $\{z_n(t)\}$ is for different ODEs with different initial conditions and rescale factors r_n and is written as

$$z_n(t) = \hat{x}(T_n) + \int_0^t h_{r_n}(z_n(s))ds. \qquad \text{(Restatement of (9.27))}$$

Ideally, we would like to see that the error of Euler's discretization diminishes asymptotically. With (9.18) and (9.21), $\forall \tau > 0$, $\forall t \in [-\tau, \tau]$,

$$\bar{x}(t(n)+t) = x_{m(t(n)+t)} = \begin{cases} \bar{x}(t(n)) + \sum_{i=n}^{m(t(n)+t)-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) & \text{if } t \geq 0 \\ \bar{x}(t(n)) - \sum_{i=m(t(n)+t)}^{n-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) & \text{if } t < 0. \end{cases} \tag{F.32}$$

Notably, the property (9.18) that $\forall t < 0, m(t) = 0$ in (F.32) ensures $\bar{x}(t(n)+t)$ is well-defined when $t(n) + t < 0$. Precisely speaking, $\forall \tau > 0$, $\forall t \in [-\tau, \tau]$, the discretization error is defined as

$$\bar{f}_n(t) \doteq \bar{x}(t(n)+t) - \bar{z}_n(t). \tag{F.33}$$

and we would like $\bar{f}_n(t)$ diminishes to 0 as $n \to \infty$ in certain sense. To this end, we study the following three sequences of functions

$$\{\bar{x}(t(n)+t)\}_{n=0}^{\infty}, \{\bar{z}_n(t)\}_{n=0}^{\infty}, \left\{\bar{f}_n(t)\right\}_{n=0}^{\infty}.$$

Equicontinuity in the extended sense on domain $(-\infty, \infty)$ is defined as following (Section 4.2.1 in Kushner and Yin (2003)).

Definition 6. *A sequence of functions $\left\{g_n : (-\infty, \infty) \to \mathbb{R}^K\right\}$ is equicontinuous in the extended sense on $(-\infty, \infty)$ if $\sup_n \|g_n(0)\| < \infty$ and $\forall \tau > 0$, $\forall \epsilon > 0$, $\exists \delta > 0$ such that*

$$\limsup_n \sup_{0 \leq |t_1 - t_2| \leq \delta, |t_1| \leq \tau, |t_2| \leq \tau} \|g_n(t_1) - g_n(t_2)\| \leq \epsilon.$$

We show $\{\bar{x}(t(n)+t)\}$, $\{\bar{z}_n(t)\}$ and $\left\{\bar{f}_n(t)\right\}$ are all equicontinuous in the extended sense.

Lemma 45. *The three sequences of functions $\{\bar{x}(t(n)+t)\}_{n=0}^{\infty}$, $\{\bar{z}_n(t)\}_{n=0}^{\infty}$, and $\left\{\bar{f}_n(t)\right\}_{n=0}^{\infty}$ are all equicontinuous in the extended sense on $t \in (-\infty, \infty)$.*

To prove those lemmas, we need the Gronwall inequality in the reverse time in Appendix F.1.2. Compared to lemmas in the main text which have domain $t \in [0, T)$, lemmas in this section have similar proofs because we first fix an arbitrary τ and prove properties on the domain $t \in [-\tau, \tau]$. We omit proofs for Lemma 45 because they are ditto to proofs of Lemma 24. Similar to Lemma 25, we now construct a particular subsequence of interest.

Lemma 46. *There exists a subsequence* $\{n_k\}_{k=0}^{\infty} \subseteq \{0, 1, 2, \dots\}$ *and some continuous functions* $\bar{f}^{\lim}(t)$ *and* $\bar{x}^{\lim}(t)$ *such that* $\forall\tau$, $\forall t \in [-\tau, \tau]$,

$$\lim_{k\to\infty} \bar{f}_{n_k}(t) = \bar{f}^{\lim}(t),$$
$$\lim_{k\to\infty} \bar{x}(T_{n_k} + t) = \bar{x}^{\lim}(t),$$

where both convergences are uniform in t *on* $[-\tau, \tau]$. *Furthermore, let* $\bar{z}^{\lim}(t)$ *denote the unique solution to the ODE* (9.13) *with the initial condition*

$$\bar{z}^{\lim}(0) = \bar{x}^{\lim}(0),$$

in other words,

$$\bar{z}^{\lim}(t) = \bar{x}^{\lim}(0) + \int_0^t h(\bar{z}^{\lim}(s))ds.$$

Then $\forall\tau$, $\forall t \in [-\tau, \tau]$, *we have*

$$\lim_{k\to\infty} \bar{z}_{n_k}(t) = \bar{z}^{\lim}(t),$$

where the convergence is uniform in t *on* $[-\tau, \tau]$.

Its proof is ditto to the proof of Lemma 25 and is omitted. We use the subsequence $\{n_k\}$ intensively in the remaining proofs. Recall that $\bar{f}_n(t)$ denotes the discretization error between $\bar{x}(t(n) + t)$ and $\bar{z}_n(t)$. We now proceed to prove that this discretization error diminishes along $\{n_k\}$. In particular, we aim to prove that $\forall\tau$, $\forall t \in [-\tau, \tau]$,

$$\lim_{k\to\infty} \left\|\bar{f}_{n_k}(t)\right\| = \left\|\bar{f}^{\lim}(t)\right\| = 0.$$

This means $\bar{x}(t(n_k) + t)$ is close to $\bar{z}_{n_k}(t)$ as $k \to \infty$. For $t \in (0, \tau]$, the proof for this part is the same as the proof we have done in Section 9.4.4. Thus, we only discuss the proof for $t \in [-\tau, 0]$. $\forall\tau$, $\forall t \in [-\tau, 0]$,

$$\begin{aligned}
&\lim_{k\to\infty} \left\|\bar{f}_{n_k}(t)\right\| \\
&= \lim_{k\to\infty} \left\| \bar{x}(t(n_k)) - \sum_{i=m(t(n_k)+t)}^{n_k-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) - \bar{z}_{n_k}(t) \right\| \quad \text{(by (F.32) and (F.33))} \\
&= \lim_{k\to\infty} \left\| - \sum_{i=m(t(n_k)+t)}^{n_k-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) - \int_0^t h(\bar{z}_{n_k}(s))ds \right\| \quad \text{(by (F.31))} \\
&\le \lim_{k\to\infty} \left\| - \sum_{i=m(t(n_k)+t)}^{n_k-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) - \int_0^t h(\bar{x}^{\lim}(s))ds \right\| \\
&\quad + \lim_{k\to\infty} \left\| \int_0^t h(\bar{x}^{\lim}(s))ds - \int_0^t h(\bar{z}_{n_k}(s))ds \right\|. \qquad \text{(F.34)}
\end{aligned}$$

We now prove that the first term in the RHS of (F.34) is 0.

Lemma 47. $\forall\tau$, $\forall t \in [-\tau, 0]$,

$$\lim_{k\to\infty} \left\| - \sum_{i=m(t(n_k)+t)}^{n_k-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) - \int_0^t h(\bar{x}^{\lim}(s))ds \right\| = 0.$$

Its proof is ditto to the proof of Lemma 27 and is omitted. This convergence is also simpler than (9.36) because here we have only a single (H, h). But in (9.36), we have a sequence $\{(H_{n_k}, h_{n_k})\}$, for which we have to split it to a double limit (9.37) and then invoke the Moore-Osgood theorem to reduce it to the single (H, h) case.

Lemma 47 confirms that the first term in the RHS of (F.34) is 0. Moreover, it also enables us to rewrite $\bar{x}^{\lim}(t)$ from a summation form to an integral form. $\forall\tau$, $\forall t \in [-\tau, 0]$

$$\begin{aligned}
&\bar{x}^{\lim}(t) \\
=&\lim_{k\to\infty} \bar{x}(t(n_k)) - \sum_{i=m(t(n_k)+t)}^{n_k-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) \\
=&\lim_{k\to\infty} \bar{x}(t(n_k)) + \int_0^t h(\bar{x}^{\lim}(s))ds. \qquad \text{(by Lemma 47)}
\end{aligned}$$

Thus, we can show the following diminishing discretization error.

Lemma 48. $\forall\tau$, $\forall t \in [-\tau, \tau]$,

$$\lim_{k\to\infty} \left\| \bar{f}_{n_k}(t) \right\| = 0.$$

Moreover, the convergence is uniform in t on $[-\tau, \tau]$.

Its proof is ditto to the proof of Lemma 29 and is omitted. This immediately implies that for any $t \in (-\infty, \infty)$

$$\lim_{k\to\infty} \bar{x}(t(n_k) + t) = \bar{z}^{\lim}(t). \tag{F.35}$$

Theorem 19 then yields that

$$\sup_{t\in(-\infty,\infty)} \left\| \bar{z}^{\lim}(t) \right\| < \infty.$$

Let X be the limit set of $\{x_n\}$, i.e., X consists of all the limits of all the convergent subsequences of $\{x_n\}$. By Theorem 19, $\sup_n \|x_n\| < \infty$, so X is bounded and nonempty.

We now prove X is an invariant set of the ODE (9.13). For any $x \in X$, there exists a subsequence $\{x_{n_k}\}$ such that

$$\lim_{k\to\infty} x_{n_k} = x.$$

Since $\{\bar{x}(t(n_k)+t)\}$ is equicontinuous in the extended sense, following the way we arrive at (F.35), we can construct a subsequence $\{n'_k\} \subseteq \{n_k\}$ such that

$$\lim_{k\to\infty} \bar{x}(t(n'_k)+t) = z^{\lim}(t), \tag{F.36}$$

where $z^{\lim}(t)$ is a solution to the ODE (9.13) and $z^{\lim}(0) = x$. The remaining is to show that $z^{\lim}(t)$ lies entirely in X. For any $t \in (-\infty, \infty)$, by the piecewise constant nature of $\bar{x}$ in (F.32), the above limit (F.36) implies that there exists a subsequence of $\{x_n\}$ that converges to $z^{\lim}(t)$, indicating $z^{\lim}(t) \in X$ by the definition of the limit set. We now have proved $\forall x \in X$, there exists a solution $z^{\lim}(t)$ to the ODE (9.13) such that $z^{\lim}(0) = x$ and $\forall t \in (-\infty, \infty), z^{\lim}(t) \in X$. This means X is an invariant set, by definition. In particular, X is a bounded invariant set.

We now prove that $\{x_n\}$ converges to X. Let $\{x_{n_k}\}$ be any convergent subsequence of $\{x_n\}$ with its limit denoted by x. We must have $x \in X$ by the definition of the limit set. So we have proved that all convergent subsequences of $\{x_n\}$ converge to a point in the bounded invariant set X. If $\{x_n\}$ does not converge to X, there must exists a subsequence $\left\{x_{n'_k}\right\}$ such that $\left\{x_{n'_k}\right\}$ is always away from X by some small $\epsilon_0 > 0$, i.e., $\forall k$,

$$\inf_{x\in X} \left\|x_{n'_k} - x\right\| \geq \epsilon_0. \tag{F.37}$$

But $\left\{x_{n'_k}\right\}$ is bounded so it must have a convergent subsequence, which, by the definition of the limit set, converges to some point in X. This contradicts (F.37). So we must have $\{x_n\}$ converges to X, which is a bounded invariant set of the ODE (9.13). This completes the proof. □

F.2.10 Proof of Theorem 20

Proof. For simplicity, we define

$$A' \doteq \begin{bmatrix} -C & A \\ -A^\top & 0 \end{bmatrix},$$
$$b' \doteq \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

We first invoke Corollary 1 to show that

$$\lim_{t\to\infty} x_t = -A'^{-1}b' \quad \text{a.s.}$$

Assumption 1 follows immediately from Lemma 32.

Assumption 2 follows immediately from Assumption 9.5.2.

For Assumption 3, define

$$H_\infty(x,y) \doteq \begin{bmatrix} -C(y) & A(y) \\ -A(y)^\top & 0 \end{bmatrix} x.$$

Then we have

$$H_c(x,y) - H_\infty(x,y) = \frac{1}{c}\begin{bmatrix} b(y) \\ 0 \end{bmatrix}.$$

After noticing

$$\|b((s,a,s',e)) - b((s,a,s',e'))\| = \rho(s,a)|r(s,a)|\|e-e'\|, \quad \forall s,a,s',e,e',$$

Assumption 3 follows immediately from Lemma 32.

For Assumption 4, it can be easily verified that both $H(x,y)$ and $H_\infty(x,y)$ are Lipschitz continuous in x for each y with the Lipschitz constant being

$$L(y) \doteq \left\| \begin{bmatrix} -C(y) & A(y) \\ -A(y)^\top & 0 \end{bmatrix} \right\|.$$

Since $A(y), b(y), C(y)$ are Lipschitz continuous in e for each (s,a,s'), Lemma 32 implies that

$$\begin{aligned} h(x) =& A'x + b', \\ h_\infty(x) =& A'x, \\ L =& \|A'\|. \end{aligned}$$

Assumption 4 then follows.

For Assumption 5, we have

$$\|h_c(x) - h_\infty(x)\| \le \frac{\|b'\|}{c},$$

the uniform convergence of h_c to h_∞ follows immediately. Proving that A' is Hurwitz is a standard exercise using the field of values of A'. We refer the reader to Section 5 of Sutton et al. (2009) for details and omit the proof. This immediately implies the globally asymptotically stability of the following two ODEs

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = A'x(t) + b', \quad \frac{\mathrm{d}x(t)}{\mathrm{d}t} = A'x(t).$$

The unique globally asymptotically equilibrium of the former is $-A'^{-1}b'$. That of the latter is 0. Assumption 5 then follows.

Assumption 6 follows immediately from Lemma 32 and Assumption 9.5.2.

Corollary 1 then implies that

$$\lim_{t\to\infty} x_t = -A'^{-1}b' \quad \text{a.s.}$$

Block matrix inversion immediately shows that the lower half of $A'^{-1}b'$ is $A^{-1}b$, yielding

$$\lim_{t\to\infty} \theta_t = -A^{-1}b \quad \text{a.s.},$$

which completes the proof. □

F.3 Auxiliary Lemmas

Lemma 49.

$$\begin{aligned}\forall n,\ T_{n+1} - T_n \geq & T,\\ \lim_{n\to\infty} T_{n+1} - T_n = & T.\end{aligned}$$

Moreover, $\forall \tau > 0, t_1, t_2$ *such that* $-\tau \leq t_1 \leq t_2 \leq \tau$*, we have*

$$\lim_{n\to\infty} \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i) = t_2 - t_1. \tag{F.38}$$

Proof. $\forall n$,

$$\begin{aligned}& T_{n+1} - T_n\\ =& t(m(T_n + T) + 1) - T_n && \text{(by (9.17))}\\ \geq & T_n + T - T_n && \text{(by (9.15))}\\ \geq & T.\end{aligned}$$

Thus,

$$\lim_{n\to\infty} T_{n+1} - T_n \geq T.$$

With

$$\begin{aligned}& \lim_{n\to\infty} T_{n+1} - T_n\\ =& \lim_{n\to\infty} t(m(T_n + T) + 1) - T_n\\ =& \lim_{n\to\infty} t(m(T_n + T)) + \alpha(m(T_n + T)) - T_n\\ \leq & \lim_{n\to\infty} T_n + T + \alpha(m(T_n + T)) - T_n && \text{(by (9.15))}\\ =& T,\end{aligned}$$

by the squeeze theorem, we have $\lim_{n\to\infty} T_{n+1} - T_n = T$.

To prove (F.38), $\forall\tau$, $\forall -\tau \le t_1 \le t_2 \le \tau$, it suffices to only consider large n such that $t(n) - \tau \ge 0$. We have

$$\begin{aligned}
&\lim_{n\to\infty} \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i) \\
=&\lim_{n\to\infty} t(m(t(n)+t_2)) - t(m(t(n)+t_1)) \\
\le&\lim_{n\to\infty} t(n) + t_2 - t(m(t(n)+t_1)) && \text{(by (9.15))} \\
\le&\lim_{n\to\infty} t(n) + t_2 - (t(n) + t_1 - \alpha(m(t(n)+t_1))) && \text{(by (9.16))} \\
=&t_2 - t_1 + \lim_{n\to\infty} \alpha(m(t(n)+t_1)) \\
=&t_2 - t_1 && \text{(by (9.3))}
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{n\to\infty} \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i) \\
=&\lim_{n\to\infty} t(m(t(n)+t_2)) - t(m(t(n)+t_1)) \\
\ge&\lim_{n\to\infty} t(n) + t_2 - \alpha(m(t(n)+t_2)) - t(m(t(n)+t_1)) && \text{(by (9.16))} \\
\ge&\lim_{n\to\infty} t(n) + t_2 - \alpha(m(t(n)+t_2)) - (t(n)+t_1) && \text{(by (9.15))} \\
=&\lim_{n\to\infty} t_2 - t_1 - \alpha(m(t(n)+t_2)) \\
=&t_2 - t_1. && \text{(by (9.3))}
\end{aligned}$$

By the squeeze theorem, we have

$$\lim_{n} \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i) = t_2 - t_1.$$

□

Lemma 50. *For any $x, x', c \ge 1$, including $c = \infty$,*

$$\|H_c(x,y) - H_c(x',y)\| \le L(y)\|x - x'\|, \tag{F.39}$$

$$\|h_c(x) - h_c(x')\| \le L\|x - x'\|. \tag{F.40}$$

Proof. To prove (F.39), we first consider $1 \le c < \infty$,

$$\begin{aligned}
&\|H_c(x,y) - H_c(x',y)\| \\
=&\left\| \frac{H(cx,y)}{c} - \frac{H(cx',y)}{c} \right\| && \text{(by (9.4))} \\
\le&\frac{\|H(cx,y) - H(cx',y)\|}{c} \\
\le&L(y)\frac{\|cx - cx'\|}{c} && \text{(by (9.7))} \\
=&L(y)\|x - x'\|.
\end{aligned}$$

By (9.8),

$$\|H_\infty(x,y) - H_\infty(x',y)\| \le L(y)\|x - x'\|.$$

To prove (F.40), $\forall x$, $\forall x'$, $\forall c \ge 1$ including $c = \infty$,

$$\begin{aligned}
&\|h_c(x) - h_c(x')\| \\
=&\left\|\mathbb{E}_{y\sim d_y}\left[H_c(x,y) - H_c(x',y)\right]\right\| \\
\le&\mathbb{E}_{y\sim d_y}\left[\|H_c(x,y) - H_c(x',y)\|\right] \\
\le&\mathbb{E}_{y\sim d_y}\left[L(y)\|x - x'\|\right] \\
\le&L\|x - x'\|.
\end{aligned}$$

□

Lemma 51. $\forall x$,

$$\sup_{c\ge 1}\|h_c(0)\| < \infty, \tag{F.41}$$

$$\sup_{c\ge 1}\limsup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)\left[H_c(x,Y_{i+1}) - h_c(x)\right]\right\| = 0 \quad a.s. \tag{F.42}$$

$$\sup_{c\ge 1}\sup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)H_c(0,Y_{i+1})\right\| < \infty \quad a.s. \tag{F.43}$$

$$\lim_{\delta\to 0^+}\sup_{c\ge 1}\limsup_n \sup_{0\le t_2-t_1\le\delta}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)H_c(0,Y_{i+1})\right\| = 0 \quad a.s. \tag{F.44}$$

Proof. **Proof of (F.41):**

$$\sup_{c\ge 1}\|h_c(0)\| = \sup_{c\ge 1}\left\|\frac{h(0)}{c}\right\| \le \sup_{c\ge 1}\|h(0)\| = \|h(0)\| < \infty.$$

Proof of (F.42)**:** $\forall x$,

$$
\begin{aligned}
&\sup_{c\geq 1}\limsup_{n}\sup_{0\leq t_1\leq t_2\leq T_{n+1}-T_n}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)\left[H_c(x,Y_{i+1})-h_c(x)\right]\right\| \\
=&\sup_{c\geq 1}\limsup_{n}\sup_{0\leq t_1\leq t_2\leq T_{n+1}-T_n}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)\left[\frac{H(cx,Y_{i+1})}{c}-\frac{h(cx)}{c}\right]\right\| \\
=&\sup_{c\geq 1}\frac{1}{c}\limsup_{n}\sup_{0\leq t_1\leq t_2\leq T_{n+1}-T_n}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)\left[H(cx,Y_{i+1})-h(cx)\right]\right\| \\
\leq&\sup_{c\geq 1}\frac{1}{c}\limsup_{n}\sup_{0\leq t_1\leq t_2\leq T+\sup_j\alpha(j)}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)\left[H(cx,Y_{i+1})-h(cx)\right]\right\| \\
&\qquad\qquad(\forall n, T_{n+1}-T_n\leq T+\sup_j\alpha(j)) \\
=&\sup_{c\geq 1}\frac{1}{c}\cdot 0 \qquad\qquad\text{(by Lemma 22)} \\
=&0. \qquad\qquad\text{(F.45)}
\end{aligned}
$$

Proof of (F.43):

$$
\begin{aligned}
&\limsup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H(0, Y_{i+1}) \right\| \\
=&\limsup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i)[H(0, Y_{i+1}) - h(0) + h(0)] \right\| \\
\le&\limsup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i)[H(0, Y_{i+1}) - h(0)] \right\| \\
&+\limsup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) h(0) \right\| \\
\le&\limsup_n \sup_{0\le t_1\le t_2\le T+\sup_j \alpha(j)} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i)[H(0, Y_{i+1}) - h(0)] \right\| \\
&+\limsup_n \sup_{0\le t_1\le t_2\le T+\sup_j \alpha(j)} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) h(0) \right\| \quad (\forall n, T_{n+1}-T_n \le T+\sup_j \alpha(j)) \\
=&\limsup_n \sup_{0\le t_1\le t_2\le T+\sup_j \alpha(j)} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) h(0) \right\| \qquad \text{(by Lemma 22)} \\
=&\|h(0)\| \limsup_n \sup_{0\le t_1\le t_2\le T+\sup_j \alpha(j)} \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) \\
=&\|h(0)\|(T+\sup_j \alpha(j)) \qquad \text{(by Lemma 49)} \\
<&\infty. \qquad \text{(F.46)}
\end{aligned}
$$

We now consider c in the above bounds. We first get

$$
\begin{aligned}
&\sup_{c\ge 1}\limsup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_c(0, Y_{i+1}) \right\| \\
=&\sup_{c\ge 1}\limsup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) \frac{H(0, Y_{i+1})}{c} \right\| \qquad \text{(by (9.4))} \\
=&\sup_n \sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H(0, Y_{i+1}) \right\| \qquad \text{(by } c\ge 1\text{)} \\
<&\infty. \qquad \text{(by (F.46))}
\end{aligned}
$$

Proof of (F.44):

$$
\begin{aligned}
&\lim_{\delta\to 0^+}\sup_{c\geq 1}\limsup_{n}\sup_{0\leq t_2-t_1\leq\delta}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)H_c(0,Y_{i+1})\right\| \\
\leq&\lim_{\delta\to 0^+}\sup_{c\geq 1}\limsup_{n}\sup_{0\leq t_2-t_1\leq\delta}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)\left[H_c(0,Y_{i+1})-h_c(0)\right]\right\| \\
&+\lim_{\delta\to 0^+}\sup_{c\geq 1}\limsup_{n}\sup_{0\leq t_2-t_1\leq\delta}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)h_c(0)\right\| \\
\leq&0+\lim_{\delta\to 0^+}\sup_{c\geq 1}\limsup_{n}\sup_{0\leq t_2-t_1\leq\delta}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)h_c(0)\right\| \qquad \text{(by (F.45))} \\
\leq&0+\lim_{\delta\to 0^+}\sup_{c\geq 1}\limsup_{n}\sup_{0\leq t_2-t_1\leq\delta}\left\|\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)\frac{h(0)}{c}\right\| \\
\leq&0+\|h(0)\|\lim_{\delta\to 0^+}\sup_{c\geq 1}\frac{1}{c}\limsup_{n}\sup_{0\leq t_2-t_1\leq\delta}\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i) \\
\leq&\|h(0)\|\lim_{\delta\to 0^+}\sup_{c\geq 1}\frac{1}{c}\delta \qquad \text{(by (F.38))} \\
=&\|h(0)\|\lim_{\delta\to 0^+}\delta \\
=&0.
\end{aligned}
$$

□

Lemma 52.

$$
\sup_{n}\sup_{0\leq t_1\leq t_2\leq T_{n+1}-T_n}\left(\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)L(Y_{i+1})\right)<\infty\quad a.s., \tag{F.47}
$$

$$
\lim_{\delta\to 0^+}\limsup_{n}\sup_{0\leq t_2-t_1\leq\delta}\left(\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)L(Y_{i+1})\right)=0\quad a.s., \tag{F.48}
$$

$$
\sup_{n}\sup_{0\leq t_1\leq t_2\leq T_{n+1}-T_n}\left(\sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1}\alpha(i)L_b(Y_{i+1})\right)<\infty\quad a.s. \tag{F.49}
$$

Its proof is similar to the proof of Lemma 51 and is thus omitted.

Lemma 53. *Fix a sample path $\{x_0, \{Y_i\}_{i=1}^{\infty}\}$, there exists a constant C_H such that*

$$LT \le C_H, \tag{F.50}$$

$$\sup_{c\ge 1} \|h_c(0)\| \le \frac{C_H}{T}, \tag{F.51}$$

$$\sup_{c\ge 1}\sup_{n}\sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_c(0, Y_{i+1}) \right\| \le C_H, \tag{F.52}$$

$$\sup_{n}\sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) L(Y_{i+1}) \le C_H. \tag{F.53}$$

Moreover, for the presentation convenience, we denote

$$C_{\hat{x}} \doteq [1 + C_H]\, e^{C_H}. \tag{F.54}$$

Proof. Fix a sample path $\{x_0, \{Y_i\}_{i=1}^{\infty}\}$,

$$LT < \infty, \qquad (L \text{ and } T \text{ are constants})$$

$$\sup_{c\ge 1} \|h_c(0)\| T < \infty, \qquad (\text{by (F.41)})$$

$$\sup_{c\ge 1}\sup_{n}\sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_c(0, Y_{i+1}) \right\| < \infty, \qquad (\text{by (F.43)})$$

$$\sup_{n}\sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) L(Y_{i+1}) < \infty. \qquad (\text{by (F.47)})$$

Thus, there exists a constant C_H such that

$$LT \le C_H$$

$$\sup_{c\ge 1} \|h_c(0)\| \le \frac{C_H}{T},$$

$$\sup_{c\ge 1}\sup_{n}\sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) H_c(0, Y_{i+1}) \right\| \le C_H,$$

$$\sup_{n}\sup_{0\le t_1\le t_2\le T_{n+1}-T_n} \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i) L(Y_{i+1}) \le C_H.$$

□

Lemma 54. $\sup_{n, t\in[0,T)} \|\hat{x}(T_n + t)\| \le C_{\hat{x}}$.

Proof. $\forall n \in \mathbb{N}, t \in [0,T)$,

$$
\begin{aligned}
&\|\hat{x}(T_n+t)\| \\
=&\left\|\hat{x}(T_n)+\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)H_{r_n}(\hat{x}(t(i)),Y_{i+1})\right\| \\
\leq&\|\hat{x}(T_n)\|+\left\|\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)H_{r_n}(\hat{x}(t(i)),Y_{i+1})\right\| \\
=&\|\hat{x}(T_n)\|+\left\|\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)\left[H_{r_n}(\hat{x}(t(i)),Y_{i+1})-H_{r_n}(0,Y_{i+1})\right]+\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)H_{r_n}(0,Y_{i+1})\right\| \\
\leq&\|\hat{x}(T_n)\|+\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)\|H_{r_n}(\hat{x}(t(i)),Y_{i+1})-H_{r_n}(0,Y_{i+1})\|+\left\|\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)H_{r_n}(0,Y_{i+1})\right\| \\
\leq&\|\hat{x}(T_n)\|+\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)L(Y_{i+1})\|\hat{x}(t(i))\|+\left\|\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)H_{r_n}(0,Y_{i+1})\right\| \\
\leq&\|\hat{x}(T_n)\|+\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)L(Y_{i+1})\|\hat{x}(t(i))\|+C_H && \text{(by (F.52))} \\
\leq&1+\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)L(Y_{i+1})\|\hat{x}(t(i))\|+C_H && \text{(by (9.24))} \\
\leq&\left[1+C_H\right]e^{\sum_{i=m(T_n)}^{m(T_n+t)-1}\alpha(i)L(Y_{i+1})} \\
&\text{(by } \hat{x}(T_n+t)=\hat{x}(t(m(T_n+t))) \text{ and discrete Gronwall inequality in Theorem F.1.3)} \\
\leq&\left[1+C_H\right]e^{C_H} && \text{(by (F.53))} \\
=&C_{\hat{x}}. && \text{(by (F.54))}
\end{aligned}
$$

□

Lemma 55. $\sup_{n,t\in[0,T)}\|z_n(t)\| \leq C_{\hat{x}}$.

Proof. $\forall n, t \in [0, T)$,

$$
\begin{aligned}
&\|z_n(t)\| \\
=&\left\| z_n(0) + \int_0^t h_{r_n}(z_n(s))ds \right\| \\
\leq&\|z_n(0)\| + \left\| \int_0^t h_{r_n}(z_n(s))ds \right\| \\
\leq&\|z_n(0)\| + \int_0^t \|h_{r_n}(z_n(s)) - h_{r_n}(0)\|ds + \int_0^t \|h_{r_n}(0)\|ds \\
\leq&\|z_n(0)\| + \int_0^t L\|z_n(s)\|ds + \int_0^t \|h_{r_n}(0)\|ds && \text{(by Lemma 50)} \\
\leq&\|z_n(0)\| + \int_0^t L\|z_n(s)\|ds + T\|h_{r_n}(0)\| \\
\leq&\|z_n(0)\| + \int_0^t L\|z_n(s)\|ds + T\frac{C_H}{T} && \text{(by (F.51))} \\
\leq&1 + \int_0^t L\|z_n(s)\|ds + C_H && \text{(by (9.24), (9.26))} \\
\leq&\left[1 + C_H\right] e^{LT} && \text{(by Gronwall inequality in Theorem F.1.1)} \\
\leq&\left[1 + C_H\right] e^{C_H} && \text{(by (F.50))} \\
=&C_{\hat{x}} && \text{(by (F.54))}
\end{aligned}
$$

□

Lemma 56. $\forall n$,

$$\|\bar{x}(T_{n+1})\| \leq \left(\|\bar{x}(T_n)\|C_H + C_H\right) e^{C_H} + \|\bar{x}(T_n)\|$$

where C_H is a positive constant defined in Lemma 53.

Proof. We first show the difference between $\bar{x}(T_{n+1})$ and $\bar{x}(T_n)$ by the following

derivations. $\forall n$, $\forall t \in [0, T_{n+1} - T_n]$,

$$
\begin{aligned}
&\|\bar{x}(T_n + t) - \bar{x}(T_n)\| \\
=&\|\bar{x}(t(m(T_n + t))) - \bar{x}(T_n)\| \\
=&\left\|\bar{x}(T_n) + \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) - \bar{x}(T_n)\right\| \\
=&\left\|\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1})\right\| \\
\leq& \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) \|H(\bar{x}(t(i)), Y_{i+1}) - H(\bar{x}(T_n), Y_{i+1})\| + \left\|\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(\bar{x}(T_n), Y_{i+1})\right\| \\
\leq& \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1}) \|\bar{x}(t(i)) - \bar{x}(T_n)\| + \left\|\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(\bar{x}(T_n), Y_{i+1})\right\| \\
\leq& \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1}) \|\bar{x}(t(i)) - \bar{x}(T_n)\| + \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) \|H(\bar{x}(T_n), Y_{i+1}) - H(0, Y_{i+1})\| \\
&+ \left\|\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(0, Y_{i+1})\right\| \\
\leq& \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1}) \|\bar{x}(t(i)) - \bar{x}(T_n)\| + \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1}) \|\bar{x}(T_n)\| \\
&+ \left\|\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(0, Y_{i+1})\right\| && \text{(by Assumption 4)} \\
=& \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1}) \|\bar{x}(t(i)) - \bar{x}(T_n)\| + \|\bar{x}(T_n)\| \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1}) \\
&+ \left\|\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(0, Y_{i+1})\right\| \\
\leq& \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1}) \|\bar{x}(t(i)) - \bar{x}(T_n)\| + \|\bar{x}(T_n)\| C_H + \left\|\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) H(0, Y_{i+1})\right\| \\
& && \text{(by (F.53))} \\
\leq& \sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1}) \|\bar{x}(t(i)) - \bar{x}(T_n)\| + [\|\bar{x}(T_n)\| C_H + C_H] && \text{(by (F.52))} \\
\leq& [\|\bar{x}(T_n)\| C_H + C_H]\, e^{\sum_{i=m(T_n)}^{m(T_n+t)-1} \alpha(i) L(Y_{i+1})} 248 \\
& && \text{(by discrete Gronwall inequality in Theorem F.1.3)} \\
\leq& [\|\bar{x}(T_n)\| C_H + C_H]\, e^{C_H} && \text{(by (F.53))} \qquad \text{(F.55)}
\end{aligned}
$$

□

Lemma 57.

$$\limsup_k r_{n_{2,k}} = \infty.$$

Proof. We use proof by contradiction. Suppose

$$\limsup_k r_{n_{2,k}} = C_r < \infty$$

where C_r is a constant. $\forall \epsilon > 0$, $\exists k_0$ such that $\forall k \geq k_0$,

$$r_{n_{2,k}} \leq C_r + \epsilon.$$

By Lemma 56, $\forall k \geq k_0$,

$$\begin{aligned}
r_{n_{1,k}} =& \max\left\{\left\|\bar{x}(T_{n_{1,k}})\right\|, 1\right\} && \text{(by (9.23))}\\
=& \max\left\{\left\|\bar{x}(T_{n_{2,k}+1})\right\|, 1\right\} && \text{(by (F.13))}\\
\leq& \left\|\bar{x}(T_{n_{2,k}+1})\right\| + 1\\
\leq& \left(\left\|\bar{x}(T_{n_{2,k}})\right\| C_H + C_H\right) e^{C_H} + \left\|\bar{x}(T_{n_{2,k}})\right\| + 1\\
\leq& \left(r_{n_{2,k}} C_H + C_H\right) e^{C_H} + r_{n_{2,k}} + 1\\
\leq& \left[(C_r + \epsilon) C_H + C_H\right] e^{C_H} + (C_r + \epsilon) + 1\\
<& \infty.
\end{aligned}$$

This contradicts (F.12). Thus,

$$\limsup_k r_{n_{2,k}} = \infty.$$

□

Lemma 58. $\sup_{n,t\in[0,T)} \|h_{r_n}(z_n(t))\| < \infty$.

Proof. $\forall n, \forall t \in [0, T)$,

$$\begin{aligned}
&\|h_{r_n}(z_n(t))\|\\
\leq& \|h_{r_n}(z_n(t)) - h_{r_n}(0)\| + \|h_{r_n}(0)\|\\
\leq& L\|z_n(t)\| + \|h_{r_n}(0)\| && \text{(by Lemma 50)}\\
\leq& LC_{\hat{x}} + \|h_{r_n}(0)\| && \text{(by Lemma 55)}\\
\leq& LC_{\hat{x}} + \frac{C_H}{T}. && \text{(by (9.23) and (F.51))}
\end{aligned}$$

Thus, because $C_{\hat{x}}, C_H$ are independent of n, t, $\sup_{n,t\in[0,T)} \|h_{r_n}(z_n(t))\| < \infty$.

□

Lemma 59. $\sup_{t\in[0,T)} \left\|z^{\lim}(t)\right\| \leq C_{\hat{x}}$.

Proof. $\forall t \in [0,T)$,

$$
\begin{aligned}
&\left\|z^{\lim}(t)\right\| \\
=&\left\|z^{\lim}(0) + \int_0^t h_\infty(z^{\lim}(s))ds\right\| \\
\leq&\left\|z^{\lim}(0)\right\| + \left\|\int_0^t h_\infty(z^{\lim}(s))ds\right\| \\
=&\left\|z^{\lim}(0)\right\| + \left\|\int_0^t \left[h_\infty(z^{\lim}(s)) - h_\infty(0)\right]ds + \int_0^t h_\infty(0)ds\right\| \\
\leq&\left\|z^{\lim}(0)\right\| + \int_0^t \left\|h_\infty(z^{\lim}(s)) - h_\infty(0)\right\|ds + \int_0^t \|h_\infty(0)\|ds \\
\leq&\left\|z^{\lim}(0)\right\| + \int_0^t L\left\|z^{\lim}(s)\right\|ds + \int_0^t \|h_\infty(0)\|ds && \text{(by Lemma 50)} \\
\leq&1 + \int_0^t L\left\|z^{\lim}(s)\right\|ds + \int_0^t \|h_\infty(0)\|ds && \text{(by (9.24), (9.26))} \\
\leq&1 + \int_0^t L\left\|z^{\lim}(s)\right\|ds + T\|h_\infty(0)\| \\
\leq&1 + \int_0^t L\left\|z^{\lim}(s)\right\|ds + C_H && \text{(by Assumption 5 and (F.51))} \\
\leq&\left[1 + C_H\right]e^{\int_0^t L ds} && \text{(by Gronwall inequality in Theorem F.1.1)} \\
\leq&\left[1 + C_H\right]e^{LT} \\
\leq&C_{\hat{x}}. && \text{(by (F.50), (F.54))}
\end{aligned}
$$

□

Lemma 60. $\lim_{k\to\infty} h_{r_{n_k}}(z^{\lim}(t)) = h_\infty(z^{\lim}(t))$ *uniformly in* $t \in [0,T)$.

Proof. By Assumption 5, $\lim_{k\to\infty} h_{r_{n_k}}(v) = h_\infty(v)$ uniformly in a compact set $\left\{v | v \in \mathbb{R}^d, \|v\| \leq C_x\right\}$. By Lemma 59, $\left\{z^{\lim}(t) | t \in [0,T)\right\} \subseteq \left\{v | v \in \mathbb{R}^d, \|v\| \leq C_x\right\}$. Therefore, $\lim_{k\to\infty} h_{r_{n_k}}(z^{\lim}(t)) = h_\infty(z^{\lim}(t))$ uniformly in $\left\{z^{\lim}(t) | t \in [0,T)\right\}$ and on $t \in [0,T)$. □

Lemma 61. $\forall t \in [0,T)$, *we have*

$$\lim_{k\to\infty} z_{n_k}(t) = z^{\lim}(t).$$

Moreover, the convergence is uniform in t *on* $[0,T)$.

Proof. By (9.33), $\forall \delta > 0$, there exists a k_1 such that $\forall k \geq k_1$, $\forall t \in [0, T)$,

$$\left\|\hat{x}(T_{n_k} + t) - \hat{x}^{\lim}(t)\right\| \leq \delta. \tag{F.56}$$

By Lemma 60, there exists a k_2 such that $\forall k \geq k_2$, $\forall t \in [0, T)$,

$$\left\|h_{r_{n_k}}(z^{\lim}(t)) - h_\infty(z^{\lim}(t))\right\| \leq \delta. \tag{F.57}$$

$\forall k \geq \max\{k_1, k_2\}$, $\forall t \in [0, T)$

$$\begin{aligned}
&\left\|z_{n_k}(t) - z^{\lim}(t)\right\| \\
=&\left\|\hat{x}(T_{n_k}) + \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds - \hat{x}^{\lim}(0) - \int_0^t h_\infty(z^{\lim}(s))ds\right\| \\
\leq&\left\|\hat{x}(T_{n_k}) - \hat{x}^{\lim}(0)\right\| + \left\|\int_0^t h_{r_{n_k}}(z_{n_k}(s))ds - \int_0^t h_\infty(z^{\lim}(s))ds\right\| \\
\leq&\delta + \left\|\int_0^t h_{r_{n_k}}(z_{n_k}(s)) - h_\infty(z^{\lim}(s))ds\right\| && \text{(by (F.56))} \\
\leq&\delta + \int_0^t \left\|h_{r_{n_k}}(z_{n_k}(s)) - h_{r_{n_k}}(z^{\lim}(s))\right\|ds + \int_0^t \left\|h_{r_{n_k}}(z^{\lim}(s)) - h_\infty(z^{\lim}(s))\right\|ds \\
\leq&\delta + L\int_0^t \left\|z_{n_k}(s) - z^{\lim}(s)\right\|ds + \int_0^t \left\|h_{r_{n_k}}(z^{\lim}(s)) - h_\infty(z^{\lim}(s))\right\|ds \\
& && \text{(by Lemma 50)} \\
\leq&\delta + t\delta + L\int_0^t \left\|z_{n_k}(s) - z^{\lim}(s)\right\|ds && \text{(by (F.57))} \\
\leq&(\delta + t\delta)e^{Lt} && \text{(by Gronwall inequality in Theorem F.1.1)} \\
\leq&(\delta + T\delta)e^{LT},
\end{aligned}$$

which completes the proof. □

Lemma 62. *For any function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, if $\lim_{\substack{a\to\infty \\ b\to\infty}} f(a, b) = L$ then $\lim_{c\to\infty} f(c, c) = L$ where L is a constant.*

Proof. By definition, $\forall \epsilon > 0, \exists a_0, b_0$ such that $\forall a > a_0, b > b_0$, $\|f(a, b) - L\| < \epsilon$. Thus, $\forall \epsilon > 0, \exists c_0 = \max\{a_0, b_0\}$ such that $\forall c > c_0$, $\|f(c, c) - L\| < \epsilon$. □

Lemma 63. $\forall t \in [0, T)$,

$$\lim_{k\to\infty} \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds = \int_0^t h_\infty(\hat{x}^{\lim}(s))ds.$$

Proof. From Lemma 54, it is easy to see that

$$\sup_{t\in[0,T)} \left\|\hat{x}^{\lim}(t)\right\| < \infty,$$

which, similar to Lemma 58, implies that

$$\sup_{k,t\in[0,T)} \left\| h_{r_{n_k}}\left(\hat{x}^{\lim}(t)\right)\right\| < \infty.$$

By the dominated convergence theorem, $\forall t \in [0,T)$,

$$\lim_{k\to\infty} \int_0^t h_{r_{n_k}}(\hat{x}^{\lim}(s))ds = \int_0^t \lim_{k\to\infty} h_{r_{n_k}}(\hat{x}^{\lim}(s))ds = \int_0^t h_\infty(\hat{x}^{\lim}(s))ds,$$

which completes the proof. □

Lemma 64. $\forall t \in [0,T)$,

$$\lim_{k\to\infty} \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds = \int_0^t h_\infty(z^{\lim}(s))ds.$$

Proof. $\forall \epsilon > 0$, by Lemma 60, $\exists k_0$ such that $\forall k \geq k_0$, $\forall t \in [0,T)$,

$$\left\| h_{r_{n_k}}(z^{\lim}(s)) - h_\infty(z^{\lim}(s))\right\| \leq \epsilon. \tag{F.58}$$

By Lemma 61, $\exists k_1$ such that $\forall k \geq k_1$, $\forall t \in [0,T)$,

$$\left\| z_{n_k}(t) - z^{\lim}(t)\right\| \leq \epsilon. \tag{F.59}$$

Thus, $\forall k \geq \max\{k_0, k_1\}$, $\forall t \in [0,T)$,

$$\begin{aligned}
&\left\| \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds - \int_0^t h_\infty(z^{\lim}(s))ds \right\| \\
\leq &\left\| \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds - \int_0^t h_{r_{n_k}}(z^{\lim}(s))ds \right\| + \left\| \int_0^t h_{r_{n_k}}(z^{\lim}(s))ds - \int_0^t h_\infty(z^{\lim}(s))ds \right\| \\
\leq &\int_0^t \left\| h_{r_{n_k}}(z_{n_k}(s)) - h_{r_{n_k}}(z^{\lim}(s))\right\| ds + \int_0^t \left\| h_{r_{n_k}}(z^{\lim}(s)) - h_\infty(z^{\lim}(s))\right\| ds \\
\leq &\int_0^t \left\| h_{r_{n_k}}(z_{n_k}(s)) - h_{r_{n_k}}(z^{\lim}(s))\right\| ds + T\epsilon && \text{(by (F.58))} \\
\leq &\int_0^t L\left\| z_{n_k}(s) - z^{\lim}(s)\right\| ds + T\epsilon && \text{(by Lemma 50)} \\
\leq &LT\epsilon + T\epsilon. && \text{(by (F.59))}
\end{aligned}$$

Thus, $\forall t \in [0,T)$,

$$\lim_{k\to\infty} \int_0^t h_{r_{n_k}}(z_{n_k}(s))ds = \int_0^t h_\infty(z^{\lim}(s))ds.$$

□

Lemma 65.

$$\lim_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i) L(Y_{i+1}) \right\| = 0, \tag{F.60}$$

$$\lim_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i) H(0, Y_{i+1}) \right\| = 0. \tag{F.61}$$

Proof.

$$\begin{aligned}
&\limsup_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i) L(Y_{i+1}) \right\| \\
=&\limsup_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i)[L(Y_{i+1}) - L] + \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i) L \right\| \\
\le&\limsup_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i)[L(Y_{i+1}) - L] \right\| + \limsup_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i) L \right\| \\
\le&\limsup_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i)[L(Y_{i+1}) - L] \right\| + L \limsup_n \alpha(m(T_{n+1}) - 1) \\
\le&\limsup_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i)[L(Y_{i+1}) - L] \right\| + 0 \qquad \text{(by (9.3))} \\
\le&\limsup_n \sup_{0\le t_1\le t_2\le T+\sup_j \alpha(j)} \left\| \sum_{i=m(T_n+t_1)}^{m(T_n+t_2)-1} \alpha(i)[L(Y_{i+1}) - L] \right\| \\
=&0. \qquad \text{(by (9.20))}
\end{aligned}$$

This implies

$$\lim_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i) L(Y_{i+1}) \right\| = 0.$$

Following a similar proof, we have

$$\lim_n \lim_{t\to T^-} \left\| \sum_{i=m(T_n+t)}^{m(T_{n+1})-1} \alpha(i) H(0, Y_{i+1}) \right\| = 0.$$

□

Lemma 66. $\lim_{k\to\infty} \frac{\left\|\bar{x}(T_{n_k+1})\right\| - \lim_{t\to T^-} \left\|\bar{x}(T_{n_k}+t)\right\|}{\left\|\bar{x}(T_{n_k})\right\|} = 0.$

Proof. We first analyze the numerator. $\forall k$,

$$
\begin{aligned}
&\left| \|\bar{x}(T_{n_k+1})\| - \lim_{t\to T^-} \|\bar{x}(T_{n_k}+t)\| \right| \\
&= \lim_{t\to T^-} \left| \|\bar{x}(T_{n_k+1})\| - \|\bar{x}(T_{n_k}+t)\| \right| \\
&\le \lim_{t\to T^-} \|\bar{x}(T_{n_k+1}) - \bar{x}(T_{n_k}+t)\| \\
&= \lim_{t\to T^-} \left\| \bar{x}(T_{n_k}) + \sum_{i=m(T_{n_k})}^{m(T_{n_k+1})-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) - \bar{x}(T_{n_k}) - \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) \right\| \\
&= \lim_{t\to T^-} \left\| \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) H(\bar{x}(t(i)), Y_{i+1}) \right\| \\
&\le \lim_{t\to T^-} \left\| \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) \left[H(\bar{x}(t(i)), Y_{i+1}) - H(0, Y_{i+1}) \right] \right\| + \left\| \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) H(0, Y_{i+1}) \right\| \\
&\le \lim_{t\to T^-} \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) L(Y_{i+1}) \|\bar{x}(t(i)\| + \left\| \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) H(0, Y_{i+1}) \right\| \\
&= \|\bar{x}(t(m(T_{n_k+1})-1)\| \left[\lim_{t\to T^-} \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) L(Y_{i+1}) \right] + \lim_{t\to T^-} \left\| \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) H(0, Y_{i+1}) \right\| \\
&\qquad\qquad (\forall k, \lim_{t\to T^-} m(T_{n_k}+t) = m(T_{n_k+1})-1) \\
&\le \left([\|\bar{x}(T_{n_k})\| C_H + C_H]\, e^{C_H} + \|\bar{x}(T_{n_k})\| \right) \left[\lim_{t\to T^-} \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) L(Y_{i+1}) \right] \\
&\quad + \lim_{t\to T^-} \left\| \sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1} \alpha(i) H(0, Y_{i+1}) \right\|. \qquad \text{(by (F.55))}
\end{aligned}
$$

By (F.15), we have

$$\lim_{k\to\infty} \|\bar{x}(T_{n_k})\| = \lim_{k\to\infty} r_{n_k} = \infty. \tag{F.62}$$

Thus,

$$
\begin{aligned}
&\lim_{k\to\infty}\left|\frac{\|\bar{x}(T_{n_k+1})\| - \lim_{t\to T^-}\|\bar{x}(T_{n_k}+t)\|}{\|\bar{x}(T_{n_k})\|}\right| \\
=&\lim_{k\to\infty}\frac{\left|\|\bar{x}(T_{n_k+1})\| - \lim_{t\to T^-}\|\bar{x}(T_{n_k}+t)\|\right|}{\|\bar{x}(T_{n_k})\|} \\
=&\lim_{k\to\infty}\frac{\left(\left[\|\bar{x}(T_{n_k})\|C_H + C_H\right]e^{C_H} + \|\bar{x}(T_{n_k})\|\right)\left[\lim_{t\to T^-}\sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1}\alpha(i)L(Y_{i+1})\right]}{\|\bar{x}(T_{n_k})\|} \\
&+\lim_{k\to\infty}\frac{\lim_{t\to T^-}\left\|\sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1}\alpha(i)H(0,Y_{i+1})\right\|}{\|\bar{x}(T_{n_k})\|} \\
\le&\left(C_He^{C_H}+1\right)\left[\lim_{k\to\infty}\lim_{t\to T^-}\sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1}\alpha(i)L(Y_{i+1})\right] + \lim_{k\to\infty}\frac{\lim_{t\to T^-}\left\|\sum_{i=m(T_{n_k}+t)}^{m(T_{n_k+1})-1}\alpha(i)H(0,Y_{i+1})\right\|}{\|\bar{x}(T_{n_k})\|} \\
&\hspace{20em}\text{(by (F.62))} \\
\le&\left(C_He^{C_H}+1\right)\cdot 0 + 0 \hspace{12em}\text{(by (F.60) and (F.61))} \\
=&0.
\end{aligned}
$$

□

F.4 Proofs for Completeness

Proofs in this section have used ideas and sketches from Kushner and Yin (2003) but are self-contained and complete.

F.4.1 Proof of Lemma 22

Proof. **Case 1:** Let Assumptions 1, 2, 4, and 6 hold.
Fixed an arbitrary $\tau > 0$. For an arbitrary $x, t \in (-\infty, \infty)$, define

$$
\begin{aligned}
\psi(i) &\doteq H(x, Y_{i+1}) - h(x), \\
S(n) &\doteq \sum_{i=0}^{n-1}\psi(i), \\
\Psi(t) &\doteq \sum_{i=0}^{m(t)-1}\alpha(i)\psi(i).
\end{aligned}
$$

Here, we use (9.18) so that $\forall t < 0, m(t) = 0$ and the convention that $\sum_{k=i}^{j}\alpha(k) = 0$ when $j < i$. Fix a sample path $\{x_0, \{Y_i\}_{i=1}^{\infty}\}$ where Assumptions 1, 2, 4, & 6 hold.

Assumption 6 implies that

$$\lim_{n\to\infty} \alpha(n)S(n+1) = 0.$$

Use subscript j to denote the jth dimension of a vector, we then have

$$\limsup_{n\to\infty} \sup_{-\tau\le t\le\tau} |\alpha(m(t(n)+t))S(m(t(n)+t)+1)_j| = 0. \tag{F.63}$$

Moreover, for $\forall t \in [-\tau, \tau]$, we have

$$\begin{aligned}
\Psi(t) =& \sum_{i=0}^{m(t)-1} \alpha(i)\psi(i) \\
=& \sum_{i=0}^{m(t)-1} \alpha(i) \left[\sum_{j=0}^{i} \psi(j) - \sum_{j=0}^{i-1} \psi(j)\right] \\
=& \sum_{i=0}^{m(t)-1} \alpha(i) \sum_{j=0}^{i} \psi(j) - \sum_{i=0}^{m(t)-1} \alpha(i) \sum_{j=0}^{i-1} \psi(j) \\
=& \sum_{i=0}^{m(t)-1} \alpha(i) \sum_{j=0}^{i} \psi(j) - \sum_{i=0}^{m(t)-2} \alpha(i+1) \sum_{j=0}^{i} \psi(j) \\
=& \alpha(m(t)-1) \sum_{i=0}^{m(t)-1} \psi(i) + \sum_{i=0}^{m(t)-2} [\alpha(i) - \alpha(i+1)] \sum_{j=0}^{i} \psi(j) \\
=& \alpha(m(t)-1) \sum_{i=0}^{m(t)-1} \psi(i) + \sum_{i=0}^{m(t)-2} S(i+1)[\alpha(i) - \alpha(i+1)] \\
=& \alpha(m(t)-1)S(m(t)) + \sum_{i=0}^{m(t)-2} S(i+1)\frac{\alpha(i) - \alpha(i+1)}{\alpha(i)}\alpha(i).
\end{aligned} \tag{F.64}$$

Thus, for any dimension j,

$$
\begin{aligned}
&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} \left| \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i)(H(x,Y_{i+1})_j - h(x)_j) \right| \\
=&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} |\Psi(t(n)+t_2)_j - \Psi(t(n)+t_1)_j| \\
\le&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} |\alpha(m(t(n)+t_2)-1)S(m(t(n)+t_2))_j| \\
&+ |\alpha(m(t(n)+t_1)-1)S(m(t(n)+t_1))_j| \\
&+ \left| \sum_{i=m(t(n)+t_1)-1}^{m(t(n)+t_2)-2} S(i+1)_j \frac{\alpha(i)-\alpha(i+1)}{\alpha(i)} \alpha(i) \right| && \text{(by (F.64))} \\
=&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} \left| \sum_{i=m(t(n)+t_1)-1}^{m(t(n)+t_2)-2} S(i+1)_j \frac{\alpha(i)-\alpha(i+1)}{\alpha(i)} \alpha(i) \right| && \text{(by (F.63))} \\
\le&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} \sum_{i=m(t(n)+t_1)-1}^{m(t(n)+t_2)-2} \left| S(i+1)_j \frac{\alpha(i)-\alpha(i+1)}{\alpha(i)} \alpha(i) \right| \\
\le&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} \sum_{i=m(t(n)+t_1)-1}^{m(t(n)+t_2)-2} |\alpha(i)S(i+1)_j| \left| \frac{\alpha(i)-\alpha(i+1)}{\alpha(i)} \right| \\
\le&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} \left(\sup_{m(t(n)+t_1)-1\le i\le m(t(n)+t_2)-2} |\alpha(i)S(i+1)_j| \right) \sum_{i=m(t(n)+t_1)-1}^{m(t(n)+t_2)-2} \left| \frac{\alpha(i)-\alpha(i+1)}{\alpha(i)} \right| \\
\le&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} \left(\sup_{m(t(n)+t_1)-1\le i\le m(t(n)+t_2)-2} |\alpha(i)S(i+1)_j| \right) C_\alpha \sum_{i=m(t(n)+t_1)-1}^{m(t(n)+t_2)-2} \alpha(i) \\
&\text{(by Assumption 2, } C_\alpha \text{ is a constant from the big } \mathcal{O} \text{ notation–} \tfrac{\alpha(n)-\alpha(n+1)}{\alpha(n)} = \mathcal{O}(\alpha(n)) \text{)} \\
=&\limsup_{n\to\infty} \left[\sup_{-\tau\le t_1\le t_2\le\tau} \left(\sup_{m(t(n)+t_1)-1\le i\le m(t(n)+t_2)-2} |\alpha(i)S(i+1)_j| \right) \right. \\
&\left. \cdot C_\alpha \left(\sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i) + \alpha(m(t(n)+t_1)-1) \right) \right] \\
=&\limsup_{n\to\infty} \sup_{-\tau\le t_1\le t_2\le\tau} \left(\sup_{m(t(n)+t_1)-1\le i\le m(t(n)+t_2)-2} |\alpha(i)S(i+1)_j| \right) C_\alpha (t_2 - t_1 + \alpha(m(t(n)+t_1)-1)) \\
& && \text{(by (F.38))} \\
\le&\limsup_{n\to\infty} \left(\sup_{m(t(n)-\tau)-1\le i} |\alpha(i)S(i+1)_j| \right) C_\alpha (t_2 - t_1 + \alpha(m(t(n)+t_1)-1)) \\
\le&2C_\alpha\tau \limsup_{n\to\infty} \left(\sup_{m(t(n)-\tau)-1\le i} |\alpha(i)S(i+1)_j| \right) \\
\le&2C_\alpha\tau \limsup_{n\to\infty} \left(\sup_{n\le i} |\alpha(i)S(i+1)_j| \right) \\
=&0. && \text{(by (F.63))}
\end{aligned}
$$

Thus, $\forall \tau > 0$, $\forall x$,

$$\limsup_n \sup_{-\tau \le t_1 \le t_2 \le \tau} \left\| \sum_{i=m(t(n)+t_1)}^{m(t(n)+t_2)-1} \alpha(i) \left[H(x, Y_{i+1}) - h(x) \right] \right\| = 0 \quad a.s.$$

The proofs for (9.19) and (9.20) follow the same logic and thus are omitted.

Case 2: Let Assumptions 1, 2, 4, and 6′ hold.

By Assumption 4 and the equivalence between norms, we have

$$\|H(x,y)\|_2 \le C \left(\|H(0,y)\|_2 + L(y)\|x\|_2 \right)$$

for some constant C independent of x, y. So for any x,

$$\sup_y \frac{\|H(x,y)\|_2^2}{v(y)} \le \sup_y \frac{2C^2\|H(0,y)\|_2^2 + 2C^2 L(y)^2 \|x\|_2^2}{v(y)} < \infty.$$

In other words, for any x,

$$y \mapsto H(x,y) \in \mathcal{L}^2_{v,\infty}.$$

Similarly, we have for any x,

$$y \mapsto L_b(y) \in \mathcal{L}^2_{v,\infty}.$$

Let g denote any of the following functions:

$$\begin{aligned} y &\mapsto H(x,y) \quad (\forall x), \\ y &\mapsto L_b(y) \quad (\forall x), \\ y &\mapsto L(y). \end{aligned}$$

We now always have $g \in \mathcal{L}^2_{v,\infty}$. Proposition 6 of Borkar et al. (2021) then confirms that

$$\sum_{i=0}^{\infty} \alpha(i)(g(Y_{i+1}) - \mathbb{E}_{y \sim d_y}[g(y)])$$

converges almost surely to a square-integrable random variable. Lemma 22 then follows immediately from the Cauchy convergence test.

□

F.4.2 Proof of Lemma 27

To prove Lemma 27, we first decompose it into three terms. Then, we prove the convergence of each term in Lemmas 67, 68, & 69. Finally, we restate Lemma 27 and connect everything.

For each t, let $\{\Delta_l\}_{l=1}^{\infty}$ be a strictly decreasing sequence of real numbers such that $\lim_{l\to\infty}\Delta_l = 0$ and $\forall l, \frac{t}{\Delta_l} - 1 \in \mathbb{N}$, e.g., $\Delta_l \doteq \frac{t}{l+1}$. Because $\forall l$,

$$\sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) = \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}),$$

we have

$$\lim_{k\to\infty} \left\| \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))ds \right\| \tag{F.65}$$

$$= \lim_{l\to\infty}\lim_{k\to\infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))ds \right\|$$

$$\leq \lim_{l\to\infty}\lim_{k\to\infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l)) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))ds \right\| \tag{F.66}$$

$$+ \lim_{l\to\infty}\lim_{k\to\infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) \left(H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l), Y_{i+1}) \right) \right\|$$

(F.67)

$$+ \lim_{l\to\infty}\lim_{k\to\infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) \left(H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l), Y_{i+1}) - h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l)) \right) \right\|.$$

(F.68)

Now, we show the limit of (F.66), (F.67), and (F.68) are 0 in Lemmas 67, 68, and 69 with proofs in Appendix F.4.3, F.4.4, and F.4.5.

Lemma 67. $\forall j, \forall t \in [0, T)$,

$$\lim_{l\to\infty}\lim_{k\to\infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l)) - \int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s)) \right\| = 0.$$

Lemma 68. $\forall j, \forall t \in [0,T)$,

$$\lim_{l\to\infty}\lim_{k\to\infty}\left\|\sum_{a=0}^{\frac{t}{\Delta_l}-1}\sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1}\alpha(i)\left(H_{r_{n_j}}(\hat{x}(t(i)),Y_{i+1})-H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l),Y_{i+1})\right)\right\|=0.$$

Lemma 69. $\forall j, \forall t \in [0,T)$,

$$\lim_{l\to\infty}\lim_{k\to\infty}\left\|\sum_{a=0}^{\frac{t}{\Delta_l}-1}\sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1}\alpha(i)\left(H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l),Y_{i+1})-h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l))\right)\right\|=0.$$

Plugging Lemmas 67, 68, and 69 back to (F.65) completes the proof of Lemma 27.

F.4.3 Proof of Lemma 67

Proof. $\forall j, \forall t \in [0,T)$,

$$\begin{aligned}
&\lim_{l\to\infty}\lim_{k\to\infty}\sum_{a=0}^{\frac{t}{\Delta_l}-1}\sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1}\alpha(i)h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l))\\
=&\lim_{l\to\infty}\sum_{a=0}^{\frac{t}{\Delta_l}-1}h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l))\lim_{k\to\infty}\sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1}\alpha(i)\\
=&\lim_{l\to\infty}\sum_{a=0}^{\frac{t}{\Delta_l}-1}h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l))\Delta_l && \text{(by (F.38))}\\
=&\int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))ds. && \text{(by definition of integral)}
\end{aligned}$$

Thus, $\forall j, \forall t \in [0,T)$,

$$\begin{aligned}
&\lim_{l\to\infty}\lim_{k\to\infty}\left\|\sum_{a=0}^{\frac{t}{\Delta_l}-1}\sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1}\alpha(i)h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l))-\int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))\right\|\\
=&\left\|\lim_{l\to\infty}\lim_{k\to\infty}\sum_{a=0}^{\frac{t}{\Delta_l}-1}\sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1}\alpha(i)h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l))-\int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))\right\|\\
=&\left\|\int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))-\int_0^t h_{r_{n_j}}(\hat{x}^{\lim}(s))\right\|\\
=&0.
\end{aligned}$$

□

F.4.4 Proof of Lemma 68

Proof. $\forall j, \forall t \in [0,T), \forall l$

$$\begin{aligned}
&\lim_{k\to\infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) \left(H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l), Y_{i+1}) \right) \right\| \\
&\leq \lim_{k\to\infty} \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) \left\| H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l), Y_{i+1}) \right\| \\
&\leq \lim_{k\to\infty} \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) L(Y_{i+1}) \left\| \hat{x}(t(i)) - \hat{x}^{\lim}(a\Delta_l) \right\| \qquad \text{(by Assumption 4)} \\
&\leq \lim_{k\to\infty} \left[\sup_{0\leq a\leq \frac{t}{\Delta_l}-1} \sup_{m(T_{n_k}+a\Delta_l)\leq i\leq m(T_{n_k}+a\Delta_l+\Delta_l)-1} \left\| \hat{x}(t(i)) - \hat{x}^{\lim}(a\Delta_l) \right\| \right] \\
&\quad \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i) L(Y_{i+1}) \\
&= \lim_{k\to\infty} \left[\sup_{0\leq a\leq \frac{t}{\Delta_l}-1} \sup_{m(T_{n_k}+a\Delta_l)\leq i\leq m(T_{n_k}+a\Delta_l+\Delta_l)-1} \left\| \hat{x}(t(i)) - \hat{x}^{\lim}(a\Delta_l) \right\| \right] \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i) L(Y_{i+1}).
\end{aligned} \tag{F.69}$$

We show the limit of the following term.

$$\lim_{k\to\infty}\left[\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{m(T_{n_k}+a\Delta_l)\le i\le m(T_{n_k}+a\Delta_l+\Delta_l)-1}\left\|\hat{x}(t(i))-\hat{x}^{\lim}(a\Delta_l)\right\|\right]$$

$$=\lim_{k\to\infty}\left[\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{t(m(T_{n_k}+a\Delta_l))\le t(i)\le t(m(T_{n_k}+a\Delta_l+\Delta_l)-1)}\left\|\hat{x}(t(i))-\hat{x}^{\lim}(a\Delta_l)\right\|\right]$$

$$\le\lim_{k\to\infty}\left[\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{t(m(T_{n_k}+a\Delta_l))\le \tau\le t(m(T_{n_k}+a\Delta_l+\Delta_l)-1)}\left\|\hat{x}(\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|\right]$$

$$=\lim_{k\to\infty}\left[\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{T_{n_k}+a\Delta_l\le \tau\le t(m(T_{n_k}+a\Delta_l+\Delta_l)-1)}\left\|\hat{x}(\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|\right]$$

($\hat{x}$ is a constant function on interval $[t(m(T_{n_k}+a\Delta_l)), T_{n_k}+a\Delta_l]$ by (9.21) and (9.22))

$$\le\lim_{k\to\infty}\left[\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{T_{n_k}+a\Delta_l\le \tau< T_{n_k}+a\Delta_l+\Delta_l}\left\|\hat{x}(\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|\right] \qquad \text{(by (9.16))}$$

$$=\lim_{k\to\infty}\left[\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{0\le \tau<\Delta_l}\left\|\hat{x}(T_{n_k}+a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|\right]. \qquad \text{(F.70)}$$

By (9.33), $\forall\delta>0$, $\exists k_0$ such that $\forall k\ge k_0$, $\forall t\in[0,T)$,

$$\left\|\hat{x}(T_{n_k}+t)-\hat{x}^{\lim}(t)\right\|\le\delta.$$

$\forall t\in[0,T), \forall l, \forall a$, $\forall k\ge k_0$,

$$\left|\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{0\le\tau<\Delta_l}\left\|\hat{x}(T_{n_k}+a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|-\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{0\le\tau<\Delta_l}\left\|\hat{x}^{\lim}(a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|\right|$$

$$\le\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{0\le\tau<\Delta_l}\left|\left\|\hat{x}(T_{n_k}+a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|-\left\|\hat{x}^{\lim}(a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|\right|$$

(by $|\sup_x f(x)-\sup_x g(x)|\le\sup_x|f(x)-g(x)|$)

$$\le\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{0\le\tau<\Delta_l}\left\|\hat{x}(T_{n_k}+a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)-\hat{x}^{\lim}(a\Delta_l+\tau)+\hat{x}^{\lim}(a\Delta_l)\right\|$$

$$\le\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{0\le\tau<\Delta_l}\left\|\hat{x}(T_{n_k}+a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l+\tau)\right\|$$

$$\le\sup_{0\le a\le \frac{t}{\Delta_l}-1}\ \sup_{0\le\tau<\Delta_l}\delta$$

$$\le\delta.$$

Thus, $\forall t \in [0,T), \forall l, \forall a$,

$$\begin{aligned}
&\lim_{k\to\infty} \sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{0\le\tau<\Delta_l} \left\|\hat{x}(T_{n_k}+a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\| \\
=&\sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{0\le\tau<\Delta_l} \left\|\hat{x}^{\lim}(a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\lim_{k\to\infty} \left[\sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{m(T_{n_k}+a\Delta_l)\le i\le m(T_{n_k}+a\Delta_l+\Delta_l)-1} \left\|\hat{x}(t(i))-\hat{x}^{\lim}(a\Delta_l)\right\|\right] \\
=&\lim_{k\to\infty} \sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{0\le\tau<\Delta_l} \left\|\hat{x}(T_{n_k}+a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\| \qquad \text{(by (F.70))} \\
=&\sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{0\le\tau<\Delta_l} \left\|\hat{x}^{\lim}(a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|. \qquad \text{(F.71)}
\end{aligned}$$

$\forall j, \forall t \in [0,T), \forall l$,

$$\begin{aligned}
&\lim_{k\to\infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l}-1} \sum_{i=m(T_{n_k}+a\Delta_l)}^{m(T_{n_k}+a\Delta_l+\Delta_l)-1} \alpha(i)\left(H_{r_{n_j}}(\hat{x}(t(i)),Y_{i+1})-H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l),Y_{i+1})\right)\right\| \\
\le&\lim_{k\to\infty} \left[\sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{m(T_{n_k}+a\Delta_l)\le i\le m(T_{n_k}+a\Delta_l+\Delta_l)-1} \left\|\hat{x}(t(i))-\hat{x}^{\lim}(a\Delta_l)\right\|\right] \sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i)L(Y_{i+1}) \\
&\qquad \text{(by (F.69))} \\
\le&\lim_{k\to\infty} \left[\sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{m(T_{n_k}+a\Delta_l)\le i\le m(T_{n_k}+a\Delta_l+\Delta_l)-1} \left\|\hat{x}(t(i))-\hat{x}^{\lim}(a\Delta_l)\right\|\right] \\
&\limsup_{k\to\infty} \left[\sum_{i=m(T_{n_k})}^{m(T_{n_k}+t)-1} \alpha(i)L(Y_{i+1})\right] \\
\le&\lim_{k\to\infty} \left[\sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{m(T_{n_k}+a\Delta_l)\le i\le m(T_{n_k}+a\Delta_l+\Delta_l)-1} \left\|\hat{x}(t(i))-\hat{x}^{\lim}(a\Delta_l)\right\|\right] C_H \\
&\qquad \text{(by (F.53))} \\
=&C_H \sup_{0\le a\le \frac{t}{\Delta_l}-1} \sup_{0\le\tau<\Delta_l} \left\|\hat{x}^{\lim}(a\Delta_l+\tau)-\hat{x}^{\lim}(a\Delta_l)\right\|. \qquad \text{(by (F.71))} \qquad \text{(F.72)}
\end{aligned}$$

By Corollary 2, $\hat{x}^{\lim}$ is continuous and $[0,t]$ is a compact set, $\forall \epsilon > 0, \exists \eta$ such that

$$\sup_{0\le|t_1-t_2|\le\eta, t_1\in[0,t], t_2\in[0,t]} \left\|\hat{x}^{\lim}(t_1)-\hat{x}^{\lim}(t_2)\right\| \le \epsilon. \qquad \text{(F.73)}$$

Thus, $\forall \epsilon > 0$, $\exists l_0$ such that $\forall l \geq l_0, \Delta_l \leq \eta$ and we will have

$$0 \leq \sup_{0 \leq a \leq \frac{t}{\Delta_l} - 1} \sup_{0 \leq \tau < \Delta_l} \left\| \hat{x}^{\lim}(a\Delta_l + \tau) - \hat{x}^{\lim}(a\Delta_l) \right\| \leq \epsilon. \qquad \text{(by (F.73))}$$

Therefore, $\forall t$,

$$\lim_{l \to \infty} \left[\sup_{0 \leq a \leq \frac{t}{\Delta_l} - 1} \sup_{0 \leq \tau < \Delta_l} \left\| \hat{x}^{\lim}(a\Delta_l + \tau) - \hat{x}^{\lim}(a\Delta_l) \right\| \right] = 0. \qquad \text{(F.74)}$$

This concludes $\forall j, \forall t \in [0, T)$,

$$\lim_{l \to \infty} \lim_{k \to \infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l} - 1} \sum_{i=m(T_{n_k} + a\Delta_l)}^{m(T_{n_k} + a\Delta_l + \Delta_l) - 1} \alpha(i) \left(H_{r_{n_j}}(\hat{x}(t(i)), Y_{i+1}) - H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l), Y_{i+1}) \right) \right\|$$

$$= \lim_{l \to \infty} C_H \sup_{0 \leq a \leq \frac{t}{\Delta_l} - 1} \sup_{0 \leq \tau < \Delta_l} \left\| \hat{x}^{\lim}(a\Delta_l + \tau) - \hat{x}^{\lim}(a\Delta_l) \right\| \qquad \text{(by (F.72))}$$

$$= C_H \cdot 0 \qquad \text{(by (F.74))}$$

$$= 0.$$

□

F.4.5 Proof of Lemma 69

Proof. By (F.42), $\forall j, \forall a, \forall l$,

$$\lim_{k \to \infty} \left\| \sum_{i=m(T_{n_k} + a\Delta_l)}^{m(T_{n_k} + a\Delta_l + \Delta_l) - 1} \alpha(i) \left[H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l), Y_{i+1}) - h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l)) \right] \right\| = 0. \qquad \text{(F.75)}$$

Thus,$\forall j, \forall t \in [0, T)$,

$$\lim_{l \to \infty} \lim_{k \to \infty} \left\| \sum_{a=0}^{\frac{t}{\Delta_l} - 1} \sum_{i=m(T_{n_k} + a\Delta_l)}^{m(T_{n_k} + a\Delta_l + \Delta_l) - 1} \alpha(i) \left[H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l), Y_{i+1}) - h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l)) \right] \right\|$$

$$\leq \lim_{l \to \infty} \sum_{a=0}^{\frac{t}{\Delta_l} - 1} \lim_{k \to \infty} \left\| \sum_{i=m(T_{n_k} + a\Delta_l)}^{m(T_{n_k} + a\Delta_l + \Delta_l) - 1} \alpha(i) \left[H_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l), Y_{i+1}) - h_{r_{n_j}}(\hat{x}^{\lim}(a\Delta_l)) \right] \right\|$$

$$= \lim_{l \to \infty} \sum_{a=0}^{\frac{t}{\Delta_l} - 1} 0 \qquad \text{(by (F.75))}$$

$$= 0.$$

□