University of California
Santa Barbara

On the Statistical Complexity of Offline Policy Evaluation for Tabular Reinforcement Learning

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics & Applied Probability

by

Ming Yin

Committee in charge:

Professor S. Rao Jammalamadaka, Co-Chair
Professor Yu-Xiang Wang, Co-Chair
Professor Tomoyuki Ichiba

March 2023

The Dissertation of Ming Yin is approved.

Professor Yu-Xiang Wang, Co-Chair

Professor Tomoyuki Ichiba

Professor S. Rao Jammalamadaka, Committee Chair

March 2023

On the Statistical Complexity of Offline Policy Evaluation for Tabular Reinforcement Learning

Ming Yin

Acknowledgements

I would like to offer my humble and grateful acknowledgment to all of the wonderful people I have had the fortune to work with during my Ph.D. career. Much of the impetus for the ideas presented in this dissertation were derived from our work together. In particular, I would like to thank my advisor, Professor Yu-Xiang Wang, for introducing me to the area of reinforcement learning. It is a privilege to work with you, and the experience of working with you definitely becomes the life-changing point of my Ph.D. career. You are a world-class expert and a kind human being, which the latter means even more to me.

I would also like to thank my advisor, Distinguished Professor S. Rao Jammalamadaka, for your tremendous support over the years. The knowledge I gained from your PSTAT 207 course-sequence has been the building blocks for my later research in Statistical Machine Learning. Your counsel, wisdom, and guidance over these years have been a constant and unwavering source of encouragement in my tough time.

I am privileged to have Professor Tomoyuki Ichiba serve on my dissertation committee. You are the first PSTATS faculty I met in person when I was taking the TA oral exam back in the summer of 2016. You are always supportive of me, and having served as your TA for the graduate course PSTAT 213A has been the major teaching experience of my time at UCSB.

Throughout the years, I have had the great pleasure of working with many brilliant researchers, Mengdi Wang, Yu Bai, Yaqi Duan, Dan Qiao, Thanh Nguyen-Tang, Sunil Gupta, Svetha Venkatesh, Raman Arora, Jiachen Li, William Yang Wang, Kaiqi Zhang, Wenjing Chen, Chong Liu, and Qinxun Bai. I cannot achieve what I have done without you.

I was fortunate to spend my graduate student life in the Department of Statistics and Applied Probability at UCSB. I would like to thank all my friends in PSTAT Department. Life would not have been the same without you.

Finally, I would like to thank my parents for their unconditional love throughout my life.

Curriculum Vitæ

Ming Yin

Education

2023, expected	Ph.D. in Computer Science, University of California, Santa Barbara.
2023	Ph.D. in Statistics and Applied Probability, University of California, Santa Barbara.
2018	M.S. in Statistics, University of California, Santa Barbara.
2016	B.S. in Applied Mathematics, University of Science and Technology of China.

Publications

ICLR 2023	Offline Reinforcement Learning with Differentiable Function Approximation is Provably Efficient, Ming Yin, Mengdi Wang, Yu-Xiang Wang. *In Proceedings of the 10th International Conference on Learning Representations, Kigali Rwanda, Africa.*
AAAI 2023	On Instance-Dependent Bounds for Offline Reinforcement Learning with Linear Function Approximation, Thanh Nguyen-Tang, Ming Yin, Sunil Gupta, Svetha Venkatesh, Raman Arora. *In Proceedings of Association for the Advancement of Artificial Intelligence, Washtington, DC, USA.*
NeurIPS WS 2022	Offline Policy Evaluation for Reinforcement Learning with Adaptively Collected Data, Sunil Madhow, Dan Qiao, Ming Yin, Yu-Xiang Wang. *In NeurIPS workshop in Offline RL, New Orleans, LA, USA.*
NeurIPS WS 2022	Offline Reinforcement Learning with Closed-Form Policy Improvement Operators, Jiachen Li, Edwin Zhang, Ming Yin, Qinxun Bai, Yu-Xiang Wang, William Yang Wang. *In Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence, Eindhoven, Netherlands.*
UAI 2022	Offline Stochastic Shortest Path: Learning, Evaluation and Towards Optimality, Ming Yin*, Wenjing Chen*, Mengdi Wang, Yu-Xiang Wang. *In Proceedings of Association for the Advancement of Artificial Intelligence, Washtington, DC, USA.*
ICML 2022	Sample-Efficient Reinforcement Learning with loglog(T) Switching Cost, Dan Qiao, Ming Yin, Ming Min, Yu-Xiang Wang. *In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA.*
ICLR 2022	Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism, Ming Yin, Yaqi Duan, Mengdi Wang, Yu-Xiang Wang. *In Proceedings of the 10th International Conference on Learning Representations, Virtual.*

NeurIPS 2021	Towards Instance-Optimal Offline Reinforcement Learning with Pessimism, Ming Yin, Yu-Xiang Wang. *In Proceedings of the 35th Conference on Neural Information Processing Systems, Vancouver, Canada.*
NeurIPS 2021	Optimal Uniform OPE and Model-based Offline Reinforcement Learning in Time Homogeneous, Reward-Free and Task-Agnostic Settings, Ming Yin, Yu-Xiang Wang. *In Proceedings of the 35th Conference on Neural Information Processing Systems, Vancouver, Canada.*
NeurIPS 2021	Near-Optimal Offline Reinforcement Learning via Double Variance Reduction, Ming Yin, Yu Bai, Yu-Xiang Wang. *In Proceedings of the 35th Conference on Neural Information Processing Systems, Vancouver, Canada.*
AISTATS 2021	Near-Optimal Provable Uniform Convergence in Offline Policy Evaluation for Reinforcement Learning, Ming Yin, Yu Bai, Yu-Xiang Wang **(Oral presentation)** *In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, Virtual.*
AISTATS 2020	Asymptotically Efficient Off-Policy Evaluation for Tabular Reinforcement Learning, Ming Yin, Yu-Xiang Wang. *In Proceedings of the 23th International Conference on Artificial Intelligence and Statistics, Sicily, Italy.*

Academic Services

Area Chair	**[NeurIPS]** Conference on Neural Information Processing Systems, 2023
Conf. Reviewer	**[ICML]** International Conference on Machine Learning, 2020,2021,2022,2023 **[AISTATS]** International Conference on Artificial Intelligence and Statistics, 2021,2022,2023 **[NeurIPS]** Conference on Neural Information Processing Systems, 2021,2022 **[ICLR]** International Conference on Learning Representations, 2022,2023 **[AAAI]** AAAI Conference on Artificial Intelligence, 2023 **[UAI]** Conference on Uncertainty in Artificial Intelligence, 2023
Journal Reviewer	**[JMLR]** Journal of Machine Learning Research **[TMLR]** Transactions on Machine Learning Research **[Ann. Stat.]** Annals of Statistics

Professional Experience

2022	Applied Scientist Intern, Amazon AWS AI, San Jose, CA
2021	Visiting Summer Research Intern, Princeton University, Princeton, NJ

Abstract

On the Statistical Complexity of Offline Policy Evaluation for Tabular Reinforcement Learning

by

Ming Yin

Offline Policy Evaluation (OPE) aims at evaluating the expected cumulative reward of a target policy π when offline data are collected by running a logging policy μ. Standard importance-sampling based approaches for this problem suffer from a variance that scales exponentially with time horizon H, which motivates a splurge of recent interest in alternatives that break the "Curse of Horizon". In the Second chapter of this thesis, we prove the modification of *Marginalized Importance Sampling* (MIS) method can achieve the Cramer-Rao lower bound, provided that the state space and the action space are finite.

In the Third chapter of the thesis, we go beyond the off-policy evaluation setting and propose a new uniform convergence for OPE. The Uniform OPE problem requires evaluating all the policies in a policy class Π simultaneously, and we obtain nearly optimal error bounds for a number of global / local policy classes. Our results imply that the model-based planning achieves an optimal episode complexity of $O(H^3/d_m\epsilon^2)$ in identifying an ϵ-optimal policy under the time-inhomogeneous episodic MDP model. Here d_m is the minimal marginal state-action visitation probability for the current MDP under the behavior policy μ. We further improve the sample complexity guarantee to $O(H^2/d_m\epsilon^2)$ under the time-homogeneous episodic MDPs, using a novel singleton-absorbing MDP technique in the Fourth chapter. Both results are known to be optimal under their respective settings. In the final part of the thesis, we summarise our work in reinforcement learning and conclude with potential future directions.

List of Frequently Used Notations

OPE	Offline Policy Evaluation
MDP	Markov Decision Processes
TMIS	Tabular Marginalized Importance Sampling
μ	Logging/Behavior policy
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
S	Number of states
A	Number of actions
H	Horizon
$d_t^{\mu}(s,a)$	$\mathbb{P}^{\pi}(s_t = s, a_t = a \vert \mu)$
d_m	$\min_{t,s,a} d_t^{\mu}(s,a)$
$Q_t^{\pi}(s,a)$	$\mathbb{E}_{\pi}[\sum_{t'=t}^{H} r_{t'} \vert s_t = s, a_t = a]$
$V_t^{\pi}(s)$	$\mathbb{E}_{\pi}[\sum_{t'=t}^{H} r_{t'} \vert s_t = s]$
v^{π}	$\mathbb{E}_{\pi}[\sum_{t=1}^{H} r_t]$

Contents

Chapter 1

Introduction

In *offline Reinforcement Learning* (offline RL [1, 2]), the goal is to learn a reward-maximizing policy in an unknown environment which forms a *Markov Decision Process* (MDP), using the historical data coming from a (fixed) behavior policy μ. Unlike an online RL, where the agent can keep interacting with the environment and gain new feedback by exploring unvisited state-action space, offline RL usually is needed when such online interplays are expensive or even unethical. Due to its nature of having no access to interact with the MDP model (which causes distributional mismatches), most of the literature that studies the sample complexity / provable efficiency of offline RL (*e.g.* [3, 4, 5, 6, 7, 8, 9, 10, 11]) relies on making different data-coverage assumptions for making the problem learnable, and provide near-optimal worst-case performance bounds that depend on their data-coverage coefficients.

Offline Policy evaluation (OPE), which predicts the performance of a policy with data only sampled by a logging/behavior policy [12], plays a key role for using reinforcement learning (RL) algorithms responsibly in many real-world decision-making problems such as marketing, finance, robotics, and healthcare. Deploying a policy without having an accurate evaluation of its performance could be costly, illegal, and can even break down the machine learning system. There is a large body of literature that studied the off-policy evaluation problem in both

theoretical and application-oriented aspects. From the theoretical perspective, OPE problem is extensively studied in contextual bandits [13, 14, 15, 16] and reinforcement learning (RL) [17, 18, 19, 20, 21] and the results of OPE studies have been applied to real-world applications including marketing [22, 23] and education [24].

In this thesis, we provide non-asymptotic analysis of the point OPE estimators, explaining how the statistical error is characterized by the sample size, distributional shift, planning horizon, and its connections to the policy optimization problems via the uniform convergence. Our contribution can be summarized as follows:

- In Chapter 2, we consider the problem of off-policy evaluation for a finite horizon, non-stationary, episodic MDP under tabular MDP setting. We propose and analyze Tabular-MIS estimator, which closes the gap between Cramer-Rao lower bound provided by [18] and the MSE upper bound of State-MIS estimator [21]. We also provide a high probability result by introducing a data-splitting type Tabular-MIS estimator, which retains the asymptotic efficiency while having an exponential tail. Moreover, the calculation of Tabular-MIS estimator and Split-TMIS does not explicitly incorporate the importance weights, which in turn implies our off-policy evaluation algorithm can be implemented without needing to know logging probabilities μ. Such logging-policy-free feature makes our Tabular-MIS estimator estimator more practical in the real-world applications. Finally, we conduct a numerical simulation to empirically validate our theoretical results. We see that Tabular-MIS estimator improves over State-MIS estimator in MSE by a factor of H, as expected.

- In Chapter 3, we represent the first systematic study of uniform convergence in offline policy evaluation. For the global policy class (deterministic or stochastic), we use fully model-based OPEMA estimator to obtain an ϵ-uniform OPE with episode complexity $\widetilde{O}(H^4 S/d_m\epsilon^2)$ (Theorem 3.5.1) and in some cases this can be reduced to $\widetilde{O}(H^4/d_m\epsilon^2)$,

where d_m is minimal marginal state-action occupancy probability depending on logging policy μ. For the global deterministic policy class, we obtain an ϵ-uniform OPE with episode complexity $\widetilde{O}(H^3S/d_m\epsilon^2)$ with an optimal dependence on H (Theorem 3.5.2). For a (data-dependent) local policy class that cover all policies are in the $O(\sqrt{H}/S)$-neighborhood of the *empirical* optimal policy (see the definition in Section 4.4), we obtain ϵ-uniform OPE with $\widetilde{O}(H^3/d_m\epsilon^2)$ episodes (Theorem 3.5.3). Our uniform OPE over the local policy class implies that ERM (VI or PI with empirically estimated MDP), as well as any sufficiently accurate model-based planning algorithm, has an optimal episode complexity of $\widetilde{O}(H^3/d_m\epsilon^2)$ (Theorem 3.6.1). To the best of our knowledge, this is the first rate-optimal algorithm in the offline RL setting.

- In Chapter 4, we study the uniform convergence problems for offline policy evaluation (OPE) and provide complete answers for their optimality behavior. We derive the $\tilde{O}(H^2/d_m\epsilon^2)$ optimal episode complexity for local uniform OPE (Theorem 4.6.1) via the model-based method and this implies optimal offline learning with the same rate. We characterize the statistical limit for the global uniform convergence by proving a minimax lower bound $\Omega(H^2S/d_m\epsilon^2)$ (over all model-based approaches) (Theorem 4.5.1). This result answers the question left by [7] that the global uniform OPE is generically harder than the local uniform OPE / offline learning by a factor of S, such a difference will dominate when the state space is exponentially large. Critically, our model-based frameworks naturally generalize to the more challenging settings like task-agnostic and reward-free settings. In particular, we establish the $\tilde{O}(H^2\log(K)/d_m\epsilon^2)$ (Theorem 4.7.1) and $\tilde{O}(H^2S/d_m\epsilon^2)$ (Theorem 4.7.2) complexities for *offline task-agnostic learning* and *offline reward-free learning*.

We conclude the thesis by summarizing our work and mentioning possible future research directions in Chapter 5.

Chapter 2

Offline Policy Evaluation for Tabular Reinforcement Learning

In this chapter, we focus on offline policy evaluation (OPE), a fundamental problem in Reinforcement Learning (RL). OPE is concerned with estimating the mean cumulative reward of a given decision policy, known as the target/evaluation policy, using historical data generated by a potentially different policy, known as the behavior/logging policy. OPE is most crucial for offline RL, where we only have access to a historical dataset and are not allowed to explore the environment.

2.1 Offline Policy Evaluation Setup

In the reinforcement learning problem the agent interacts with an underlying unknown dynamic which is modeled as a Markov decision process (MDP). An MDP is defined by the quantities $M = (\mathcal{S}, \mathcal{A}, r, P, d_1, H)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel with $P_t(s'|s, a)$ representing the probability of seeing state s' after taking action a at state s, $r_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the mean reward function with $r_t(s, a)$ being the

average immediate reward of (s, a) at time t, d_1 denotes the initial state distribution, and H is the time horizon. The subscript t in P_t indicates that the transition dynamics are non-stationary and could be different at each time t. A (non-stationary) policy $\pi : \mathcal{S} \to \mathbb{P}_{\mathcal{A}}^{H}$[1] assigns each state $s_t \in \mathcal{S}$ a distribution over actions at each time t, *i.e.* $\pi_t(\cdot|s_t)$ is a probability simplex with dimension $|\mathcal{S}|$.

Given a target policy of interest π, then the distribution of one H-step trajectory $\tau = (s_1, a_1, r_1, ..., s_H, a_H, r_H, s_{H+1})$ is specified by $\pi := (d_1, \pi)$[2] as follows: $s_1 \sim d_1^{\pi}$, for $t = 1, ..., H$, $a_t \sim \pi_t(\cdot|s_t)$ and random reward r_t has mean $r_t(s_t, a_t)$. Then value function under policy π is defined as:

$$v^{\pi} = \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} r_t\right].$$

The OPE problem aims at estimating v^{π} while given that n episodic data[3] $\mathcal{D} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\right\}_{i\in[n]}^{t\in[H]}$ are actually coming from a different logging policy μ.

2.2 Traditional Methods for OPE and Related Settings

The classical way to tackle the problem of OPE relies on incorporating importance sampling weights (IS), which corrects the mismatch in the distributions under the behavior policy and target policy. Specifically, define the t-step importance ratio as $\rho_t := \pi_t(a_t|s_t)/\mu_t(a_t|s_t)$, then it

[1] Here $\mathbb{P}_{\mathcal{A}}^{H} = \mathbb{P}_{\mathcal{A}} \times \mathbb{P}_{\mathcal{A}} \times \mathbb{P}_{\mathcal{A}} \times ... \times \mathbb{P}_{\mathcal{A}}$, where "×" represents Cartesian product and the product is performed for H times.

[2] For brevity, $\forall\pi$ we use π to denote the pair (d_1, π). This can be understood as: $\forall\pi,\ d_1^{\pi} = d_1$.

[3] To distinguish the data from different episodes, we use superscript to denote which episode they belong to throughout the rest of the work.

uses the cumulative importance ratio $\rho_{1:t} := \prod_{t'=1}^{t} \rho_{t'}$ to create IS based estimators:

$$\widehat{V}_{\text{IS}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{V}_{\text{IS}}^{(i)}, \quad \widehat{V}_{\text{IS}}^{(i)} := \rho_{1:H}^{(i)} \cdot \sum_{t=1}^{H} r_t^{(i)};$$
$$\widehat{V}_{\text{step-IS}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{V}_{\text{step-IS}}^{(i)}, \quad \widehat{V}_{\text{step-IS}}^{(i)} := \sum_{t=1}^{H} \rho_{1:t}^{(i)} r_t^{(i)},$$

where $\rho_{1:t}^{(i)} = \prod_{t'=1}^{t} \pi_{t'}(a_{t'}^{(i)}|s_{t'}^{(i)})/\mu_{t'}(a_{t'}^{(i)}|s_{t'}^{(i)})$. There are different versions of IS estimators including weighted IS estimators and doubly robust estimators [25, 26, 14, 18].

Even though IS-based off-policy evaluation methods possess a lot of advantages (*e.g.* unbiasedness), the variance of the cumulative importance ratios $\rho_{1:t}$ may grow exponentially as the horizon goes long. Attempts to break the barriers of horizon have been tried using model-based approaches [27, 28], which builds the whole MDP using either parametric or nonparametric models for estimating the value of target policy. [29] considers breaking the curse of horizon of time-invariant MDPs by deploying importance sampling on the average visitation distribution of state-action pairs, [30] considers leveraging the stationary ratio of state-action pairs to replace the trajectory weights in an online fashion and [31] further applies the same idea in the deep reinforcement learning regime. Recently, [32, 33] propose double reinforcement learning (DRL), which is based on doubly robust estimator with cross-fold estimation of q-functions and marginalized density ratios. It was shown that DRL is asymptotically efficient when both components are estimated at fourth-root rates, however no finite sample error bounds are given.

Markov Decision Processes have a long history of associated research [34, 35], but many theoretical problems in the basic tabular setting remain an active area of research as of today. In particular, other than off-policy setting, there are two types of questions: *Regret bound and sample complexity in the online setting* and *Sample complexity with a generative model*. A detailed discussion can be found in Section A.1 in appendix.

The OPE setting is different in two ways compared to those mentioned above. First, we

consider a fixed pair of logging and target policy μ and π, so our bounds can depend explicitly on π and μ instead of S, A. Second, we do not have either online access to the environment (to change policies) or a generative model. Our high-probability bound with a direct union bound argument, implies a sample complexity of $\widetilde{O}(H^3S^2A/\epsilon^2)$ for identifying the optimal policy, which is suboptimal up to a factor of S, but notably has the optimal dependence in H. We remark that achieving the optimal dependence in the planning horizon H is generally tricky (see, e.g., the COLT open problem [36] for more details). The current thesis is among the few instances where we know how to obtain the optimal parameters.

2.3 Our Goal in Tabular OPE and Assumptions

In this chapter, our goal is to obtain the optimality of IS-based methods through marginalized importance sampling (MIS). In an earlier attempt, [21] constructs MIS estimator by aggregating all trajectories that share the same state transition patterns to directly estimate the state distribution shifts after the change of policies from the behavioral to the target. However, as pointed in Remark 4 in [21], the MSE upper bound of MIS estimator is asymptotically inefficient by a multiplicative factor of H. [21] conjectures that the lower bound is not achievable in their infinite action setting. To bridge the gap and ultimately achieve the optimality, we consider the Tabular MDPs, where both the state space and action space are finite (*i.e.* $S = |\mathcal{S}| < \infty, A = |\mathcal{A}| < \infty$) and each state-action pair can be visited frequently as long as the logging policy μ does sufficient exploration (which corresponds to Assumption 4.4.1). Under the Tabular MDP setting, we can show the MSE upper bound of MIS estimator matches the Cramer-Rao lower bound provided by [18]. To distinguish the difference, throughout the rest of paper we call the modified MIS estimator Tabular-MIS (TMIS) and the MIS estimator in [21] State-MIS (SMIS).

2.3.1 Notions, Objective and Assumptions

In addition to the non-stationary, finite horizon tabular MDP $M = (S, \mathcal{A}, r, T, d_1, H)$ (where $S := |S| < \infty$ and $A := |\mathcal{A}| < \infty$), non-stationary logging policy μ and target policy π, we denote $d_t^\mu(s_t, a_t)$ and $d_t^\pi(s_t, a_t)$ the induced joint state-action distribution at time t and the state distribution counterparts $d_t^\mu(s_t)$ and $d_t^\pi(s_t)$, satisfying $d_t^\pi(s_t, a_t) = d_t^\pi(s_t) \cdot \pi(a_t|s_t)$.[4] The initial distributions are identical $d_1^\mu = d_1^\pi = d_1$. Moreover, we use $P_{i,j}^\pi \in \mathbb{R}^{S \times S},\ \forall j < i$ to represent the state transition probability from step j to step i under policy π, where $P_{t+1,t}^\pi(s'|s) = \sum_a P_{t+1,t}(s'|s, a)\pi_t(a|s)$. The marginal state distribution vector $d_t^\pi(\cdot)$ satisfies $d_t^\pi = P_{t,t-1}^\pi d_{t-1}^\pi$.

Historical data $\mathcal{D} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\right\}_{i\in[n]}^{t\in[H]}$ was obtained by logging policy μ and we can only use $\mathcal{D}$ to estimate the value of target policy π, *i.e.* v^π. Suppose we only assume knowledge about π and *do not observe* $r_t(s_t, a_t)$ for any actions other than the noisy immediate reward $r_t^{(i)}$ after observing $s_t^{(i)}, a_t^{(i)}$. The goal is to find an estimator which minimizes the mean-square error (MSE), namely:

$$\mathrm{MSE}(\pi, \mu, M) = \mathbb{E}_\mu[(\widehat{v}^\pi - v^\pi)^2].$$

Assumption 2.3.1 (Bounded rewards)**.** $\forall\ t = 1, ..., H$ *and* $i = 1, ..., n$, $0 \le r_t^{(i)} \le R_{\max}$.

The bounded reward assumption can be relaxed to: $\exists R_{\max}, \sigma < +\infty$ such that $0 \le \mathbb{E}[r_t|s_t, a_t, s_{t+1}] \le R_{\max}$, $\mathrm{Var}[r_t|s_t, a_t, s_{t+1}] \le \sigma^2$ (as in [21]), for achieving Cramer-Rao lower bound. However, the boundedness will become essential for applying concentrate inequalities in deriving high probability bounds.

Assumption 2.3.2 (Sufficient exploration)**.** *Logging policy* μ *obeys that* $d_m := \min_{t,s_t} d_t^\mu(s_t) > 0$.

In fact this assumption can be relaxed to: require $d_t^\mu(s_t) > 0$ whenever $d_t^\pi(s_t) > 0$, and the corresponding $d_m := \min_{t,s_t}\{d_t^\mu(s_t) : d_t^\mu(s_t) > 0\}$. However, for the illustration purpose

[4] For μ, $d_t^\mu(s_t, a_t) = d_t^\mu(s_t) \cdot \mu(a_t|s_t)$.

we stick to the above assumption. This assumption is always required for the consistency of off-policy evaluation estimator.

Assumption 2.3.3 (Bounded weights)**.** $\tau_s := \max_{t,s_t} \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)} < +\infty$ *and* $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)} < +\infty$.

Assumption 2.3.3 is also necessary for discrete state and actions, as otherwise the second moments of the importance weight would be unbounded and the MSE of estimators will become intractable . The bound on τ_s is natural since $\tau_s \leq \max_{t,s_t} \frac{1}{d_t^\mu(s_t)} = \frac{1}{\min_{t,s_t} d_t^\mu(s_t)} = \frac{1}{d_m}$ and it is finite by the Assumption 4.4.1; similarly, $\tau_a < \infty$ is also automatically satisfied if $\min_{t,s_t,a_t} \mu(a_t|s_t) > 0$. Finally, as we will see in the results, explicit dependence on τ_s, τ_a and d_m only appear in the low-order terms of the error bound.

2.4 Tabular-MIS estimator

To overcome the barrier caused by cumulative importance weights in IS type estimators, marginalized importance sampling directly estimates the marginalized state visitation distribution $\widehat{d}_t$ and defines the MIS estimator:

$$\widehat{v}^\pi_{MIS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{H} \frac{\widehat{d}_t^\pi(s_t^{(i)})}{\widehat{d}_t^\mu(s_t^{(i)})} \widehat{r}_t^\pi(s^{(i)}). \tag{2.1}$$

and $\widehat{d}_t^\mu(\cdot)$ is directly estimated using the empirical mean, *i.e.* $\widehat{d}_t^\mu(s_t) := \frac{1}{n} \sum_i \mathbf{1}(s_t^{(i)} = s_t) := \frac{n_{s_t}}{n}$ whenever $n_{s_t} > 0$ and $\widehat{d}_t^\pi(s_t)/\widehat{d}_t^\mu(s_t) = 0$ when $n_{s_t} = 0$. Then the MIS estimator (2.1) becomes:

$$\widehat{v}^\pi_{MIS} = \sum_{t=1}^{H} \sum_{s_t} \widehat{d}_t^\pi(s_t) \widehat{r}_t^\pi(s_t) \tag{2.2}$$

Construction of State-MIS estimator. Based on the estimated marginal state transition $\widehat{d}_t^\pi = \widehat{P}_t^\pi \widehat{d}_{t-1}^\pi$, State-MIS estimator in [21] directly estimates the state transition $P_t^\pi(s_t|s_{t-1})$ and state

reward $r_t^\pi(s_t)$ as:

$$\widehat{P}_t^\pi(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}} \sum_{i=1}^{n} \frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})} \tag{2.3}$$

$$\cdot \mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}, a_t^{(i)}) = (s_{t-1}, s_t, a_t)); \tag{2.4}$$

$$\widehat{r}_t^\pi(s_t) = \frac{1}{n_{s_t}} \sum_{i=1}^{n} \frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)} r_t^{(i)} \cdot \mathbf{1}(s_t^{(i)} = s_t). \tag{2.5}$$

State-MIS estimator directly constructs state transitions $\widehat{P}_t^\pi(s_t|s_{t-1})$ without explicitly modeling actions. Therefore, it is still valid when action space $\mathcal{A}$ is unbounded. However, importance weights must be explicitly utilized for compensating the discrepancy between μ and π and the knowledge of $\mu(a|s)$ at each state-action pair (s, a) is required.

Construction of Tabular-MIS estimator. Alternatively, we can go beyond importance weights and construct empirical estimates for $\widehat{P}_{t+1}(s_{t+1}|s_t, a_t)$ and $\widehat{r}_t(s_t, a_t)$ as:

$$\begin{aligned} \widehat{P}_{t+1}(s_{t+1}|s_t, a_t) &= \frac{\sum_{i=1}^{n} \mathbf{1}[(s_{t+1}^{(i)}, a_t^{(i)}, s_t^{(i)}) = (s_{t+1}, s_t, a_t)]}{n_{s_t, a_t}} \\ \widehat{r}_t(s_t, a_t) &= \frac{\sum_{i=1}^{n} r_t^{(i)} \mathbf{1}[(s_t^{(i)}, a_t^{(i)}) = (s_t, a_t)]}{n_{s_t, a_t}}, \end{aligned} \tag{2.6}$$

where we set $\widehat{P}_{t+1}(s_{t+1}|s_t, a_t) = 0$ and $\widehat{r}_t(s_t, a_t) = 0$ if $n_{s_t, a_t} = 0$, with n_{s_t, a_t} the empirical visitation frequency to state-action (s_t, a_t) at time t. The corresponding estimation of $\widehat{P}_t^\pi(s_t|s_{t-1})$ and $\widehat{r}_t^\pi(s_t)$ are defined as:

$$\begin{aligned} \widehat{P}_t^\pi(s_t|s_{t-1}) &= \sum_{a_{t-1}} \widehat{P}_t(s_t|s_{t-1}, a_{t-1}) \pi(a_{t-1}|s_{t-1}), \\ \widehat{r}_t^\pi(s_t) &= \sum_{a_t} \widehat{r}_t(s_t, a_t) \pi(a_t|s_t), \ \widehat{d}_t^\pi = \widehat{P}_t^\pi \widehat{d}_{t-1}^\pi. \end{aligned} \tag{2.7}$$

In conclusion, by using the same estimator for $\widehat{d}_t^\mu$, $\widehat{v}_{\text{TMIS}}^\pi$ and $\widehat{v}_{\text{SMIS}}^\pi$ share the same form of (2.2). However, Tabular-MIS estimator constructs a different estimation of component $\widehat{d}_t^\pi$ though (2.6)-(2.7) by leveraging the fact that each state-action pair is visited frequently under

Tabular setting.

The motivation of MIS-type estimators comes from the fact that we have a nonstationary MDP model and its underlying state marginal transition follows $d_t^\pi = P_t^\pi d_{t-1}^\pi$. The MIS estimators are then obtained by using corresponding plug-in estimators for each different components (*i.e.* $\widehat{d}_t^\pi$ for d_t^π, $\widehat{P}_t^\pi$ for P_t^π). On the other hand, IS-type estimators design the value function in a more straightforward way without needing to estimate the transition environment [37]. Therefore in this sense MIS-type estimators are essentially model-based estimators with the model of interactive environment $M = (S, \mathcal{A}, r, T, d_1, H)$.

2.5 Mean-Square Error Bound for TMIS

We now show that our Tabular-MIS estimator achieves the asymptotic Cramer-Rao lower bound for DAG-MDP [18] and therefore is asymptotically sample efficient. To formalize our statement, we pre-specify the following boundary conditions: $r_0(s_0) \equiv 0$, $\sigma_0(s_0, a_0) \equiv 0$, $\frac{d_0^\pi(s_0)}{d_0^\mu(s_0)} \equiv 1$, $\frac{\pi(a_0|s_0)}{\mu(a_0|s_0)} \equiv 1$, $V_{H+1}^\pi \equiv 0$, and, as a reminder, $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$ and $\tau_s := \max_{t,s_t} \frac{d_t^\pi(s_t)}{d_t^\mu(s_t)}$.

Theorem 2.5.1. *Suppose the n episodic historical data* $\mathcal{D} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\right\}_{i=1,\dots,n}^{t=1,\dots,H}$ *is obtained by running a logging policy* μ *and* π *is the new target policy which we want to test. If the number of episodes n satisfies*

$$n > \max\left[\frac{16\log n}{\min_{t,s_t,a_t} d_t^\mu(s_t,a_t)}, \frac{4H\tau_a\tau_s}{\min_{t,s_t}\max\{d_t^\pi(s_t), d_t^\mu(s_t)\}}\right]$$

then under Assumption 3.3.1-2.3.3 our Tabular-MIS estimator $\widehat{v}_{\text{TMIS}}^\pi$ *has the following Mean-*

Square-Error upper bound:

$$\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi})^2] \leq \frac{1}{n}\sum_{h=0}^{H}\sum_{s_h,a_h}\frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\cdot \mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\Big| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right]$$
$$\cdot\left(1 + \sqrt{\frac{16\log n}{n\min_{t,s_t} d_t^{\mu}(s_t)}}\right) + O(\frac{\tau_a^2\tau_s H^3}{n^2\cdot d_m}), \tag{2.8}$$

where the value function is defined as: $V_h^{\pi}(s_h) := \mathbb{E}_{\pi}\left[\sum_{t=h}^{H} r_t^{(1)}\Big| s_h^{(1)} = s_h\right],\ \forall h \in \{1,2,...,H\}.$

The proof of this theorem and related technical results that are presented in this section, are deferred to the Appendix. We summarize the novel ingredients in the proof in Section 2.6. Before that, we make a few remarks about this interesting result.

Remark 1 (Asymptotic efficiency and local minimaxity)**.** *The error bound implies that* $\lim_{n\to\infty} n\cdot\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi})^2]$

$$\sum_{t=0}^{H}\mathbb{E}_{\mu}\left[\frac{d^{\pi}(s_t^{(1)},a_t^{(1)})^2}{d^{\mu}(s_t^{(1)},a_t^{(1)})^2}\mathrm{Var}\left[V_{t+1}^{\pi}(s_{t+1}^{(1)}) + r_t^{(1)}\Big| s_t^{(1)}, a_t^{(1)}\right]\right].$$

This exactly matches the CR-lower bound in [18, Proposition 3] for DAG-MDP[5]*. In contrast, the State-MIS estimator in [21] achieves an asymptotic MSE of*

$$\sum_{t=0}^{H}\mathbb{E}_{\mu}\left[\frac{d^{\pi}(s_t^{(1)})^2}{d^{\mu}(s_t^{(1)})^2}\mathrm{Var}\left[\frac{\pi(a_t^{(1)}|s_t^{(1)})}{\mu(a_t^{(1)}|s_t^{(1)})}(V_{t+1}^{\pi}(s_{t+1}^{(1)}) + r_t^{(1)})\Big| s_t^{(1)}\right]\right]. \tag{2.9}$$

We note that while in classical literature CR-lower bound is often used as the lower bound for the variance of *unbiased* estimators, the modern theory of estimation establishes that it is also the correct asymptotic minimax lower bound for the MSE of *all* estimators in every local neighborhood of the parameter space [38, Chapter 8]. In other words, our results imply that

[5] [18] focused on the special case with deterministic reward only at $t = H$. It is straightforward to show that the above expression is the CR-lower bound in the general tabular setting.

Tabular-MIS estimator is asymptotically, locally, uniformly minimax optimal, namely, optimal for every problem instance separately.

Algorithm 1 Tabular MIS Off-Policy Evaluation

Input: Logging data $\mathcal{D} = \{\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{t=1}^H\}_{i=1}^n$ from the behavior policy μ. A target policy π which we want to evaluate its cumulative reward.

1: Calculate the on-policy estimation of initial distribution $d_1(\cdot)$ by $\widehat{d}_1(s) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(s_1^{(i)} = s)$, and set $\widehat{d}_1^{\mu}(\cdot) := \widehat{d}_1(\cdot)$, $\widehat{d}_1^{\pi}(s) := \widehat{d}_1(\cdot)$.

2: **for** $t = 2, 3, \ldots, H$ **do**

3: Choose all transition data at time step t, $\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{i=1}^n$.

4: Calculate the on-policy estimation of $d_t^{\mu}(\cdot)$ by $\widehat{d}_t^{\mu}(s) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(s_t^{(i)} = s)$.

5: Set the off-policy estimation of $\widehat{P}_t(s_t|s_{t-1}, a_{t-1})$:

$$\begin{aligned}&\widehat{P}_t(s_t|s_{t-1}, a_{t-1})\\ :=&\frac{\sum_{i=1}^n \mathbf{1}[(s_t^{(i)}, a_{t-1}^{(i)}, s_{t-1}^{(i)}) = (s_t, s_{t-1}, a_{t-1})]}{n_{s_{t-1},a_{t-1}}}\end{aligned}$$

when $n_{s_{t-1},a_{t-1}} > 0$. Otherwise set it to be zero.

6: Estimate the reward function

$$\widehat{r}_t(s_t, a_t) := \frac{\sum_{i=1}^n r_t^{(i)} \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}{\sum_{i=1}^n \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}.$$

when $n_{s_t,a_t} > 0$. Otherwise set it to be zero.

7: Set $\widehat{d}_t^{\pi}(\cdot)$ according to $\widehat{d}_t^{\pi} = \widehat{P}_t^{\pi}\widehat{d}_{t-1}^{\pi}$, with $\widehat{P}_t^{\pi}$ defined according to (2.7). Also, set $\widehat{r}_t^{\pi}(\cdot)$ according to (2.7).

8: **end for**

9: Substitute all the estimated values above into (2.1) to obtain $\widehat{v}^{\pi}$, the estimated value of π.

While asymptotically efficient estimators for this problem in related settings have been pro-

posed in independent recent work [32, 33], our estimator is the first that comes with finite sample guarantees with an explicit expression on the low-order terms. Moreover, our estimator demonstrates that doubly robust estimation techniques is not essential for achieving asymptotic efficiency.

Remark 2 (Simplified finite sample error bound)**.** *The theory implies that there is universal constants C_1, C_2 such that for all $n \geq C_1 H \frac{\tau_a}{d_m}$, i.e., when we have a just visited every state-action pair for $\Omega(H)$ times, $\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi})^2] = C_2 H^2 \tau_a \tau_s R^2_{\max}/n$.*

In deriving the above remark, we used the somewhat surprising observation that

$$\sum_{t=1}^{H} \mathbb{E}_{\pi}\left[\mathrm{Var}\left[V^{\pi}_{t+1}(s^{(1)}_{t+1}) + r^{(1)}_t \middle| s^{(1)}_t, a^{(1)}_t\right]\right] \leq H^2 R^2_{\max}.$$

Note that we are summing H quantities that are potentially on the order of $H^2 R^2_{\max}$, yet no additional factors of H shows up. This observation is folklore and has been used in deriving tight results for tabular RL in e.g. [39]. It can be proven using the following decomposition of the variance of the empirical mean estimator and the fact it is bounded by $H^2 R^2_{\max}$.

Lemma 2.5.1. *For any policy π and any MDP.*

$$\begin{aligned}\mathrm{Var}_{\pi}\left[\sum_{t=1}^{H} r^{(1)}_t\right] = \sum_{t=1}^{H}\Big(&\mathbb{E}_{\pi}\left[\mathrm{Var}\left[r^{(1)}_t + V^{\pi}_{t+1}(s^{(1)}_{t+1})\middle| s^{(1)}_t, a^{(1)}_t\right]\right] \\ &+ \mathbb{E}_{\pi}\left[\mathrm{Var}\left[\mathbb{E}[r^{(1)}_t + V^{\pi}_{t+1}(s^{(1)}_{t+1})| s^{(1)}_t, a^{(1)}_t]\middle| s^{(1)}_t\right]\right]\Big).\end{aligned}$$

The proof, which applies the law-of-total-variance recursively, is deferred to the appendix.

Remark 3 (When $\pi = \mu$)**.** *One surprising observation is that Tabular-MIS estimator improves the efficiency even for the on-policy evaluation problem when $\pi = \mu$. In other words, the natural Monte Carlo estimator of the reward in the on-policy evaluation problem is in fact asymptotically inefficient.*

2.6 Outline of Proof of Theorem 2.5.1

At a higher level the techniques we used include the idea of fictitious estimator and peeling the variance (expectation) of fictitious estimator $\widetilde{v}^{\pi}$ from behind by applying the total law of variances (expectations) repeatedly, as in [21].

In addition to the above techniques, we leverage the fact of frequent state-action visitations in our design of TMIS estimator and based on that we are able to achieve an asymptotic lower Mean Square Error (MSE) bound. The main components are the following.

Fictitious Tabular-MIS estimator. Fictitious Tabular-MIS estimator $\widetilde{v}^{\pi}_{\text{TMIS}}$ is a modified version of $\widehat{v}^{\pi}_{\text{TMIS}}$ with $\widehat{P}^{\pi}_{t+1}(\cdot|s_t, a_t)$, $\widehat{r}^{\pi}_t(s_t, a_t)$ replaced by the underlying true $P^{\pi}_{t+1}(\cdot|s_t, a_t)$, $r^{\pi}_t(s_t, a_t)$ when the visitation frequency of state-action pair (s_t, a_t) is insufficient (*e.g.* $n_{s_t,a_t} < O(nd^{\mu}_t(s_t, a_t))$). Specifically, fictitious Tabular-MIS estimator $\widetilde{v}^{\pi}_{\text{TMIS}}$ remains every part of $\widehat{v}^{\pi}_{\text{TMIS}}$ unchanged except the following:

$$\widetilde{r}_t(s_t, a_t) = \begin{cases} \widehat{r}_t(s_t, a_t) & \text{if } n_{s_t,a_t} \geq nd^{\mu}_t(s_t, a_t)(1-\theta) \\ r_t(s_t, a_t) & \text{otherwise;} \end{cases} \tag{2.10}$$

and

$$\widetilde{P}_{t+1,t}(\cdot|s_t, a_t) = \begin{cases} \widehat{P}_{t+1,t} & \text{if } n_{s_t,a_t} \geq nd^{\mu}_t(s_t, a_t)(1-\theta) \\ P_{t+1,t} & \text{otherwise,} \end{cases} \tag{2.11}$$

where θ is the parameter constrained by $0 < \theta < 1$, which we will choose later in the proof.

This slight modification makes $\widetilde{v}^{\pi}_{\text{TMIS}}$ no longer implementable using the logging data $\mathcal{D}$, but it does provide an unbiased estimator of v^{π} (Lemma A.2.5 in appendix) and, most importantly, it is easier to do theoretical analysis on $\widetilde{v}^{\pi}_{\text{TMIS}}$ than on $\widehat{v}^{\pi}_{\text{TMIS}}$. Moreover, Multiplicative Chernoff bound helps to find the connection between $\widetilde{v}^{\pi}_{\text{TMIS}}$ and $\widehat{v}^{\pi}_{\text{TMIS}}$.

Peeling arguments using the total law of variance (expectation). The core idea in analyzing

the variance of $\widetilde{v}^{\pi}$ is to peel the variance from behind (start from time H to 1) and the peeling tool we used here is through marrying the standard Bellman equations with the total law of variance. Lemma A.2.2 (in appendix) shows this spirit and it is used repeatedly throughout the whole analysis. Beyond that, the peeling argument can be used to prove the dependence in H is only H^2 for our Tabular-MIS estimator. This result explicates that H^2 is enough for TMIS to evaluate a particular policy and this is different from SMIS, which in general requires the dependence of H^3 for off-policy evaluation.

2.7 A High-Probability Bound with Data-Splitting TMIS.

Tabular-MIS estimator provides the asymptotic optimal variance bound of order $O(H^2SA/n)$ and based on that it is natural to ask the related learning question: whether TMIS can further achieve a high probability bound with the same sample complexity? We figure out that the standard concentration inequalities (*e.g.* Hoeffding's inequality, Bernstein inequality) cannot be directly applied because of the highly correlated structures of the Tabular-MIS estimator. To address this problem we design the following data split version of TMIS and as we will see, the original TMIS is essentially a special case of data-splitting TMIS.

Data splitting Tabular-MIS estimator. Assume the total number of episodes n can be factorized as $n = M \cdot N$, where $M, N > 1$ are two integers,[6] and we can partition the data $\mathcal{D}$ into N folds with each fold $\mathcal{D}^{(i)}$ ($i = 1, ..., N$) has M different episodes, or in other words, we split the n episodes evenly. Then by the i.i.d. nature of n episodes, we have $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, ..., \mathcal{D}^{(N)}$ are independent collections.

For each $\mathcal{D}^{(i)}$, we can create a Tabular-MIS estimator $\widehat{v}^{\pi(i)}_{\text{TMIS}}$ (for notation simplicity we use $\widehat{v}^{\pi}_{(i)}$ to denote $\widehat{v}^{\pi(i)}_{\text{TMIS}}$ in the future discussions) using its own M episodes. Then $\widehat{v}^{\pi}_{(1)}, \widehat{v}^{\pi}_{(2)}, ..., \widehat{v}^{\pi}_{(N)}$ are independent of each other and we can use the empirical mean to define the data splitting

[6] In general this might not be true, *e.g.* if n is prime number. However, we can resolve it by choosing $M = \lfloor n/N \rfloor$.

Tabular-MIS estimatorand the corresponding fictitious version:

$$\widehat{v}^{\pi}_{\text{split}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{v}^{\pi}_{(i)}, \qquad \widetilde{v}^{\pi}_{\text{split}} = \frac{1}{N}\sum_{i=1}^{N}\widetilde{v}^{\pi}_{(i)}, \tag{2.12}$$

where each $\widetilde{v}^{\pi}_{(i)}$ is the fictitious estimator of $\widehat{v}^{\pi}_{(i)}$.

The data splitting TMIS estimator explicitly characterizes the independence of n different episodes by grouping them into N chunks. Chunks are independent of each other and taking the average over all $\widehat{v}^{\pi}_{(i)}$ $i = 1, ..., N$ will guarantee the validity of using concentration inequalities.

More importantly, the data splitting TMIS estimator holds the same information-theoretical variance lower bound as the non-data splitting TMIS estimator, which is not surprising since the non-data splitting TMIS estimator is just the special case of the data splitting Tabular-MIS estimator with $N = 1$. This idea is summarized into the following theorem:

Theorem 2.7.1. *Using n i.i.d. episodic data from a near-uniform*[7] *logging policy μ and suppose M, the number of episodes for each $\mathcal{D}^{(i)}$, satisfies:*

$$M > \max\left[O(SA \cdot \text{Polylog}(S, H, A, n)), O(H\tau_a\tau_s)\right],$$

then the data splitting Tabular-MIS estimator obeys:

$$\mathbb{E}[(\widehat{v}^{\pi}_{\text{split}} - v^{\pi})^2] \leq O(\frac{H^2SA}{n}). \tag{2.13}$$

Remark 4. *The condition in Theorem 2.7.1 is achievable. For example, choose $M \approx \sqrt{n}$, then the condition holds when n is sufficiently large.*

High probability bound. By coupling the data splitting techniques with the boundedness of Tabular-MIS estimator (*i.e.* $\widehat{v}^{\pi} \leq HR_{\max}, \widetilde{v}^{\pi} \leq HR_{\max}$, see Lemma A.2.3 in appendix), we

[7]Near-uniform here means: $\min_{t,s_t,a_t} d_t^{\mu}(s_t, a_t) > \Omega(1/(SA))$.

can apply concentration inequalities to show the difference between $\widehat{v}^{\pi}_{\text{split}}$ and v^{π} is bounded by order $\widetilde{O}(\sqrt{H^2SA/n})$, which is summarized into the following theorem.

Theorem 2.7.2. *Suppose n i.i.d. episodic historical data comes from a near-uniform logging policy μ and suppose M, the number of episodes in each $\mathcal{D}^{(i)}$, satisfies: $\widetilde{O}(\sqrt{n \cdot SA}) \geq M$ and $M > \max\left[O(SA \cdot \text{Polylog}(S, H, A, n, 1/\delta)), O(H\tau_a\tau_s)\right]$. Then we have with probability $1-\delta$, the data splitting Tabular-MIS estimator obeys:*

$$|\widehat{v}^{\pi}_{\text{split}} - v^{\pi}| \leq \widetilde{O}(\sqrt{\frac{H^2SA}{n}}).$$

The proof Theorem 2.7.2 relies on bounding the difference between $\widehat{v}^{\pi}_{\text{split}}$ and $\widetilde{v}^{\pi}_{\text{split}}$ using Multiplicative Chernoff bound and bounding the difference between $\widetilde{v}^{\pi}_{\text{split}}$ and v^{π} using Bernstein inequality. During the process of bounding $|\widehat{v}^{\pi}_{\text{split}} - \widetilde{v}^{\pi}_{\text{split}}|$ we observe that a stronger uniform bound can be derived. In fact, this bound is 0. We formalize it into the following lemma.

Lemma 2.7.1. *Suppose n i.i.d. episodic historical data comes from a near-uniform logging policy μ and suppose M, the number of episodes in each $\mathcal{D}^{(i)}$, satisfies:*

$$M > \max\left[O(SA \cdot \text{Polylog}(S, H, A, N, 1/\delta)), O(H\tau_a\tau_s)\right].$$

Then we have with probability $1-\delta$,

$$\sup_{\pi \in \prod} |\widehat{v}^{\pi}_{\text{split}} - \widetilde{v}^{\pi}_{\text{split}}| = 0$$

Since $n = N \cdot M$, therefore let $N = 1$, $M = n$, then if

$$M > \max\left[O(SA \cdot \text{Polylog}(S, H, A, 1/\delta)), O(H\tau_a\tau_s)\right],$$

$$\sup_{\pi \in \prod} |\widehat{v}^{\pi}_{\text{TMIS}} - \widetilde{v}^{\pi}_{\text{TMIS}}| = 0$$

holds with probability $1-\delta$*, where* $\prod$ *consists of all the* H*-step nonstationary policies.*

Remark 5. *The uniform difference bound between* $\widehat{v}^{\pi}_{\text{TMIS}}$ *and* $\widetilde{v}^{\pi}_{\text{TMIS}}$ *is obtained by observing the construction of fictitious estimator* (2.10) *and* (2.11) *are independent of the specific target policy* π*. This result tells the* $\sup_{\pi\in\prod}|\widehat{v}^{\pi}_{\text{TMIS}}-\widetilde{v}^{\pi}_{\text{TMIS}}|$ *can be arbitrarily small with high probability and therefore does not depend on* H *factor. This fact will help us to derive the correct dependence in* H *for uniform convergence problem.*

2.8 Empirical validation

In this section, we present some empirical studies to demonstrate that our main theoretical results about Tabular-MIS estimator presented in Theorem 2.5.1 are empirically verified.

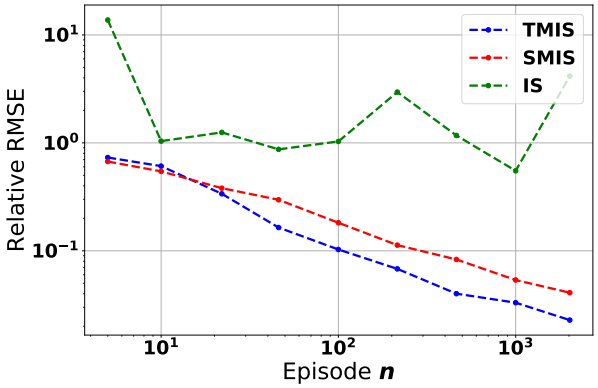

Figure 2.1: Different Episode n, Relative RMSE ($\sqrt{\text{MSE}}/v^{\pi}$) on Non-stationary Non-mixing MDP

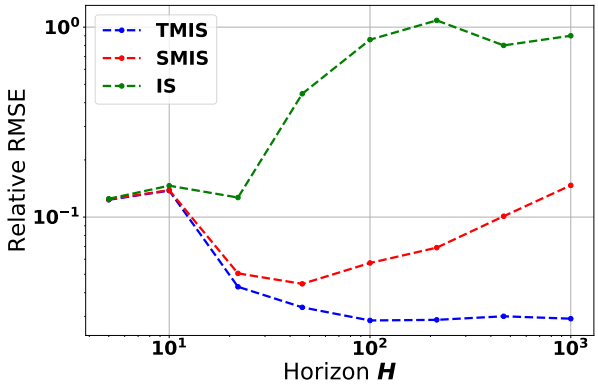

Figure 2.2: Different Horizon H, Relative RMSE ($\sqrt{\text{MSE}}/v^{\pi}$) on Non-stationary Non-mixing MDP

Time-varying, non-mixing Tabular MDP. We test our approach in simulated MDP environment where both the states and the actions are binary. Concretely, there are two states s_0 and s_1 and two actions a_1 and a_2. State s_0 always has probability 1 going back to itself, regardless of the actions, *i.e.* $P_t(s_0|s_0,a_1)=1$ and $P_t(s_0|s_0,a_2)=1$. For state s_1, at each time step there is one action (we call it a) that has probability $2/H$ going to s_0 and the other action (we call it a') has

probability 1 going back to s_1, *i.e.* $P_t(s_0|s_1, a) = 2/H = 1 - P_t(s_1|s_1, a)$ and $P_t(s_1|s_1, a') = 1$. Moreover, which action will make state s_1 go to state s_0 with probability $2/H$ is decided by a random parameter $p_t \in [0, 1]$. If $p_t < 0.5$, $a = a_1$ and if $p_t \geq 0.5$, $a = a_2$. One can receive reward 1 at each time step if $t > H/2$ and is in state s_0, and will receive reward 0 otherwise. Lastly, for state s_0, we set $\mu(\cdot|s_0) = \pi(\cdot|s_0)$; for state s_1, we set $\mu(a_1|s_1) = \mu(a_2|s_1) = 1/2$ and $\pi(a_1|s_1) = 1/4 = 1 - \pi(a_2|s_1)$.

Figure 2.1 shows the asymptotic convergence rates of relative RMSE with respect to the number of episodes, given fixed horizon $H = 100$. Both SMIS and TMIS has a $O(1/\sqrt{n})$ convergence rate. The saving of $\sqrt{H}$ of TMIS over SMIS in this log-log plot is reflected in the intercept. Figure 2.2 has fixed $n = 1024$ with varying horizon H. Note since $v^\pi \approx O(H)$, therefore for TMIS our theoretical result implies $\sqrt{\text{MSE}}/v^\pi = O(\sqrt{H^2}/H) = O(1)$, which is consistent with the horizontal line when H is large. Moreover, for SMIS $\sqrt{\text{MSE}}/v^\pi = O(\sqrt{H^3}/H) = O(\sqrt{H})$, so after taking the $\log(\cdot)$ we should have asymptotic linear trend with coefficient $1/2$. The red line in Figure 2.2 empirically verifies this result. More empirical study discussions are deferred to Appendix A.4.

2.9 Discussion

Logging policy free algorithm. We point out that the implementation of Tabular-MIS estimator does not require the knowledge of logging policy μ, as shown in Algorithm 3,2.[8] This is critical in the sense that in the real-world sequential decision making problems, it is very likely the complete information about logging policy is not available. This may happen due to mis-records or the lack of maintenance. By only using the historical data, tabular MIS off-policy evaluation is able to achieve the asymptotic efficiency. In contrast, the state MIS estimator always requires the full information about the logging policy.

[8] Algorithm 2 is deferred to appendix due to space constraint.

Connection to approximate MDP estimation. Our TMIS is essentially an approximate MDP estimator (with the non-stationary dynamic transitions P_t estimated by *maximum likelihood estimator* (MLE)) except that we marginalize out the action in both $\widehat{r}_t^{\pi}(s)$ and $\widehat{d}_t^{\pi}(s)$ and provide an importance sampling interpretation. To the best of our knowledge, existing analysis of the fully model-based approach does not provide tight bounds. We give two examples. The seminal "simulation lemma" in [40] together with a naive concentration-type analysis gives only an $\widetilde{O}(\sqrt{H^4S^3A/n})$ bound in our setting. In a very recent compilation of improvements over this bound [41], this bound can be improved to either $\widetilde{O}(\sqrt{H^4S^2A/n})$ or $\widetilde{O}(\sqrt{H^6SA/n})$. Our result is the first that achieves the optimal $\widetilde{O}(\sqrt{H^2SA/n})$ rate regardless of whether it is the model-based or model-free approach.

From off-policy evaluation to offline learning. A real offline reinforcement learning system is equipped with both offline learning algorithms and off-policy evaluation algorithms. The decision maker should first run the offline learning algorithm to find a near optimal policy and then use off-policy evaluation methods to check if the obtained policy is good enough. Under our tabular MDP setting, we point out it is possible to find a ϵ-optimal policy in near optimal time and sample complexity $O(H^3SA/\epsilon^2)$using the Q-value iteration (QVI) based algorithm designed by [42]. Their QVI algorithm assumes a generative model which can provide independent sample of the next state s' given any current state-action (s, a). At a first glance, this assumption seems too strong for offline learning since we cannot force the agent to stay in any arbitrary location. In fact, the Assumption 4.4.1 on μ actually reveals that the underlying logging policy can be considered as the surrogate of the generative model. As n gets large, the visitation frequency of any (s_t, a_t) will be large enough with high probability, as guaranteed by Multiplicative Chernoff bound.

From off-policy evaluation to uniform off-policy evaluation. The high probability result achieves $\widetilde{O}(\sqrt{H^2SA/n})$ complexity. Following this discovery line, then it is natural to ask whether uniform convergence over a class of policies (*e.g.* all deterministic policies) can be

achieved with optimal sample complexity. This problem is interesting since it will guarantee the strong performance of off-policy evaluation methods over all policies in certain policy class $\prod$. By a direct application of union bound, we can obtain the following result:

Theorem 2.9.1. *Let $\prod$ contains all the deterministic H-step policies. Then under the same condition as Theorem 2.7.2, the data splitting Tabular-MIS estimator satisfies:*

$$\sup_{\pi \in \prod} |\widehat{v}^{\pi}_{\text{split}} - v^{\pi}| \leq \widetilde{O}(\sqrt{\frac{H^3 S^2 A}{n}}),$$

with probability $1 - \delta$.

The uniform convergence bound implies that the empirical best policy $\hat{\pi} = \text{argmax}_{\pi} \widehat{v}^{\pi}_{\text{split}}$ is within $\epsilon = O(\sqrt{\frac{H^3 S^2 A}{n}})$ of the optimal policy. This matches the sample complexity lower bound for learning the optimal policy [43] in all parameters except a factor of S.

In this chapter, we proposed a new marginalized importance sampling estimator for the off-policy evaluation (OPE) problem under the episodic tabular setting. We show that this estimator has a finite sample error bound that matches the exact Cramer-Rao lower bound up to low-order factors. We also provide an extension with high probability error bound. To the best of our knowledge, these results are the first of their kind. Future work includes resolving the open problems mentioned above and generalizing the results to more practical settings.

Chapter 3

Uniform Convergence in Offline Policy Evaluation

3.1 Introduction

In offline reinforcement learning (offline RL), there are mainly two fundamental parts: *offline policy evaluation* (OPE) and *offline learning* (also known as *batch RL*) [12]. OPE addresses the statistical estimation problem of predicting the performance of a fixed target policy π with only data collected by a logging/behavioral policy μ. On the other hand, offline learning is a *statistical learning* problem that aims at learning a near-optimal policy using an offline dataset alone [2].

As offline RL methods do not require interacting with the task environments or having access to a simulator, they are more suitable for real-world applications of RL such as those in marketing [23], targeted advertising [44, 45], finance [46], robotics [47, 48], language [49] and health care [50, 51, 52, 53]. In these tasks, it is usually not feasible to deploy an online RL algorithm to trials-and-error with the environment. Instead, we are given a large offline dataset of historical interaction to come up with a new policy π and to demonstrate that this new policy

π will perform better using the same dataset without actually testing it online.

In this chapter, we present our solution via a statistical learning perspective by studying the *uniform convergence* in OPE under *non-stationary transition, finite horizon, episodic Markov decision process (MDP)* model with finite states and actions. Informally, given a policy class Π and a logging policy μ, uniform convergence problem in OPE (Uniform OPE for short) focuses on coming up with OPE estimator $\widehat{v}^{\pi}$ and characterizing the number of episodes n we need (from μ) in order for $\widehat{v}^{\pi}$ to satisfies that with high probability

$$\sup_{\pi \in \Pi} \left| \widehat{v}^{\pi} - v^{\pi} \right| \leq \epsilon.$$

The focus of research would be to characterizing the *episode complexity*: the number of episodes n needed as a function of ϵ, failure probability δ, the parameters of the MDP as well as the logging policy μ.

We highlight that even though uniform convergence is the main workhorse in statistical learning theory see, e.g.,[54], few analogous results have been established for the offline reinforcement learning problem. The overarching theme of this work is to understand what a natural complexity measure is for policy classes in reinforcement learning and its dependence in the size of the state-space and planning horizon.

In addition, uniform OPE has two major consequences (which we elaborate in detail in the following motivation section), but briefly: (1) allowing any accurate planning algorithm to work as sample efficient offline learning algorithm with our model-based method; (2) providing finite sample guarantee for offline evaluation uniformly for all policies in the policy class.

3.1.1 Motivation of Uniform Convergence in OPE

Existing research in offline RL usually focuses on designing specific algorithms that learn the optimal policy $\pi^{\star} := \operatorname{argmax}_{\pi} v^{\pi}$ with given static offline data $\mathcal{D}$. In the rich literature of

statistical learning theory, however, learning bounds are often obtained via a stronger uniform convergence argument which ensures an arbitrary learner to output a model that generalizes. Specifically, the *empirical risk minimizer* (ERM) that outputs the *empirical optimal policy* has been shown to be sufficient and necessary for efficiently learning almost all learnable problems [54, 55].

The natural analogy of ERM in the RL setting would be to find the *empirical optimal policy* $\widehat{\pi}^\star := \operatorname{argmax}_\pi \widehat{v}^\pi$ for some OPE estimator $\widehat{v}^\pi$. If we could establish a uniform convergence bound for $\widehat{v}^\pi$, then it implies that $\widehat{\pi}^\star$ is nearly optimal too via

$$\begin{aligned} 0 \leq v^{\pi^\star} - v^{\widehat{\pi}^\star} &= v^{\pi^\star} - \widehat{v}^{\widehat{\pi}^\star} + \widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star} \\ &\leq |v^{\pi^\star} - \widehat{v}^{\pi^\star}| + |\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}| \leq 2 \sup_\pi |v^\pi - \widehat{v}^\pi|. \end{aligned}$$

Thus, uniform OPE is a stronger setting than offline learning with the additional benefit of accurately evaluating any other (possibly heuristic) policy optimization algorithms that are used in practice.

From the OPE perspective, there is often a need to evaluate the performance of a *data-dependent* policy, and uniform OPE becomes useful. For example, when combined with existing methods, it will allow us to evaluate policies selected by safe-policy improvements, proximal policy optimization, UCB-style exploration-bonus as well as any heuristic exploration criteria such as curiosity, diversity and reward-shaping techniques.

Model-based estimator for OPE. The OPE estimator we consider here is the standard model-based estimator, i.e. estimating the transition dynamics and immediate rewards, then simply plug in the parameters of empirically estimated MDP $\widehat{M}$ to obtain $\hat{v}^\pi$ for any π. This model-based approach has several benefits. **1.** It enables flexible choice of policy search methods since it converts the problem to planning over the estimated MDP $\widehat{M}$. **2.** Uniform OPE with model-based estimator avoids the use of data-splitting that leads to inefficient data use. For example,

[42] learns the ϵ-optimal policy with the optimal rate in the generative model setting, where in each subroutine new independent data $s_{s,a}^{(1)}, ..., s_{s,a}^{(m)}$ need to be sampled to estimate $P_{s,a}$ and samples from previous rounds cannot be reused. A uniform convergence result could completely avoid data splitting during the learning procedure.

3.2 Related Literature

1. OPE: Most existing work on OPE focuses on the *Importance Sampling* (IS) methods [13, 14, 17, 19] or their doubly robust variants [18, 56]. These methods are more generally applicable even if the the Markovian assumption is violated or the states are not observable, but has an error (or sample complexity) that depends exponential dependence in horizon H. Recently, a family of estimators based on *marginalized importance sampling* (MIS) [29, 21, 32, 33, 57] have been proposed in order to overcome the "*curse of horizon*" under the additional assumption of state observability. In the tabular setting, [57] design the Tabular-MIS estimator which matches the Cramer-Rao lower bound constructed by [18] up to a low order term for every instance (π, μ and the MDP), which translates into an $O(H^2/d_m\epsilon^2)$ episode complexity in the (pointwise) OPE problem we consider for all π (as we discussed in Chapter 2). Tabular-MIS, however, is identical to the model-based plug-in estimator we use, *off-policy empirical model approximator* (OPEMA), as we will discuss further in this chapter.

2. Offline Learning: For the offline learning, most theoretical work considers the infinite horizon discounted setting with function approximation. [4, 3] first raises the information-theoretic considerations for offline learning and uses Fitted Q-Iteration (FQI) to obtain $\epsilon V_{\max}$-optimal policy using sample complexity $\widetilde{O}((1-\gamma)^{-4}C_\mu/\epsilon^2)$ where C_μ is *concentration coefficient* [58] that is similar to our $1/d_m$. More recently, [5] improves the result to $\tilde{O}((1-\gamma)^{-2}C_\mu/\epsilon^2)$.

However, these bounds are not tight in terms of the dependence on the effective horizon[1] $(1-\gamma)^{-1}$. More recently, [6, 59] explore weaker settings for batch learning but with sub-optimal sample complexity dependencies. Our result is the first that achieves the optimal rate, although restricted to the finite horizon episodic setting.

3. Uniform convergence in RL: There are few existing work that deals with uniform convergence in OPE. However, we notice that the celebrated simulation lemma [40] is actually an uniform bound with an episode complexity of $O(H^4S^2/d_m\epsilon^2)$. Several existing work uses uniform-convergence arguments over value function classes for online RL [60]. The closest to our work is perhaps [61], which studies model-based planning in the generative model setting. We are different in that we are in the offline learning setting. In addition, our local policy class is optimal for a larger region of ϵ_{opt} (independent to n), while their results (Lemma 10) imply optimal OPE only for empirically optimal policy with $\epsilon_{\text{opt}} \leq \sqrt{(1-\gamma)^{-5}SA/n}$. Lastly, we discovered the thesis of [62, Ch.3 Theorem 1], which discusses the pseudo-dimension of policy classes. The setting is not compatible to ours, and does not imply a uniform OPE bound in our setting.

3.3 Uniform Convergences Problems

We first review the tabular RL setting we discussed in Chapter 2. RL environment is usually modeled as a *Markov Decision Process* (MDP) which is denoted by $M = (\mathcal{S}, \mathcal{A}, r, P, d_1, H)$. The MDP consists of a state space $\mathcal{S}$, an action space $\mathcal{A}$ and a transition kernel $P_t : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \mapsto [0,1]$ with $P_t(s'|s,a)$ representing the probability transition from state s, action a to next state s' at time t. In particular here we consider non-stationary transition dynamics so P_t varies over time t. Besides, $r_t : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}$ is the expected reward function and given (s_t, a_t), $r_t(s_t, a_t)$

[1]The optimal rate should be $(1-\gamma)^{-1}C/\epsilon^2$, analogous to our $H^3/d_m\epsilon^2$ bound. The additional H^2 is due to scaling — we are obtaining ϵ-optimal policy and they obtain $\epsilon V_{\max}$-optimal policy ($V_{\max} = H$ in our case). See Table 3.1 for a consistent comparison.

specifies the average reward obtained at time t. d_1 is the initial state distribution and H is the horizon. Moreover, we focus on the case where state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite, *i.e.* $S := |\mathcal{S}| < \infty, A := |\mathcal{A}| < \infty$. A (non-stationary) policy is formulated by $\pi := (\pi_1, \pi_2, ..., \pi_H)$, where π_t assigns each state $s_t \in \mathcal{S}$ a probability distribution over actions at each time t. Any fixed policy π together with MDP M induce a distribution over trajectories of the form $(s_1, a_1, r_1, s_2, ..., s_H, a_H, r_H, s_{H+1})$ where $s_1 \sim d_1$, $a_t \sim \pi_t(\cdot|s_t)$, $s_{t+1} \sim P_t(\cdot|s_t, a_t)$ and r_t has mean $r_t(s_t, a_t)$ for $t = 1, ..., H$.[2]

In addition, we denote $d_t^\pi(s_t, a_t)$ the induced marginal state-action distribution and $d_t^\pi(s_t)$ the marginal state distribution, satisfying $d_t^\pi(s_t, a_t) = d_t^\pi(s_t) \cdot \pi(a_t|s_t)$. Moreover, $d_1^\pi = d_1 \; \forall \pi$. We use the notation $P_t^\pi \in \mathbb{R}^{S\cdot A \times S\cdot A}$ to represent the state-action transition $(P_t^\pi)_{(s,a),(s',a')} := P_t(s'|s, a)\pi_t(a'|s')$, then the marginal state-action vector $d_t^\pi(\cdot, \cdot) \in \mathbb{R}^{S\times A}$ satisfies the expression $d_{t+1}^\pi = P_{t+1}^\pi d_t^\pi$. We define the quantity $V_t^\pi(s) = \mathbb{E}_\pi[\sum_{t'=t}^H r_{t'}|s_t = s]$ and the Q-function $Q_t^\pi(s, a) = \mathbb{E}_\pi[\sum_{t'=t}^H r_{t'}|s_t = s, a_t = a]$ for all $t = 1, ..., H$. The ultimate measure of the performance of policy π is the value function:

$$v^\pi = \mathbb{E}_\pi \left[\sum_{t=1}^H r_t \right].$$

Lastly, for the standard OPE problem, the goal is to estimate v^π for a given π while assuming that n episodic data $\mathcal{D} = \left\{ (s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)}) \right\}_{i\in[n]}^{t\in[H]}$ are rolling from a *different* policy μ.

Uniform OPE extends the pointwise OPE to a family of policies. Specifically, for an policy class Π of interest, we aim at showing that $\sup_{\pi\in\Pi} |\widehat{v}^\pi - v^\pi| < \epsilon$ with high probability with optimal dependence in all parameters. In this paper, we consider three policy classes.

The global policy class. The policy class Π we considered here consists of all the non-stationary policies, deterministic or stochastic. This is the largest possible class we can consider and hence the hardest one.

[2]Here r_t without any argument is random reward and $\mathbb{E}[r_t|s_t, a_t] = r_t(s_t, a_t)$.

The global deterministic policy class. Here class consists of all the non-stationary deterministic policies. By the standard results in reinforcement learning, there exists at least one deterministic policy that is optimal [12]. Therefore, the deterministic policy class is rich enough for evaluating any learning algorithm (*e.g.* Q-value iteration in [42]) that wants to learn to the optimal policy.

The local policy class: in the neighborhood of empirical optimal policy. Given empirical MDP $\widehat{M}$ (*i.e.* the transition kernel is replaced by $\widehat{P}_t(s_{t+1}|s_t, a_t) := n_{s_{t+1},s_t,a_t}/n_{s_t,a_t}$ if $n_{s_t,a_t} > 0$ and 0 otherwise, where n_{s_t,a_t} is the number of visitations to (s_t, a_t) among all n episodes[3]), it is convenient to learn the empirical optimal policy $\widehat{\pi}^\star := \text{argmax}_\pi \widehat{v}^\pi$ since the full empirical transition $\widehat{P}$ is known. Standard methods like Policy Iteration (PI) and Value Iteration (VI) can be leveraged for finding $\widehat{\pi}^\star$. This observation allows us to consider the following interesting policy class: $\Pi_1 := \{\pi : s.t. \ ||\widehat{V}_t^\pi - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \leq \epsilon_{\text{opt}}, \ \forall t = 1, ..., H\}$ with $\epsilon_{\text{opt}} \geq 0$ a parameter. Here we consider $\widehat{\pi}^\star$ (instead of $\pi^\star$) since by defining with empirical optimal policy, we can use data $\mathcal{D}$ to really check class Π_1, therefore this definition is more practical.

3.3.1 Assumptions

Next we present some mild necessary regularity assumptions for uniform convergence OPE problem.

Assumption 3.3.1 (Bounded rewards)**.** $\forall \ t = 1, ..., H$ *and* $i = 1, ..., n$, $0 \leq r_t^{(i)} \leq 1$.

Assumption 3.3.2 (Exploration requirement)**.** *Logging policy* μ *obeys that* $\min_{t,s_t} d_t^\mu(s_t) > 0$*, for any state* s_t *that is "accessible". Moreover, we define quantity* $d_m := \min\{d_t^\mu(s_t, a_t) : d_t^\mu(s_t, a_t) > 0\}$.

State s_t is "accessible" means there exists a policy π so that $d_t^\pi(s_t) > 0$. If for any policy π we always have $d_t^\pi(s_t) = 0$, then state s_t can never be visited in the given MDP. Assumption 4.4.1

[3]Similar definition holds for n_{s_{t+1},s_t,a_t}.

simply says μ have the right to explore all "accessible" states. This assumption is required for the consistency of uniform convergence estimator since we have "$\sup_{\pi\in\Pi}$" and is similar to the standard *concentration coefficient* assumption made by [58, 3]. As a short comparison, offline learning problems (*e.g.* offline policy optimization in [63]) only require $d_t^{\mu}(s_t) > 0$ for any state s_t satisfies $d_t^{\pi^\star}(s_t) > 0$. Last but not least, even though our target policy class is deterministic, by the above assumptions μ is always stochastic.

3.4 Method: Offline Policy Empirical Model Approximator

The method we use for doing OPE in uniform convergence is the *offline policy empirical model approximator* (OPEMA). OPEMA uses off-policy data to build the empirical estimators for both the transition dynamic and the expected reward and then substitute the related components in real value function by its empirical counterparts. First recall for any target policy π, by definition: $v^\pi = \sum_{t=1}^H \sum_{s_t,a_t} d_t^\pi(s_t,a_t) r_t(s_t,a_t)$, where the marginal state-action transitions satisfy $d_{t+1}^\pi = P_{t+1}^\pi d_t^\pi$. OPEMA then directly construct empirical estimates for $\widehat{P}_{t+1}(s_{t+1}|s_t,a_t)$ and $\widehat{r}_t(s_t,a_t)$ as:

$$\widehat{P}_{t+1}(s_{t+1}|s_t,a_t) = \frac{\sum_{i=1}^n \mathbf{1}[(s_{t+1}^{(i)}, a_t^{(i)}, s_t^{(i)}) = (s_{t+1}, s_t, a_t)]}{n_{s_t,a_t}}, \quad \widehat{r}_t(s_t,a_t) \quad = \frac{\sum_{i=1}^n r_t^{(i)} \mathbf{1}[(s_t^{(i)}, a_t^{(i)}) = (s_t,a_t)]}{n_{s_t,a_t}}.$$

and $\widehat{P}_{t+1}(s_{t+1}|s_t,a_t) = 0$ and $\widehat{r}_t(s_t,a_t) = 0$ if $n_{s_t,a_t} = 0$ (recall n_{s_t,a_t} is the visitation frequency to (s_t,a_t) at time t) and then the estimates for state-action transition $\widehat{P}_t^\pi$ is defined as: $\widehat{P}_t^\pi(s_{t+1},a_{t+1}|s_t,a_t) = \widehat{P}_t(s_{t+1}|s_t,a_t)\pi(a_{t+1}|s_{t+1})$.The initial distribution is also constructed using empirical estimator $\widehat{d}_1^\pi(s_1) = n_{s_1}/n$. Based on the construction, the empirical marginal state-action transition follows $\widehat{d}_{t+1}^\pi = \widehat{P}_{t+1}^\pi \widehat{d}_t^\pi$ and the final estimator for v^π is:

$$\widehat{v}_{\mathrm{OPEMA}}^\pi = \sum_{t=1}^H \sum_{s_t,a_t} \widehat{d}_t^\pi(s_t,a_t)\widehat{r}_t(s_t,a_t). \tag{3.1}$$

OPEMA is model-based method as it uses plug-in estimators ($\widehat{d}_t^{\pi}$ and $\widehat{r}_t$) for each model components (d_t^{π} and r_t). Traditionally, the error of OPEMA is obtained via the simulation lemma [40], with $O(H^4S^2/d_m\epsilon^2)$-episode complexity. Recent work in [21, 57, 64] reveals that there is an importance sampling interpretation of OPEMA

$$\widehat{v}^{\pi}_{\text{OPEMA}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\hat{d}^{\pi}(s_t^{(i)})}{\hat{d}_t^{\mu}(s_t^{(i)})}\hat{r}_t^{\pi}(s^{(i)}), \tag{3.2}$$

and the effectiveness of MIS of recent work partially explains why OPEMA could work, even for the Uniform OPE problem.

3.5 Main Results for Uniform OPE

In this section, we present our results for uniform OPE problems. For brevity, we use $\widehat{v}^{\pi}$ to denote $\widehat{v}^{\pi}_{\text{OPEMA}}$ in the rest of the presentattion. Proofs of all technical results are deferred to the appendix. We start with the following Lemma:

Lemma 3.5.1 (martingale decomposition)**.** *For fixed π:*

$$\sum_{t=1}^{H}\langle\widehat{d}_t^{\pi} - d_t^{\pi}, r_t\rangle = \sum_{h=2}^{H}\langle V_h^{\pi}, (\widehat{T}_h - T_h)\widehat{d}_{h-1}^{\pi}\rangle + \langle V_1^{\pi}, \widehat{d}_1^{\pi} - d_1^{\pi}\rangle$$

where $T_{h+1} \in \mathbb{R}^{S\times(SA)}$ be the one step transition matrix, i.e. $T_{s_{h+1},(s_h,a_h)} = P_{h+1}(s_{h+1}|s_h,a_h)$. the inner product on the left hand side is taken w.r.t state-action and the inner product on the left hand side is taken w.r.t state only. Proof can be found in Appendix B.

Remark 6. *Note when the reward is deterministic, the left hand side is simply $\widehat{v}^{\pi} - v^{\pi}$, and the right hand side has a martingale structure which enables the applicability of concentration analysis that gives rise to the following theorems. Moreover, this decomposition is essentially "primal-dual" formulation since the LHS can be viewed as the primal form through marginal*

distribution representation and RHS is the dual form with value function representation.

3.5.1 Uniform OPE for global policy class

We present the following result Theorem 3.5.1 for global policy class.

Theorem 3.5.1. *Let* Π *consists of all policies, then there exists an absolute constant* c *such that if* $n > c \cdot 1/d_m \cdot \log(HSA/\delta)$*, then with probability* $1-\delta$*, we have:*

$$\sup_{\pi\in\Pi} \left|\widehat{v}^{\pi} - v^{\pi}\right| \leq c\left(\sqrt{\frac{H^4 \log(\frac{HSA}{\delta})}{d_m \cdot n}} + \sqrt{\frac{H^4 S \log(nHSA)}{d_m \cdot n}}\right).$$

Moreover, if failure probability $\delta < e^{-S}$*, then above can be further bounded by* $2c\sqrt{\frac{H^4}{d_m \cdot n}\log(\frac{nHSA}{\delta})}$.

The first term in the bound reflects the concentration of $\sup_{\pi\in\Pi}\left|\widehat{v}^{\pi} - v^{\pi}\right|$ around its mean, via McDiarmid inequality. The second term is a bound of $\mathbb{E}[\sup_{\pi\in\Pi}\left|\widehat{v}^{\pi} - v^{\pi}\right|]$. The analysis of both terms rely on the Martingale decomposition from Lemma 3.5.1.

Our result improves over the simulation lemma by a factor of HS but is sub-optimal by another factor HS compared to the lower bound (Theorem 3.5.4). In the small failure probability regime ($\delta < e^{-S}$) we can get rid of the dependence on S except for the implicit dependence through d_m. This is meaningful since we usually consider deriving results with high confidence.

3.5.2 Uniform OPE for deterministic policies

The Martingale decomposition also allows us to derive a high-probability OPE bound via a concentration argument, which complements the optimal bounds on mean square error from [57].

Lemma 3.5.2 (Convergence for fixed policy)**.** *Fix any policy π. Then there exists absolute constants c, c_1, c_2 such that if $n > c \cdot 1/d_m \cdot \log(HSA/\delta)$, then with probability $1-\delta$, we have:*

$$\left|\widehat{v}^{\pi} - v^{\pi}\right| \leq c_1 \sqrt{\frac{H^2 \log(\frac{c_2 HSA}{\delta})}{n \cdot d_m}} + \tilde{O}\left(\frac{H^2\sqrt{SA}}{n \cdot d_m}\right).$$

Note if we absorb the higher order term, our result implies sample complexity of $\widetilde{O}(H^2/d_m\epsilon^2)$ for evaluating any fixed target policy π. Notice that the total number of deterministic policies is A^{HS} in our problem, a standard union bound over all deterministic policies yields the following result.

Theorem 3.5.2. *Let Π consists of all deterministic policies, then there exists absolute constants c, c_1, c_2 such that if $n > c \cdot 1/d_m \cdot \log(HSA/\delta)$, then with probability $1-\delta$, we have:*

$$\sup_{\pi \in \Pi} \left|\widehat{v}^{\pi} - v^{\pi}\right| \leq c_1 \sqrt{\frac{H^3 S \log(\frac{c_2 HSA}{\delta})}{n \cdot d_m}} + \tilde{O}\left(\frac{H^3 S^{1.5} A^{0.5}}{n \cdot d_m}\right).$$

Theorem 3.5.2 implies an episode complexity of $\widetilde{O}(H^3 S/d_m\epsilon^2)$, which is optimal in H but suboptimal by a factor of S. While the deterministic policy class seems restrictive, it could be useful in many cases because the optimal policy is deterministic, and many exploration-bonus based exploration methods use deterministic policy throughout.

3.5.3 Uniform OPE for the local (near *empirically optimal*) policy class

For the local (near *empirically optimal*) policy class, the following theorem obtains the optimal episode complexity.

Theorem 3.5.3. *Suppose $\epsilon_{opt} \leq \sqrt{H}/S$ and $\Pi_1 := \{\pi : s.t.\ ||\widehat{V}_t^{\pi} - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \leq \epsilon_{opt},\ \forall t = 1, ..., H\}$. Then there exists constant c_1, c_2 such that for any $0 < \delta < 1$, when $n > c_1 H^2 \log(HSA/\delta)/d_m$,*

we have with probability $1-\delta$*,*

$$\sup_{\pi\in\Pi_1}\left\|\widehat{Q}_1^\pi - Q_1^\pi\right\|_\infty \le c_2\sqrt{\frac{H^3\log(HSA/\delta)}{n\cdot d_m}}.$$

This uniform convergence result is presented with l_∞ norm over (s,a). A direct corollary is $\sup_{\pi\in\Pi_1}\left\|\widehat{V}_1^\pi - V_1^\pi\right\|_\infty$ achieves the same rate. Theorem 3.5.3 provides the sample complexity of $O(H^3\log(HSA/\delta)/d_m\epsilon^2)$ and the dependence of all parameters are optimal up to the logarithmic term. Note that our bound does not explicitly depend on ϵ_{opt}, which is an improvement over agarwal2020model as they have an additional $O(\epsilon_{\text{opt}}/(1-\gamma))$ error in the infinite horizon setting. Besides, our assumption on ϵ_{opt} is mild since the required upper bound is proportional to $\sqrt{H}$. Lastly, this result implies a $O(\epsilon+\epsilon_{\text{opt}})$-optimal policy for offline/batch learning of the optimal order $O(H^3\log(HSA/\delta)/d_m\epsilon^2)$, which means statistical learning result enables offline learning.

3.5.4 Information-theoretical lower bound

Finally, we present a fine-grained sample complexity lower bound of the uniform OPE problem that captures the dependence of all parameters including d_m.

Theorem 3.5.4 (Minimax lower bound for uniform OPE)**.** *For all* $0 < d_m \le \frac{1}{SA}$*. Let the class of problems be*

$$\mathcal{M}_{d_m} := \left\{(\mu, M) \;\middle|\; \min_{t,s_t,a_t} d_t^\mu(s_t,a_t) \ge d_m\right\}.$$

There exists universal constants c_1, c_2, c_3, p *(with* $H, S, A \ge c_1$ *and* $0 < \epsilon < c_2$*) such that*

$$\inf_{\widehat{v}}\sup_{(\mu,M)\in\mathcal{M}_{d_m}}\mathbb{P}_{\mu,M}\left(\sup_{\pi\in\Pi}|\widehat{v}^\pi - v^\pi| \ge \epsilon\right) \ge p$$

if $n \le c_3H^3/d_m\epsilon^2$*. Here* Π *consists of all deterministic policies.*

The proof uses a reduction argument that shows if a stronger uniform OPE bound exists, then it implies an algorithm that breaks an offline learning lower bound (Theorem B.6.2), which itself is proven by embedding many stochastic multi-armed bandits problems in a family of hard MDPs. Our construction is inspired by the MDPs in [65] and a personal communication with Christopher Dann but involve substantial modifications to account for the differences in the assumption about rewards. The part in which we obtain explicit dependence on d_m is new and it certifies that the offline learning (and thus uniform OPE) problem strictly more difficult than their online counterpart.

On optimality. The above result provides the minimax lower bound of complexity $\Omega(H^3/d_m\epsilon^2)$. As a comparison, Theorem 3.5.2 gives $\widetilde{O}(H^3S/d_m\epsilon^2)$ is a factor of S away from the lower bound and Theorem 3.5.3 has the same rate of the lower bound up to logarithmic factor.

3.6 Main Results for Offline Learning

In this section we discuss the implication of our results on offline learning. As we discussed earlier in the introduction, a uniform OPE bound of ϵ implies that the corresponding ERM algorithm finds a 2ϵ-suboptimal policy. But it also implies that all other offline policy-learning algorithms that are not ERM, we could gracefully decompose their error into optimization error and statistical (generalization) error.

Theorem 3.6.1. *Let $\hat{\pi}^* = \operatorname{argmax}_\pi \hat{v}^\pi$ — the empirically optimal policy. Let $\hat{\pi}$ be any data-dependent choice of policy such that $\hat{v}^{\hat{\pi}^*} - \hat{v}^{\hat{\pi}} \le \epsilon_{opt}$, then. There is a universal constant c such that w.p. $\ge 1-\delta$*

1. $v^{\pi^*} - v^{\hat{\pi}} \le c\sqrt{\frac{H^4 S \log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$.
2. *If $\delta < e^{-S}$, the bound improves to $c\sqrt{\frac{H^4 S \log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$. And if in addition $\hat{\pi}$ is deterministic, the bound further improves to $c\sqrt{\frac{H^3 \min\{H,S\} \log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$.*

3. *If* $\epsilon_{opt} \le \sqrt{H}/S$ *and that* $||\widehat{V}_t^{\hat{\pi}} - \widehat{V}_t^{\hat{\pi}^\star}||_\infty \le \epsilon_{opt}$, $\forall t = 1, ..., H$, *then* $v^{\pi^*} - v^{\hat{\pi}} \le c\sqrt{\frac{H^3 \log(HSA/\delta)}{d_m \cdot n}} + \epsilon_{opt}$.

Table 3.1: A comparison of related offline policy learning results.

Method/Analysis	Setting	Guarantee	Sample complexity[b]
[61]	Generative model	$\epsilon + O(\epsilon_{\text{opt}}/(1-\gamma))$-optimal	$\widetilde{O}(SA/(1-\gamma)^3\epsilon^2)$
[3, 4]	∞-horizon offline	ϵ-optimal policy	$\widetilde{O}((1-\gamma)^{-6}C_\mu/\epsilon^2)$
[5]	∞-horizon offline	ϵ-optimal policy	$\widetilde{O}((1-\gamma)^{-4}C_\mu/\epsilon^2)$
SIMPLEX for exact empirical optimal[a]	H-horizon offline	ϵ-optimal policy	$\widetilde{O}(H^3/d_m\epsilon^2)$
PI/VI for ϵ_{opt}-empirical optimal	H-horizon offline	$(\epsilon + \epsilon_{\text{opt}})$-optimal policy	$\widetilde{O}(H^3/d_m\epsilon^2)$
Minimax lower bound (Theorem B.6.2)	H-horizon offline	over class $\mathcal{M}_{d_m}$	$\Omega(H^3/d_m\epsilon^2)$

[a] PI/VI or SIMPLEX is not essential and can be replaced by any efficient empirical MDP solver.
[b] *Episode* complexity in H-horizon setting is comparable to *step* complexity in ∞-horizon setting because our finite-horizon MDP is *time-inhomogeneous*. Informally, we can just take $(1-\gamma)^{-1} \asymp H$ and $C_\mu \asymp 1/d_m$.

The third statement implies that all sufficiently accurate planning algorithms based on the empirically estimated MDP are optimal. For example, we can run value iteration or policy iteration to the point that $\epsilon_{\text{opt}} \le O(H^3/nd_m)$.

Comparison to existing work. Previously no algorithm is known to achieve the optimal sample complexity in the offline setting. Our result also applies to the related generative model setting by replacing $1/d_m$ with SA, which avoids the data-splitting procedure usually encountered by specific algorithm design [42]. The analogous policy-learning results In the generative model setting [61, Theorem 1] , achieves a suboptimality of $\tilde{O}((1-\gamma)^{-3}SA/n + (1-\gamma)^{-1}\epsilon_{\text{opt}})$ with no additional assumption on ϵ_{opt}. Informally, if we replace $(1-\gamma)^{-1}$ with H, then our result improves the bound from $H\epsilon_{\text{opt}}$ to just ϵ_{opt} for $\epsilon_{\text{opt}} \le \sqrt{H}/S$. These results are summarized in Table 3.1.

Sparse MDP estimate. We highlight that the result does not require the estimated MDP to be an accurate approximation in any sense. Recall that the true MDP has $O(S^2)$ parameters (ignoring the dependence on H, A and logarithmic terms), but our result is valid provided that $n = \tilde{\Omega}(1/d_m)$ which is $\Omega(S)$. This suggests that we may not even exhaustively visit all pairs to state-transitions and that the estimator of $\hat{P}_t$ is allowed to be zero in many coordinates.

Optimal computational complexity. Lastly, from the computational perspective, we can leverage the best existing solutions for solving optimization $\widehat{\pi}^\star := \operatorname{argmax}_{\pi\in\Pi}\widehat{v}^\pi$. For example, with $\epsilon_{\text{opt}} > 0$, as explained by [61], value iteration ends in $O(H \log \epsilon_{\text{opt}}^{-1})$ iteration and takes at most $O(HSA)$ time after the model has been estimated with one pass of the data ($O(nH)$ time). We have a total computational complexity of $O(H^4/(d_m\epsilon^2) + H^2SA\log(1/\epsilon))$ time algorithm for obtaining the ϵ-suboptimal policy using $n = O(H^4/(d_m\epsilon^2)$ episodes. This is essentially optimal because the leading term H^4SA/ϵ^2 is required even to just process the data needed for the result to be information-theoretically possible. In comparison, the algorithm that obtains an exact empirical optimal policy $\widehat{\pi}^\star$, the SIMPLEX policy iteration runs in time $O(\text{poly}(H, S, A, n))$ [66].

3.7 An Overview of the Proof

Our uniform convergence analysis in Section 3.5.1, relies on creating an unbiased version of $\widehat{v}_{\text{OPEMA}}$ (which we call it $\widetilde{v}_{\text{OPEMA}}$) artificially and use concentration to guarantee $\widehat{v}_{\text{OPEMA}}$ is identical to $\widetilde{v}_{\text{OPEMA}}$ in most situations. By doing so we can reduce our analysis from $\sup_{\pi\in\Pi}\left|\widehat{v}^\pi - v^\pi\right|$ to $\sup_{\pi\in\Pi}\left|\widetilde{v}^\pi - v^\pi\right|$. Specifically, $\widetilde{v}^\pi$ replaces $\widehat{P}_t$, $\widehat{r}_t$ in $\widehat{v}^\pi$ by its fictitious counterparts $\widetilde{P}_t$, $\widetilde{r}_t$, defined as:

$$\widetilde{r}_t(s_t, a_t) = \widehat{r}_t(s_t, a_t)\mathbf{1}(E_t) + r_t(s_t, a_t)\mathbf{1}(E_t^c),$$
$$\widetilde{P}_{t+1}(\cdot|s_t, a_t) = \widehat{P}_{t+1}(\cdot|s_t, a_t)\mathbf{1}(E_t) + P_{t+1}(\cdot|s_t, a_t)\mathbf{1}(E_t^c).$$

where E_t denotes the event $\{n_{s_t,a_t} \geq nd_t^\mu(s_t, a_t)/2\}$. This is saying, if observation n_{s_t,a_t} is large enough (E_t is true), we use $\widehat{P}$; otherwise we directly use P instead. This track helps dealing with out-of-sample state-action pairs. The next key is the martingale decomposition (Lemma 3.5.1). On one hand, by using the structure of $\sup_{\pi\in\Pi}\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle$ we can relax it into a "Rademacher-type complexity" which corresponds to $\tilde{O}(\sqrt{H^4S/d_m n})$ term in Theo-

rem 3.5.1. On the other hand, this decomposition has a natural martingale structure so martingale concentration inequalities can be appropriately applied, *i.e.* Theorem 3.5.2. In addition, each term $\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi \rangle$ separates the non-stationary policy into two parts with empirical distribution only depends on $\pi_{1:h-1}$ that governs how the data "roll in" and the long term value function V_h^π only depends on $\pi_{h:H}$ that governs how the reward "roll out".

For local uniform convergence, by Bellman equations we can obtain a similar decomposition on Q-function:

$$\widehat{Q}_t^\pi - Q_t^\pi = \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi (\widehat{P}_h - P_h)\widehat{V}_h^\pi,$$

where $\Gamma_{t:h}^\pi = \prod_{i=t}^{h} P_i^\pi$ is the multi-step state-action transition and $\Gamma_{t+1:t}^\pi := I$. Since π is any policy in Π_1 which may dependent on $\mathcal{D}'$ so we cannot directly apply concentration inequalities on $(\widehat{P}_h - P_h)\widehat{V}_h^\pi$. Instead, we overcome this hurdle by doing concentration on $(\widehat{P}_h - P_h)\widehat{V}_h^{\widehat{\pi}^*}$ since $\widehat{V}_h^{\widehat{\pi}^*}$ and $\widehat{P}_h$ are independent, and we connect $\widehat{V}_h^{\widehat{\pi}^*}$ back to $\widehat{V}_h^\pi$ by using they are ϵ_{opt} close (Theorem 3.5.3). This idea helps avoiding the technicality of absorbing MDP used in [61] for infinite horizon case because of our non-stationary transition setting. For the uniform convergence lower bound, our analysis relies on reducing the problem to identifying ϵ-optimal policy and proving any algorithm that learns a ϵ-optimal policy requires at least $\Omega(H^3/d_m\epsilon^2)$ episodes in the non-stationary episodic setting. Previously, [65] proves the $\Omega(HSA/\epsilon^2)$ lower bound with assumption $\sum_{i=1}^{H} r_i \leq 1$. Our proof uses a modified version of their hard-to-learn MDP instance to achieve the desired result. To produce extra H^2 dependence, we leverage the Assumption 3.3.1 that $\sum_{i=1}^{H} r_i$ may be of order $O(H)$. We only present the high-level ideas here due the space constraint, detailed proofs are in Appendix B.

3.8 Numerical Simulations

In this section we use a simple simulated environment to empirically demonstrate the correct scaling in H. Direct evaluating $\sup_{\pi\in\Pi} |\widehat{v}^{\pi} - v^{\pi}|$ empirically is computationally infeasible since the policy classes we considered here contains either A^{HS} or ∞ many policies. Instead, in the experiment we will plot the sub-optimality gap $|v^{\star} - v^{\widehat{\pi}^{\star}}|$ with $\hat{\pi}^{\star}$ being the outputs of policy planning algorithms. The sub-optimality gap is considered as a surrogate for the lower bound of $\sup_{\pi\in\Pi} |\widehat{v}^{\pi} - v^{\pi}|$. Concretely, the non-stationary MDP has 2 states s_0, s_1 and 2 actions a_1, a_2 where action a_1 has probability 1 going back the current state and for action a_2, there is one state s.t. after choosing a_2 the dynamic transitions to both states with equal probability $\frac{1}{2}$ and the other one has asymmetric probability assignment ($\frac{1}{4}$ and $\frac{3}{4}$). The transition after choosing a_2 is changing over different time steps therefore the MDP is non-stationary and the change is decided by a sequence of pseudo-random numbers (Figure 3.1 shows the transition kernel at a particular time step). Moreover, to make the learning problem non-trivial we use non-stationary rewards with 4 categories, *i.e.* $r_t(s, a) \in \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\}$ and assignment of $r_t(s, a)$ for each value is changing over time (see Section B in appendix for more details). Lastly, the logging policy in Figure 3.2 is uniform with $\mu_t(a_1|s) = \mu_t(a_2|s) = \frac{1}{2}$ for both states.

Figure 3.2 use a fixed number of episodes $n = 2048$ while varying H to examine the horizon dependence for uniform OPE. We can see for fixed pointwise OPE with OPEMA (blue line), $|v^{\pi} - \widehat{v}^{\pi}|$ scales as $O(\sqrt{H^2})$ which reflects the bound of Lemma 3.5.2; for the model-based planning, we ran both VI and PI until they converge to the empirical optimal policy $\widehat{\pi}^{\star}$. The figure shows that for this MDP example $|v^{\star} - v^{\widehat{\pi}^{\star}}|$ scales as $O(\sqrt{H^3/d_m})$ for fixed n since it is parallel to the reference magenta line. This fact empirically shows $O(\sqrt{H^3/d_m})$ bound is required confirms the scaling of our theoretical results.

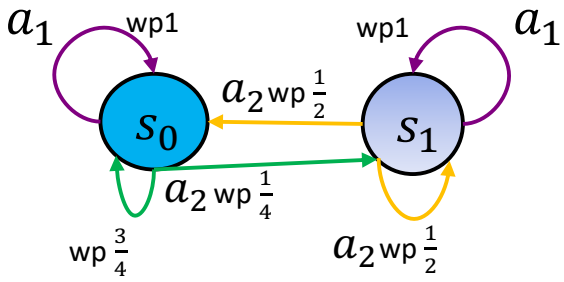

Figure 3.1: A non-stationary MDP

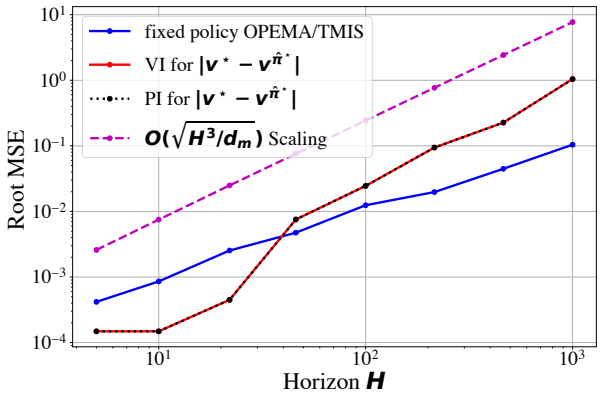

Figure 3.2: Log-log plot showing the dependence on horizon of uniform OPE and pointwise OPE via learning ($|v^\star - v^{\hat{\pi}^\star}|$)

3.9 Discussion

The efficiency of model-based methods. There had been a long-lasting debate about model-based vs model-free methods in RL. The model-based methods were considered inefficient in both space and sample complexity, due to the need to represents the transition kernel in $O(HS^2A)$. Most sample-efficient methods with the right dependence in S are model-free methods that directly represents and updates the Q-function. Our analysis reveals that direct model-based plug-in estimator is optimal in both pointwise and uniform prediction problems, which helps to correct the commonly held misunderstanding that purely model plug-in estimator is loose due to simulation lemma.

Simulation Lemma. Our result can be viewed as a strengthened version of the *simulation lemma* [40] (see also the exposition in [41], which uses similar notations to us). The OPE bound that can be obtained by applying the simulation lemma is

$$|\widehat{v}^\pi - v^\pi| \le H^2 \sup_{t,s_t,a_t} \left\| \widehat{P}(\cdot|s_t,a_t) - P(\cdot|s_t,a_t) \right\|_1 \le \widetilde{O}\left(\sqrt{\frac{H^4 S^2}{n d_m}}\right)$$

which implies an episode complexity[4] of $\widetilde{O}(H^4S^2/d_m\epsilon^2)$. The main limitation of the simulation lemma is that it does not distinguish between pointwise / uniform convergence (and their bound is in fact a uniform OPE bound), thus will suffer from a loose bound when applied to fixed policies or data-dependent policies that qualify for the smaller policy classes that we considered. For example, our Lemma 3.5.2 shows that for the same plug-in estimator, the bound improves to $\widetilde{O}(H^2/d_m\epsilon^2)$ for pointwise OPE and Theorem 3.5.3 shows that we can knock out a factor of HS^2 in the uniform convergence of *near empirically optimal* policies. Finally, there is a factor of Simprovement in the global policy class unconditionally. These savings can be used as drop-in replacements to many instances where the simulation lemma is applied to improve the parameters of the analysis therein.

This chapter represents the first systematic study of uniform convergence in offline policy evaluation. We derive near optimal results for three representative policy classes. By viewing offline policy evaluation from the uniform convergence perspective, we are able to unify two central topics in offline RL, OPE and offline learning while establishing optimal rates in a subset of these settings including the first rate-optimal offline reinforcement learning method. The work focuses on the episodic tabular MDP with nonstationary transitions. Carrying out the same analysis for the stationary transition case, infinite horizon case, as well as the linear MDP setting is highly tractable with the techniques presented. Formalizing these is left as a future direction of work. More generally, a natural complexity measure for the policy class of RL remains elusive. We hope this work would inspire a more general statistical learning theory for RL in the near future.

[4]See Section B for more calculation details.

Chapter 4

Optimal Uniform Offline Policy Evaluation in Time-Homogeneous Tabular MDPs

4.1 Introduction

In the last chapter, which is the work in [7] initiates studies for offline RL from the new perspective of *uniform convergence* in OPE (uniform OPE for short) which unifies OPE and offline learning tasks. Generally speaking, given a policy class Π and offline data with n episodes, uniform OPE seeks to come up with OPE estimators $\widehat{V}_1^\pi$ and $\widehat{Q}_1^\pi$ that satisfy with high probability

$$\sup_{\pi\in\Pi} ||\widehat{Q}_1^\pi - Q_1^\pi||_\infty < \epsilon. \tag{4.1}$$

The task is to achieve (4.1) with the optimal episode complexity: the "minimal" number of episodes n needed as a function of ϵ, failure probability δ, the parameters of the MDP as well as the behavior policy μ in the minimax sense.

To further motivate the readers why uniform OPE is important, we state its relation to offline learning. Indeed, uniform OPE in RL is analogous to uniform convergence of empirical risk in statistical learning theory [67]. In supervised learning, it has been proven that almost all learnable problems are learned by an (asymptotic) *empirical risk minimizer* (ERM) [55].

In offline RL, the natural counterpart would be the *empirical optimal policy* $\widehat{\pi}^\star := \operatorname{argmax}_\pi \widehat{V}_1^\pi$ and with uniform OPE it further ensures $\widehat{\pi}^\star$ is a near optimal policy for offline learning task via (element-wise):

$$\begin{aligned} 0 \le Q_1^{\pi^\star} - Q_1^{\widehat{\pi}^\star} &= Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star} + \widehat{Q}_1^{\pi^\star} - \widehat{Q}_1^{\widehat{\pi}^\star} + \widehat{Q}_1^{\widehat{\pi}^\star} - Q_1^{\widehat{\pi}^\star} \\ &\le |Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star}| + |\widehat{Q}_1^{\widehat{\pi}^\star} - Q_1^{\widehat{\pi}^\star}| \le 2 \sup_\pi |Q_1^\pi - \widehat{Q}_1^\pi|. \end{aligned} \tag{4.2}$$

On the *policy evaluation* side, there is often a need to evaluate the performance of a *data-dependent* policy. Uniform OPE suffices for this purpose since it will allow us to evaluate policies selected by safe-policy improvements, proximal policy optimization, UCB-style exploration-bonus as well as any heuristic exploration criteria (for further discussion and motivation, we refer to [7] and the references therein).

In this chapter, we study the uniform OPE problem under the *finite horizon episodic MDP with stationary transitions* and focus on the model-based approaches. Specifically, we consider two representative class: global policy class Π_g (contains all (deterministic) policies) and local policy class Π_l (contains policies near the empirical optimal one). We ask the following question:

What is the statistical limit for global/local uniform OPE

and what is its connection to optimal offline learning?

We answer the first part by showing global uniform OPE requires a lower bound of $\Omega(H^2 S/d_m\epsilon^2)$[1]

[1]Here d_m is the minimal marginal state-action occupancy, see Assumption 4.4.1.

for the family of model-based approach and answer the second part by showing the model-based offline plug-in estimator for local uniform OPE achieves $\tilde{O}(H^2/d_m\epsilon^2)$ minimax rate and it implies optimal offline learning. Importantly, the procedure of the model-based approach via learning $\widehat{\pi}^\star$ through planning over the empirical MDP has a wider range of use in offline setting as it naturally adapts to the new challenging tasks like *offline task-agnostic learning* and *offline reward-free learning*. Specifically, we have the following contributions, stated as Theorems and Corollaries in this chapter.

4.1.1 Optimal local uniform OPE

First and foremost, we derive the $\tilde{O}(H^2/d_m\epsilon^2)$ optimal episode complexity for local uniform OPE (Theorem 4.6.1) via the model-based method and this implies optimal offline learning with the same rate (Corollary 4.6.1); this result strictly improves upon the Theorem 3.7 in [7] ($\tilde{O}(H^3/d_m\epsilon^2)$) by a factor H in a non-trivial way through our new *singleton-absorbing MDP* technique.

4.1.2 Information-theoretical characterization of the global uniform OPE

We explicitly characterize the statistical limit for the global uniform convergence by proving global uniform OPE has minimax lower bound $\Omega(H^2S/d_m\epsilon^2)$ (over the family of all model-based approaches) (Theorem 4.5.1). This result answers the question left open in [7] that the global uniform OPE is generically harder than local uniform OPE / offline learning due to the required additional dependence on S, and such a difference will be dominant when the state space is large.

4.1.3 Generalize to the new offline settings

Critically, our model-based frameworks naturally generalize to the more challenging settings like task-agnostic and reward-free settings. In particular, we establish the $\tilde{O}(H^2 \log(K)/d_m\epsilon^2)$ (Theorem 4.7.1) and $\tilde{O}(H^2S/d_m\epsilon^2)$ (Theorem 4.7.2) complexity for *offline task-agnostic learning* and *offline reward-free learning* respectively. Both results are new and optimal.

4.1.4 Singleton-absorbing MDP: a sharp analysis tool for episodic stationary transition case

On the technical end, our major contribution is the novel design of *singleton-absorbing MDP* which handles the data-dependence hurdle encountered in the stationary transition setting. To decouple the data-dependence between $\widehat{P}_{s,a} - P_{s,a}$ and $\widehat{V}$, the traditional s-absorbing MDP proposed in [61] uses s-absorbing MDP $\widehat{V}_s$ (in lieu of $\widehat{V}$) for each state to recover the independence. Further, to control the error propagation between $\widehat{V}_s$ and $\widehat{V}$, standard ϵ-net covering were used such that the value of $\widehat{V}_s$ traverse the evenly-spaced grid over $[0, (1-\gamma)^{-1}]$ in their infinite horizon setting. However, when applied to finite horizon case, there are H different quantities $(V_1, ..., V_H)$ and the covering argument need to cover H-dimensional space $[0, H]^H$. This result in a exponential-H covering number and the metric entropy blows up by a factor H which makes sample complexity suboptimal. In contrast, the *singleton-absorbing MDP* technique designs a single absorbing MDP that can also control the error propagation sufficiently well. This sharp analysis tool negates the conjecture of [68] that absorbing MDP is not well suitable for finite horizon stationary MDP (Section 4.6.3).

4.1.5 Significance: Unifying different offline settings

In addition to studying the statistical limit of uniform OPE, this work solves the sample optimality problems for local uniform OPE (Theorem 4.6.1), offline task-agnostic (Theorem 4.7.1) and offline reward-free (Theorem 4.7.2) problems. If we take a deeper look, the algorithmic frameworks utilized are all based on the model-based empirical MDP construction and planning. Therefore, as long as we can analyze such framework sharply (*e.g.* via novel absorbing-MDP technique), then it is hopeful that our techniques can be generalized to tackle more sophisticated settings. On the other hand, things could be more tricky for online RL since the exploration phases need to be specifically designed for each settings and there may not be one general algorithmic pattern that dominates. Our findings reveal the model-based framework is fundamental for offline RL as it subsumes settings like local uniform OPE, offline task-agnostic and offline reward-free learning into the identical learning pattern. Considering these tasks were originally proposed in the online regime under different contexts, such a unified view from the model-based perspective offers a new angle for understanding offline RL.

4.2 Related Literature

Offline reinforcement learning. Information-theoretical considerations for offline RL are first proposed for *infinite horizon discounted setting* via Fitted Q-Iteration (FQI) type function approximation algorithms [4, 3, 6, 5] which can be traced back to [58, 69, 70, 71]. Later, [6] considered the offline RL under only the *realizability* assumption and [59] considers the offline RL *without good exploration*. Those are all challenging problems but with they only provide sub-optimal polynomial complexity in terms of $(1-\gamma)^{-1}$.

For the finite horizon case, [7] first achieves $\tilde{O}(H^3/d_m\epsilon^2)$ complexity under non-stationary transition but their results cannot be further improved in the stationary setting. Concurrent

with our work, a recently released work [8] designs the offline variance reduction algorithm for achieving the optimal $\tilde{O}(H^2/d_m\epsilon^2)$ rate. Their result is for a specific algorithm that uses data splitting while our results work for any algorithms that returns a nearly empirically optimal policy via a uniform convergence guarantee. Our results on the offline task-agnostic and the reward-free settings are entirely new. Another concurrent work [9] considers the horizon-free setting but does not provide uniform convergence guarantee. Even more recently, [10] considers the single concentrability coefficient $C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^\mu(s,a)}$ and obtains the sample complexity $\tilde{O}[(1-\gamma)^{-5}SC^\star/\epsilon^2]$.

In the linear MDP case, [72] studies the pessimism-based algorithms for offline policy optimization under the weak compliance assumption and [73, 74] provide some negative results (exponential lower bound) for offline RL with linear MDP structure.

Model-based approaches with minimaxity. It is known model-based methods are minimax-optimal for online RL with regret $\tilde{O}(\sqrt{HSAT})$ (*e.g.* [39, 75]). For linear MDP, In the generative model setting, [61] shows model-based approach is still minimax optimal $\tilde{O}((1-\gamma)^{-3}SA/\epsilon^2)$ by using a s-absorbing MDP construction and this model-based technique is later reused for other more general settings (*e.g.* Markov games [76] and linear MDPs [68]) and also for improving the sample size barrier [77]. In offline RL, [78, 79] use model-based approaches for continuous policy optimization and [7] uses the model-based methods to achieve $\tilde{O}(H^3/d_m\epsilon^2)$ complexity.

Task-agnostic and Reward-free problems. The reward-free problem is initiated in the online RL [80] where the agent needs to efficiently explore an MDP environment *without* using any reward information. It requires high probability guarantee for learning optimal policy for *any* reward function, which is strictly stronger than the standard learning task that one only needs to learn to optimal policy for a fixed reward. Later, [81, 82] establish the $\tilde{O}(H^3S^2A/\epsilon^2)$ complexity and [83] further tightens the dependence to $\tilde{O}(H^2S^2A/\epsilon^2)$.[2] Recently, [84] proposes

[2]We translate [83] their dimension-free result to $\tilde{O}(H^2S^2A/\epsilon^2)$ under the standard assumption $r \in [0,1]$.

the task-agnostic setting where one needs to use exploration data to simultaneously learn K tasks and provides a upper bound with complexity $\tilde{O}(H^5SA\log(K)/\epsilon^2)$. For linear MDP setting, [85] achieves the sample complexity $\tilde{O}(d^3H^6/\epsilon^2)$ and [86] considers such problem in the online two-player Markov game. However, although these settings remain critical in the offline regime, no statistical result has been formally derived so far.

4.3 The Setup for Time-Homogeneous MDPs

Episodic time-homogeneous reinforcement learning. A finite-horizon *Markov Decision Process* (MDP) is denoted by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, H, d_1)$, which is identical to the Definition in Section 3.3. The only exception is the time-homogeneous transition kernel that has the form $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ with $P(s'|s, a)$ representing the probability transition from state s, action a to next state s'. This is different from non-stationary setting where P_t can change across different times. Besides, $r : \mathcal{S} \times A \mapsto \mathbb{R}$ is the expected reward function and given (s, a) which satisfies $0 \le r \le 1$.[3] A policy is formulated by $\pi = (\pi_1, ..., \pi_H)$, where each π_h assigns each state $s \in \mathcal{S}$ a probability distribution over actions, *i.e.* $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ (where $\Delta(\mathcal{A})$ is the set of probability distributions over the actions) $\forall h \in [H]$. We emphasize although transition P is stationary, policy $\pi := \pi_{1:H}$ itself can be non-stationary. An MDP together with a policy π induces a random trajectory $s_1, a_1, r_1, \dots, s_H, a_H, r_H, s_{H+1}$ with the following data generating process: $s_1 \sim d_1, a_t \sim \pi(\cdot|s_t), r_t = r(s_t, a_t), s_{t+1} \sim P(\cdot|s_t, a_t), \forall t \in [H]$.

[3] Note here we assume mean reward function is known. It is widely-known that the randomness in the reward has lower order influence on the error than the randomness in the transition P in RL.

Q-values and Bellman (optimality) equations. For any policy π and any $h \in [H]$, we incorporate standard value function $V_h^\pi(\cdot) \in \mathbb{R}^S$ and Q-value function $Q_h^\pi(\cdot,\cdot) \in \mathbb{R}^{S\times A}$ as:

$$V_h^\pi(s) = \mathbb{E}_\pi[\sum_{t=h}^{H} r_t | s_h = s], \quad Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^{H} r_t | s_h, a_h = s, a], \quad \forall s, a \in \mathcal{S}, \mathcal{A}.$$

We always enumerate V_h^π, Q_h^π as column vectors and the (s,a)-th row of P as the row vector $P(\cdot|s,a)$, then Bellman (optimality) equation follows $\forall h \in [H]$:

$$\begin{aligned} Q_h^\pi(s,a) &= r(s,a) + P(\cdot|s,a)V_{h+1}^\pi, \quad V_h^\pi = \mathbb{E}_{a\sim\pi_h}[Q_h^\pi] \\ Q_h^\star(s,a) &= r(s,a) + P(\cdot|s,a)V_{h+1}^\star, \; V_h^\star = \max_a Q_h^\star(\cdot,a) \end{aligned}$$

The goal of RL is to find a policy $\pi^\star$ such that $v^\pi := \mathbb{E}_\pi\left[\sum_{t=1}^H r_t\right]$ is maximized, which is equivalent to simultaneously maximize $V_1^\pi(s)$ (or $Q_1^\pi(s,a)$) for all s (or s,a) [12]. Therefore, for a targeted accuracy $\epsilon > 0$ it suffices to find a policy π_{alg} such that $\left\|Q_1^\star - Q_1^{\pi_{\text{alg}}}\right\|_\infty \le \epsilon$.

Additional notations. We denote the per-step marginal state-action occupancy $d_t^\pi(s,a)$ as:

$$d_t^\pi(s,a) := \mathbb{P}[s_t = s | s_1 \sim d_1, \pi] \cdot \pi_t(a|s), \tag{4.3}$$

which is the marginal state-action probability at time t. Moreover, we define state-action transition matrix $P^{\pi_h} \in \mathbb{R}^{SA\times SA}$ with $P^{\pi_h}_{(s,a),(s',a')} = P(s'|s,a)\pi_h(a'|s')$, note π_h is indexed by h since policy $\pi_{1:H}$ can be non-stationary.

Offline setting. The offline RL assumes that episodes $\mathcal{D} = \left\{\left(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)}\right)\right\}_{i\in[n]}^{t\in[H]}$ are rolling from some behavior policy μ a priori. In particular, we cannot change μ and do not assume the functional knowledge of μ.

Model-based RL. We focus our attention on model-based methods, which has witnessed numerous successes (both theoretically and empirically) and is one of the most critical components of theoretical RL as a whole (as introduced in Section A.1). To make the presentation precise, we define the following:

Definition 4.3.1. *Model-based RL: Solving RL problems (either learning or evaluation) through learning / modeling transition dynamic P.*

We emphasize that the model-based approaches in general (*e.g.* [87, 88, 79]) follow the procedure of modeling the full MDP $M = (S, \mathcal{A}, P, r, H, d_1)$ instead of only the transition P. Nevertheless, we (by convention) assume the mean reward function is known and the initial state distribution d_1 will not affect the choice of optimal policy $\pi^\star$. Thus, Definition 4.3.1 suffices for our purposes.

4.4 Uniform convergence in offline RL Recap

We study offline RL from uniform convergence offline policy evaluation (uniform OPE) perspective. Concretely, uniform OPE extends the point-wise (fixed target policy) OPE to a family of policies Π. The goal is to construct estimator $\widehat{Q}_1^\pi$ such that $\sup_{\pi\in\Pi} \left\| Q_1^\pi - \widehat{Q}_1^\pi \right\| < \epsilon$, which automatically ensures point-wise OPE for any $\pi \in \Pi$. More importantly, uniform OPE directly implies offline learning when Π contains optimal policies. Let $\widehat{\pi}^\star := \operatorname{argmax}_\pi \widehat{V}_1^\pi$ be the *empirical optimal policy* for some OPE estimator $\widehat{v}^\pi$, then by (4.2) $\widehat{\pi}^\star$ is a near-optimal policy given uniform OPE guarantee.

We consider the following two policy classes that are of the interests and explain why they should be considered.

Definition 4.4.1 (The global (deterministic) policy class.)**.** *The global policy class Π_g consists of all the non-stationary (deterministic) policies.*

It is well-known [12] there exists at least one deterministic policy that is optimal, therefore Π_g is sufficiently rich for evaluating algorithms that aim at learning the optimal policy.

Definition 4.4.2 (The local policy class)**.** *Given empirical MDP $\widehat{M}$ and $\widehat{V}_h^\pi$ is the value under $\widehat{M}$. Let $\widehat{\pi}^\star := \text{argmax}_\pi \widehat{V}_1^\pi$ be the empirical optimal policy, then the local policy class Π_l is defined as:*

$$\Pi_l := \left\{ \pi : \ s.t. \ \left\| \widehat{V}_h^\pi - \widehat{V}_h^{\widehat{\pi}^\star} \right\|_\infty \leq \epsilon_{opt}, \forall h \in [H] \right\}$$

where $\epsilon_{opt} \geq 0$ is a parameter. This class characterizes all the policies in the neighborhood of empirical optimal policy.

In above $\widehat{M}$ transition kernel uses $\widehat{P}$ in lieu of P where $\widehat{P}(s'|s,a) = \frac{n_{s',s,a}}{n_{s,a}}$ if $n_{s,a} > 0$ and $1/S$ otherwise.[4] Moreover, once given $\widehat{P}$, it is efficient to obtain $\widehat{\pi}^\star$ using Value/Policy Iteration, therefore it is more practical to consider the neighborhood of $\widehat{\pi}^\star$ (instead of $\pi^\star$) since practitioners can use data $\mathcal{D}$ to really check Π_l whenever needed.

Next we present the regularity assumption required for uniform convergence OPE problem.

Assumption 4.4.1 (Exploration requirement)**.** *Logging policy μ obeys that $\min_s d_t^\mu(s) > 0$, for any state s that is "accessible". Moreover, we define the quantity $d_m := \min\{d_t^\mu(s,a) : d_t^\mu(s,a) > 0\}$ (recall $d_t^\mu(s,a)$ in (4.3)) to be the minimal marginal state-action probability.*

This is identical to 4.4.1. State s is "accessible" means there exists a policy π so that $d^\pi(s) > 0$. If for any policy π we always have $d^\pi(s) = 0$, then state s can never be visited in the given MDP. Assumption 4.4.1 says μ have the right to explore all "accessible" states. Assumption 4.4.1 is the minimal assumption needed for the consistency of uniform OPE task and is similar to *concentrability coefficient* data coverage assumption [58] made for function approximation learning. This assumption can be potentially relaxed for pure offline learning problems, *e.g.* [63, 10], where they only require $d^\mu(s)(d^\mu(s,a)) > 0$ for any state s (s,a) satisfies $d^{\pi^\star}(s)(d^{\pi^\star}(s,a)) > 0$.

[4]Here $n_{s,a}$ is the number of pair (s,a) being visited among n episodes. $n_{s',s,a}$ is defined similarly.

4.5 Statistical Hardness for Model-based Global Uniform OPE

From (4.2) and Definition 4.4.1, it is clear the global uniform OPE implies offline RL, therefore it is natural to wonder whether they just are *"the same task"* (their sample complexities have the same minimax rates). If this conjecture is true, then deriving sample efficient global OPE method is just as important as deriving efficient offline learning algorithm (plus the additional benefit of evaluating data-dependent algorithms)! [7] proves the $\tilde{O}(H^3S/d_m\epsilon^2)$ upper bound and $\Omega(H^3/d_m\epsilon^2)$ lower bound for global uniform OPE, but it is unclear whether the additional S is essential. We answer the question affirmatively by providing a tight lower bound result with a concise proof to show no model-based algorithm can surpass $\Omega(S/d_m\epsilon^2)$ information-theoretical limit.

Theorem 4.5.1 (Minimax lower bound for global uniform OPE)**.** *Let d_m be a parameter such that $0 < d_m \leq \frac{1}{SA}$. Let the problem class be $\mathcal{M}_{d_m} := \left\{(\mu, M) \mid \min_{t,s_t,a_t} d_t^\mu\left(s_t, a_t\right) \geq d_m\right\}$. Then there exists universal constants $c, C, p > 0$ such that: for any $n \geq cS/d_m \cdot \log(SAp)$,*

$$\inf_{\widehat{Q}_{1,mb}} \sup_{\mathcal{M}_{d_m}} \mathbb{P}_{\mu,M}\left(\sup_{\pi\in\Pi_g}\left\|\widehat{Q}^\pi_{1,mb} - Q^\pi_1\right\|_\infty \geq C\sqrt{\frac{H^2S}{nd_m}}\right) \geq p,$$

where $\widehat{Q}_{1,mb}$ is the output of any model-based algorithm and Π_g is defined in Definition 4.4.1.

By setting $\epsilon := \sqrt{\frac{H^2S}{nd_m}}$, Theorem 4.5.1 establishes the global uniform convergence lower bound of $\Omega(H^2S/d_m\epsilon^2)$ over model-based methods, which builds the hard statistical threshold between the global uniform OPE and the local uniform OPE tasks by a factor of S since the local case has achievable $\tilde{O}(1/d_m\epsilon^2)$ rate on the dependence for state-actions. This result also reveals the global uniform convergence bound in [7] ($\tilde{O}(H^3S/d_m\epsilon^2)$) is essentially minimax rate-optimal for their *non-stationary setting*[5] and complements the story on the optimality be-

[5]To be rigorous, we remark that it is rate-optimal since for the non-stationary setting the dependence for horizon is higher by a factor H.

havior for global uniform OPE. Moreover, from the generative model view the lower bound degenerates to $S/d_m\epsilon^2 \approx \Theta(S^2A/\epsilon^2)$ which is linear in the model size S^2A. This means in order to achieve global uniform convergence any algorithm needs to estimate each coordinate of transition kernel $P(s'|s,a)$ accurately. We now provide a brief sketch of the proof with the full proof being deferred to Appendix C.

Proof: [Proof Sketch] We only explain the case where $H = 2$ in this proof sketch. Our proof relies on the following novel reduction to l_1 density estimation

$$\sup_{\pi\in\Pi_g} \left\|\widehat{Q}_1^\pi - Q_1^\pi\right\|_\infty \geq \sup_{s,a} \frac{1}{2} \left\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\right\|_1$$

and leverages the Minimax rate for estimating discrete distribution under l_1 loss is $O(\sqrt{S/n_{s,a}})$ [89]. Concretely, by Definition 4.3.1, let $\widehat{P}$ be the learned transition by any arbitrary model-based method. Since we assume r is known and by convention $Q_{H+1}^\pi = 0$ for any π, then by Bellman equation

$$\widehat{Q}_h^\pi = r_h + \widehat{P}^{\pi_{h+1}}\widehat{Q}_{h+1}^\pi,\ \forall h \in [H].$$

In particular, $\widehat{Q}_{H+1}^\pi = Q_{H+1}^\pi = 0$, and this implies $\widehat{Q}_H^\pi = Q_H^\pi = r_H$. Now, again by definition of Bellman equation $\widehat{Q}_{H-1}^\pi = r_{H-1} + \widehat{P}^{\pi_H}\widehat{Q}_H^\pi = r_{H-1} + \widehat{P}^{\pi_H} r_H$ and $Q_{H-1}^\pi = r_{H-1} + P^{\pi_H} r_H$, therefore (recall $H = 2$ and note $r_H \in \mathbb{R}^{S\cdot A}, r_H^{\pi_H} \in \mathbb{R}^S$)

$$\begin{aligned}
&\sup_{\pi\in\Pi_g} \left\|\widehat{Q}_{H-1}^\pi - Q_{H-1}^\pi\right\|_\infty = \sup_{\pi\in\Pi_g} \left\|\left(\widehat{P}^{\pi_H} - P^{\pi_H}\right) r_H\right\|_\infty \\
&= \sup_{\pi\in\Pi_g} \left\|\left(\widehat{P} - P\right) r_H^{\pi_H}\right\|_\infty \approx \sup_{r\in\{0,1\}^S} \left\|\left(\widehat{P} - P\right) r\right\|_\infty \\
&\geq \sup_{s,a} \frac{1}{2} \left\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\right\|_1 \geq O(\sqrt{S/n_{s,a}});
\end{aligned}$$

Lastly, using exponential tail bound to obtain $O(\sqrt{S/n_{s,a}}) \gtrsim O(\sqrt{S/nd_m})$ with high probability. See Appendix C for the full proof for the general H.

4.6 Optimal local uniform OPE via model-based plug-in method

Global uniform OPE is intrinsically harder than the offline learning problem due to the additional state-space dependence and such a gap will amplify when S is (exponentially) large. This motivates us to switch to the local uniform convergence regime that enables optimal learning but also has sub-linear state-action size $\tilde{O}(1/d_m)$ in the policy evaluation. [7] Theorem 3.7 first obtains the $\tilde{O}(H^3/d_m\epsilon^2)$ local uniform convergence for Π_l (recall Definition 4.4.2) and also obtains the same rate for the learning task. Unfortunately, their technique cannot further reduces the dependence of H for stationary transition case. In this section we show the model-based plug-in approach ensures optimal local uniform OPE and further implies optimal offline learning with episode complexity $\tilde{O}(H^2/d_m\epsilon^2)$. To this end, we design the new *singleton-absorbing MDP* to handle the challenge in the stationary transition setting, which uses the absorbing MDP with one single H-dimensional reference point and is our major technical contribution. The *singleton-absorbing MDP* technique avoids the exponential H cover used in [68] and answers their conjecture that absorbing MDP is not well suitable for finite horizon stationary MDP.[6]

4.6.1 Model-based Offline Plug-in Estimator

Recall $n_{s,a} := \sum_{i=1}^{n}\sum_{h=1}^{H}\mathbf{1}[s_h^{(i)}, a_h^{(i)} = s, a]$ be the total counts that visit (s, a) pair, then the model-based offline plug-in estimator constructs estimator $\widehat{P}$ as:

$$\widehat{P}(s'|s,a) = \frac{\sum_{i=1}^{n}\sum_{h=1}^{H}\mathbf{1}[(s_{h+1}^{(i)}, a_h^{(i)}, s_h^{(i)}) = (s', s, a)]}{n_{s,a}},$$

if $n_{s,a} > 0$ and $\widehat{P}(s'|s,a) = \frac{1}{S}$ if $n_{s,a} = 0$. As a consequence, the estimators $\widehat{Q}_h^\pi, \widehat{V}_h^\pi$ are computed as:

$$\widehat{Q}_h^\pi = r + \widehat{P}^{\pi_{h+1}}\widehat{Q}_{h+1}^\pi = r + \widehat{P}\widehat{V}_{h+1}^\pi,$$

[6] See their Section 7, first bullet point for a discussion.

with the initial distribution $\widehat{d}_1(s) = n_s/n$. Under the above setting, we can define the empirical Bellman optimality equations (as well as the population version for completeness) as $\forall s \in S, h \in [H]$:

$$V_h^\star(s) = \max_a \left\{ r(s,a) + P(\cdot|s,a)V_{h+1}^\star \right\},$$
$$\widehat{V}_h^\star(s) = \max_a \left\{ r(s,a) + \widehat{P}(\cdot|s,a)\widehat{V}_{h+1}^\star \right\}.$$

Now we can state our local uniform OPE result with this construction.

4.6.2 Main results for local uniform OPE and offline learning

Recall $\widehat{\pi}^\star := \operatorname{argmax}_\pi \widehat{V}_1^\pi$ is the empirical optimal policy and the local policy class $\Pi_l := \{\pi : \text{ s.t. } \left\| \widehat{V}_h^\pi - \widehat{V}_h^{\widehat{\pi}^\star} \right\|_\infty \le \epsilon_{\text{opt}}, \forall h \in [H]\}$.

Theorem 4.6.1 (optimal local uniform OPE)**.** *Let $\epsilon_{opt} \le \sqrt{H/S}$ and denote $\iota = \log(HSA/\delta)$. For any $\delta \in [0,1]$, there exists universal constants c, C such that when $n > cH\cdot\log(HSA/\delta)/d_m$, with probability $1-\delta$,*

$$\sup_{\pi\in\Pi_l} \left\| \widehat{Q}_1^\pi - Q_1^\pi \right\|_\infty \le C \left[\sqrt{\frac{H^2\iota}{nd_m}} + \frac{H^{2.5}S^{0.5}\iota}{nd_m} \right].$$

Theorem 4.6.1 establishes the $\tilde{O}(H^2/d_m\epsilon^2)$ complexity bound and directly implies the upper bound for $\sup_{\pi\in\Pi_l} ||\widehat{V}_1^\pi - V_1^\pi||_\infty$ with the same rate. This result improves the local uniform convergence rate $\tilde{O}(H^3/d_m\epsilon^2)$ in [7] (Theorem 3.7) by a factor of H and is near-minimax optimal (up to the logarithmic factor). Such result is first achieved by our novel *singleton absorbing MDP* technique. We explain this technique in detail in the next section.

On the other hand, characterizing policy class through the distance in value (like Π_l) is more flexible than characterizing the distance between policies themselves (*e.g.* via total variation).

Figure 4.1: Related comparisons of sample complexities for offline RL

Result/Method	Setting	Type	Complexity	Uniform guarantee?
[3]	∞-horizon	FQI variants	$\tilde{O}((1-\gamma)^{-6}\beta_\mu/\epsilon^2)$	No
FQI [4]	∞-horizon	FQI variants	$\tilde{O}((1-\gamma)^{-6}C/\epsilon^2)$	No
MSBO/MABO [5]	∞-horizon	FQI variants	$\widetilde{O}((1-\gamma)^{-4}C_\mu/\epsilon^2)$	No
OPEMA [7]	H-horizon	Non-splitting	$\widetilde{O}(H^3/d_m\epsilon^2)$	$\sqrt{H}/S$-local uniform
OPDVR [8]	H-horizon	Data splitting	$\widetilde{O}(H^2/d_m\epsilon^2)$	No
Model-based Plug-in (Corollary 4.6.1)	H-horizon	Non-splitting	$\widetilde{O}(H^2/d_m\epsilon^2)$	$\sqrt{H/S}$-local uniform
Task-Agnostic (Theorem 4.7.1)	H-horizon	Non-splitting	$\widetilde{O}(H^2\log(K)/d_m\epsilon^2)$	—
Reward-Free (Theorem 4.7.2)	H-horizon	Non-splitting	$\widetilde{O}(H^2S/d_m\epsilon^2)$	—

* K is the number of tasks for Task-agnostic setting and β_μ, C and $1/d_m$ are data coverage parameters that measure the state-action dependence and are qualitative similar under their respective assumptions.

This is because: if two policies are "close", then their values are also similar; but the reverse may not be true since two very different policies could possibly generate similar values. Therefore the consideration of Π_l is generic and conceptually reflects the fundamental principle of RL: as long as two policies yield the same value, they are considered "equally good", no matter how different they are.[7]

Most importantly, Theorem 4.6.1 guarantees near-minimax optimal offline learning:

Corollary 4.6.1 (optimal offline learning)**.** *If* $\epsilon_{opt} \le \sqrt{H/S}$ *and that* $\sup_t ||\widehat{V}_t^{\widehat{\pi}} - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \le \epsilon_{opt}$, *when* $n > O(H \cdot \iota/d_m)$, *then with probability* $1-\delta$, *element-wisely,*

$$V_1^\star - V_1^{\widehat{\pi}} \le C\left[\sqrt{\frac{H^2\iota}{nd_m}} + \frac{H^{2.5}S^{0.5}\iota}{nd_m}\right]\mathbf{1} + \epsilon_{opt}\,\mathbf{1}.$$

Corollary 4.6.1 first establishes the minimax rate for offline learning for any policy $\widehat{\pi}$ with the measurable gap $\epsilon_{\text{opt}} \le \sqrt{H/S}$. This extends the standard concept of offline learning by allowing any empirical planning algorithm (*e.g.* VI/PI) to find an *inexact* $\widehat{\pi}$ as an $(\tilde{O}\sqrt{H^2/nd_m} + \epsilon_{\text{opt}})$-optimal policy (instead of finding exact $\widehat{\pi}^\star$). The use of *inexact* $\widehat{\pi}$ could encourage early stopping (*e.g.* for VI/PI) therefore saves computational iterations. Besides, we leverage full data to construct empirical MDP for planning and, on the contrary, [8] uses data-splitting (split data

[7] We recognize that in the specific settings (*e.g.* safe policy improvement) some of the policies that yield high values are not feasible. These considerations are beyond the scope of this paper.

into mini-batches and only apply each mini-batch at each specific iteration) to enable Variance Reduction technique, which could cause inefficient data use for the practical purpose. By the following lower bound result from [8], our Corollary 4.6.1 is near minimax optimal.

Theorem 4.6.2 (Theorem 4.2. [8]). *Let $\mathcal{M}_{d_m}$ be the same as Theorem 4.5.1. There exists universal constants c_1, c_2, c, p (with $H, S, A \geq c_1$ and $0 < \epsilon < c_2$) such that when $n \leq cH^2/d_m\epsilon^2$,*[8]

$$\inf_{V_1^{\pi_{alg}}} \sup_{(\mu,M)\in\mathcal{M}_{d_m}} \mathbb{P}_{\mu,M}\left(||V_1^\star - V_1^{\pi_{alg}}||_\infty \geq \epsilon\right) \geq p.$$

In the rest of the section, we briefly explain the main ideas needed for the proof by introducing the *singleton-absorbing MDP* technique, and the full proofs of Theorem 4.6.1, Corollary 4.6.1 are given in Appendix C.

4.6.3 Singleton absorbing MDP for finite horizon MDP

For the ease of illustration, we explain our idea via bounding $||\widehat{Q}_h^{\widehat{\pi}^\star} - Q_h^{\widehat{\pi}^\star}||_\infty$ (instead of $\sup_{\pi\in\Pi_l} ||\widehat{Q}_1^\pi - Q_1^\pi||_\infty$) and choose related quantity $\widehat{\pi}^\star$ (instead of $\widehat{\pi}$) and $\widehat{V}_h^\star$ (instead of $\widehat{V}_h^{\widehat{\pi}}$) to discuss. Essentially, the key challenge in obtaining the optimal dependence in stationary setting is the need to decouple the dependence between $P - \widehat{P}$ and $\widehat{V}_h^\star$ as we aggregate all data for constructing both $\widehat{P}$ and $\widehat{V}_h^\star$. This issue is not encountered in the non-stationary setting in general due to the flexibility to estimate different transition P_t at each time [7] and $\widehat{P}_t$ and $\widehat{V}_{t+1}^\star$ preserve conditional independence. However, when confined to stationary case, their complex $\tilde{O}(H^3/d_m\epsilon^2)$ becomes sub-optimal. Moreover, the direct use of s-absorbing MDP in [61] does not yield tight bounds for the finite horizon stationary setting, as it requires s-absorbing MDPs with H-dimensional fine-grid cover to make sure $\widehat{V}_h^\star$ is close to one of the elements in the cover (which has size $\approx H^H$ and it is not optimal [68]). We overcome this hurdle by choosing *only*

[8]The original Theorem uses $v^\star$ but we use $V_1^\star$ here. It does not matter since we can manually add a default state at the beginning of the MDP and obtain the result for our version.

one delicate absorbing MDP to approximate $\widehat{V}_h^\star$ which will not incur additional dependence on horizon H caused by the union bound. We begin with the general definition of absorbing MDP initialized in [61] and then introduce the *singleton absorbing MDP*.

Standard s-absorbing MDP in the finite horizon setting. The general s-absorbing MDP is defined as follows: for a fixed state s and a sequence $\{u_t\}_{t=1}^H$, MDP $M_{s,\{u_t\}_{t=1}^H}$ is identical to M for all states except s, and state s is absorbing in the sense $P_{M_{s,\{u_t\}_{t=1}^H}}(s|s,a) = 1$ for all a, and the instantaneous reward at time t is $r_t(s,a) = u_t$ for all $a \in \mathcal{A}, t \in [H]$. For convenience, we use the shorthand notation $V^\pi_{\{s,u_t\}}$ to denote $V^\pi_{s,M_{s,\{u_t\}_{t=1}^H}}$ and similarly for Q_t, r and transition P. Also, $V^\star_{\{s,u_t\}}$ ($Q^\star_{\{s,u_t\}}$) is the optimal value under $M_{s,\{u_t\}_{t=1}^H}$.

Before defining singleton absorbing MDP, we first present the following Lemma 4.6.1 and Lemma 4.6.2 which support the our design.

Lemma 4.6.1. $V_t^\star(s) - V_{t+1}^\star(s) \geq 0$, $\forall s \in \mathcal{S}, t \in [H]$.

Lemma 4.6.2. *Fix a state s. If we choose $u_t^\star := V_t^\star(s) - V_{t+1}^\star(s)$, then we have the following vector form equation*

$$V^\star_{h,\{s,u_t^\star\}} = V^\star_{h,M} \quad \forall h \in [H].$$

Similarly, if we choose $\hat{u}_t^\star := \widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s)$, then $\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}} = \widehat{V}^\star_{h,M}$, $\forall h \in [H]$.

The proofs are deferred to Appendix C. Note by Lemma 4.6.1 the assignment of $u_t^\star (:= r_{t,\{s,u_t^\star\}})$ is well-defined. Lemma 4.6.2 is crucial since, under the specification of $u_t^\star$, the optimal value in $M_{s,\{u_t^\star\}_{t=1}^H}$ is identical to the optimal value in original M. Based on these, we define the following:

Definition 4.6.1 (Singleton-absorbing MDP). *For each state s, the singleton-absorbing MDP is chosen to be $M_{s,\{u_t^\star\}_{t=1}^H}$, where $u_t^\star := V_t^\star(s) - V_{t+1}^\star(s)$ for all $t \in [H]$.*

Using Definition 4.6.1, for each (s,a) row the term $(\widehat{P}_{s,a} - P_{s,a})\widehat{V}_h^\star$ can be substituted by $(\widehat{P}_{s,a} - P_{s,a})\widehat{V}^\star_{h,\{s,u_t^\star\}}$, where $\widehat{P}_{s,a}$ and $\widehat{V}^\star_{h,\{s,u_t^\star\}}$ are independent by construction and Bernstein concentration applies. Furthermore, by the selection of $u_t^\star$, we can control the error of $||\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}||_\infty$ to have rate $O(\sqrt{\frac{1}{n}})$ which forces the term $(\widehat{P}_{s,a} - P_{s,a})(\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}})$ to have higher order error. These are the critical building blocks for bounding $||\widehat{Q}_h^{\hat{\pi}^\star} - Q_h^{\hat{\pi}^\star}||_\infty$.

Indeed, by Bellman equations we have the decomposition:

$$\widehat{Q}_h^{\hat{\pi}^\star} - Q_h^{\hat{\pi}^\star} = \ldots = \sum_{t=h}^{H} \Gamma^{\hat{\pi}^\star}_{h+1:t}\left(\widehat{P} - P\right)\widehat{V}^\star_{t+1},$$

where $\Gamma^\pi_{h+1:t} = \prod_{i=h+1}^t P^{\pi_i}$ is multi-step state-action transition and $\Gamma_{h+1:h} := I$. Then for each (s,a) row

$$\begin{aligned}
&(\widehat{P}_{s,a} - P_{s,a})\widehat{V}_h^\star \\
=&(\widehat{P}_{s,a} - P_{s,a})(\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}) + (\widehat{P}_{s,a} - P_{s,a})\widehat{V}^\star_{h,\{s,u_t^\star\}} \\
\lesssim&||\widehat{P}_{s,a} - P_{s,a}||_1||\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}||_\infty + \sqrt{\frac{\mathrm{Var}_{s,a}(\widehat{V}^\star_{h,\{s,u_t^\star\}})}{n_{s,a}}} \qquad (4.4)\\
\lesssim&\sqrt{\frac{S}{n_{s,a}}}\left\|\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty + \sqrt{\frac{\mathrm{Var}_{s,a}(\widehat{V}_h^\star)}{n_{s,a}}} \quad (\star)
\end{aligned}$$

where $(\star)$ is the place where the traditional technique uses the union bound over their *exponential large* ϵ-net, which we do not have! Next, by Lemma 4.6.2 and Lemma C.1.2 in Appendix

$$\begin{aligned}
||\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}||_\infty &= ||\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}} - \widehat{V}^\star_{h,\{s,u_t^\star\}}||_\infty \\
\leq H\max_t\left|\hat{u}_t^\star - u_t^\star\right| &\leq 2H\max_t|\widehat{V}_t^\star - V_t^\star|,
\end{aligned}$$

by a crude bound (Lemma D.0.11), $\max_t|\widehat{V}_t^\star - V_t^\star| \lesssim H^2\sqrt{\frac{S}{n_{s,a}}}$ which makes $\sqrt{\frac{1}{n_{s,a}}}||\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}||_\infty$ have order $1/n_{s,a}$. Finally, to reduce the horizon dependence we apply

$\sum_{t=h}^{H} \Gamma^{\pi}_{h+1:t} \sqrt{\mathrm{Var}_{s,a}\left(V^{\pi}_{t+1}\right)} \leq \sqrt{(H-h)^3}$ for any π. This (informally) bounds $\widehat{Q}^{\widehat{\pi}^\star}_h - Q^{\widehat{\pi}^\star}_h$ by

$$||\widehat{Q}^{\widehat{\pi}^\star}_h - Q^{\widehat{\pi}^\star}_h||_\infty \lesssim \sqrt{\frac{H^3}{n_{s,a}}} + \frac{Poly(H,S)}{n_{s,a}}.$$

Lastly, use $\min_{s,a} n_{s,a} \gtrsim H \cdot d_m$ to finish the proof.

Remark 7. *We emphasize the appropriate selection of $M_{s,\{u^\star_t\}_{t=1}^H}$ ($\widehat{M}_{s,\{u^\star_t\}_{t=1}^H}$) is the key for achieving optimality. It guarantees two things: 1. $\widehat{V}^\star_{h,\{s,u^\star_t\}}$ approximates $\widehat{V}^\star_h$ with sufficient accuracy (has rate $\sqrt{1/n_{s,a}}$); 2. it avoids the fine-grid design with exponential union bound in the dominate term ($\sqrt{\frac{\mathrm{Var}_{s,a}(\widehat{V}^\star_h)\log(|U_{s,a}|/\delta)}{N}}$ with $|U_{s,a}|$ to be at least H^H [68].)*

4.7 New Settings: Offline Task-Agnostic and Offline Reward-Free Learning

From Corollary 4.6.1, our model-based offline learning algorithm has two steps: 1. constructing offline empirical MDP $\widehat{M}$ using the offline dataset $\mathcal{D} = \{(s^{(i)}_t, a^{(i)}_t, r(s^{(i)}_t, a^{(i)}_t), s^{(i)}_{t+1})\}^{t\in[H]}_{i\in[n]}$; 2. performing any accurate black-box *planning* algorithm and returning $\widehat{\pi}^\star$(or $\widehat{\pi}$) as the final output. However, the only *effective* data (data that contains stochasticity) is $\mathcal{D}' = \{(s^{(i)}_t, a^{(i)}_t)\}^{t\in[H]}_{i\in[n]}$. This indicates we are essentially using the state-action space exploration data $\mathcal{D}'$ to solve the task-specific problem with reward r. With this perspective in mind, it is natural to ask: given only the offline exploration data $\mathcal{D}'$, can we efficiently learn a set of potentially conflicting K tasks (K rewards) simultaneously? Even more, can we efficiently learn all tasks (any reward) simultaneously? This brings up the following definitions.

Definition 4.7.1 (Offline Task-agnostic Learning)**.** *Given a offline exploration datatset $\mathcal{D}' = \{(s^{(i)}_t, a^{(i)}_t)\}^{t\in[H]}_{i\in[n]}$ by μ with n episodes. Given K tasks with reward $\{r_k\}_{k=1}^K$ and the corresponding*

K MDPs $M_k = (S, \mathcal{A}, P, r_k, H, d_1)$. *Can we use* $\mathcal{D}'$ *to output* $\hat{\pi}_1, \ldots, \hat{\pi}_K$ *such that*

$$\mathbb{P}\left[\forall r_k, k \in [K], \left\|V^{\star}_{1,M_k} - V^{\hat{\pi}_k}_{1,M_k}\right\|_\infty \leq \epsilon\right] \geq 1 - \delta?$$

Definition 4.7.2 (Offline Reward-free Learning)**.** *Given a offline exploration datatset* $\mathcal{D}' = \{(s_t^{(i)}, a_t^{(i)})\}_{i\in[n]}^{t\in[H]}$ *by* μ *with n episodes. For any reward r and the corresponding MDP* $M = (S, \mathcal{A}, P, r, H, d_1)$. *Can we use* $\mathcal{D}'$ *to output* $\hat{\pi}$ *such that*

$$\mathbb{P}\left[\forall r, \left\|V^{\star}_{1,M} - V^{\hat{\pi}}_{1,M}\right\|_\infty \leq \epsilon\right] \geq 1 - \delta?$$

Definition 4.7.1 and Definition 4.7.2 are the offline counterparts of [84] and [80] in online RL. Those settings are of practical interests in the offline regime as well since in practice reward functions are often iteratively engineered to encourage desired behavior via trial and error and using one shot of offline exploration data $\mathcal{D}'$ to tackle problems with different reward functions (different tasks) could help improve sample efficiency significantly.

Our singleton absorbing MDP technique adapts to those settings and we have the following two theorems, Theorems 4.7.1, 4.7.2, whose proofs are found in Appendix C.

Theorem 4.7.1 (optimal offline task-agnostic learning)**.** *Given* $\mathcal{D}' = \{(s_t^{(i)}, a_t^{(i)})\}_{i\in[n]}^{t\in[H]}$ *by* μ. *Given K tasks with reward* $\{r_k\}_{k=1}^K$ *and the corresponding K MDPs* $M_k = (S, \mathcal{A}, P, r_k, H, d_1)$. *Denote* $\iota = \log(HSA/\delta)$. *Let* $\widehat{\pi}_k^{\star} := \operatorname{argmax}_\pi \widehat{V}^{\pi}_{1,M_k}$ $\forall k \in [K]$, *when* $n > O(H \cdot [\iota + \log(K)]/d_m)$, *then with probability* $1 - \delta$,

$$\left\|V^{\star}_{1,M_k} - V^{\widehat{\pi}_k^{\star}}_{1,M_k}\right\|_\infty \leq O\left[\sqrt{\frac{H^2(\iota + \log(K))}{nd_m}} + \frac{H^{2.5}S^{0.5}(\iota + \log(K))}{nd_m}\right]. \ \forall k \in [K]$$

Theorem 4.7.2 (optimal offline reward-free learning)**.** *Given* $\mathcal{D}' = \{(s_t^{(i)}, a_t^{(i)})\}_{i\in[n]}^{t\in[H]}$ *by* μ. *For*

any reward r denote the corresponding MDP $M = (S, \mathcal{A}, P, r, H, d_1)$. *Denote* $\iota = \log(HSA/\delta)$. *Let* $\widehat{\pi}^\star_M := \operatorname{argmax}_\pi \widehat{V}^\pi_{1,M}$ $\forall r$, *when* $n > O(HS \cdot \iota/d_m)$, *then with probability* $1-\delta$,

$$\left\| V^\star_{1,M} - V^{\widehat{\pi}^\star_M}_{1,M} \right\|_\infty \le O\left[\sqrt{\frac{H^2 S \cdot \iota}{n d_m}} + \frac{H^2 S \cdot \iota}{n d_m} \right]. \quad \forall r, M.$$

By a direct translation of both theorems, we have sample complexity of order $\widetilde{O}(H^2 \log(K)/d_m\epsilon^2)$ and $\widetilde{O}(H^2 S/d_m\epsilon^2)$. All the parameters have the optimal rates, see the lower bounds in [84] and [80].[9] The higher order dependence in Theorem 4.7.2 is also tight comparing to Theorem 4.7.1. Such statistically optimal results reveal the model-based methods generalize well to those seemingly challenging problems in the offline regime. Changing to these harder problems would not affect the optimal statistical efficiency of the model-based approach.

4.7.1 A Visualization of the Singleton Absorbing MDP

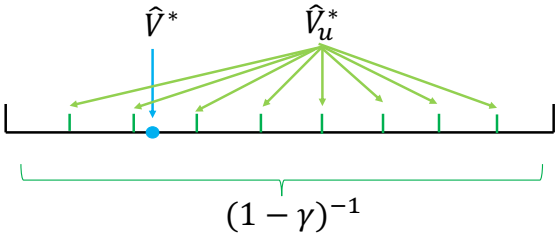

Figure 4.2: Covering-based Absorbing: 1-D case

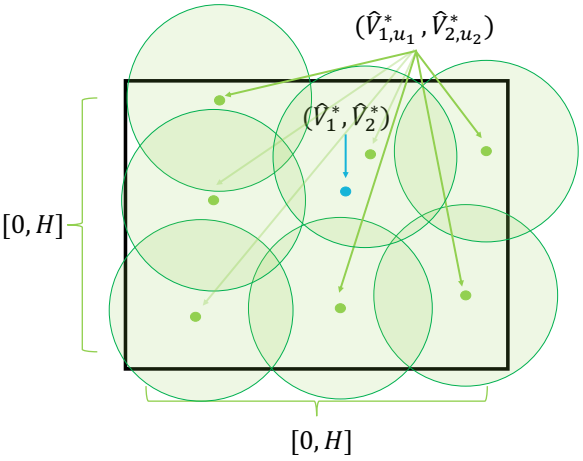

Figure 4.3: Covering-based Absorbing: 2-D case

[9] To be rigorous, we add a discussion in Appendix C to explain more clearly why our rates are optimal for these problems. We do not formalize these lower bounds in the offline cases as theorems since they are not novel contributions.

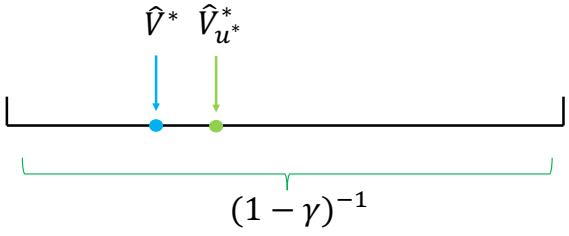

Figure 4.4: Singleton Absorbing: 1-D case

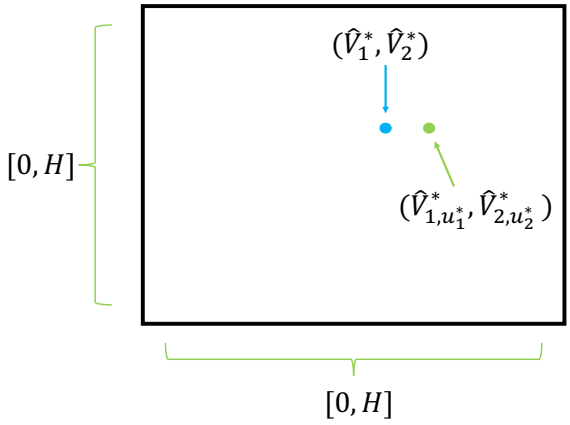

Figure 4.5: Singleton Absorbing: 2-D case

We provide a visualization for understanding the *singleton absorbing MDP* technique. 4.2, 4.4 demonstrate the infinite horizon case and 4.3, 4.5 demonstrate the finite horizon case. In particular, it should be a H-dimensional hypercube $[0, H]^H$ (that contains $\widehat{V}^\star_1, \ldots, \widehat{V}^\star_h$) instead of only the square $[0, H] \times [0, H]$ $(\widehat{V}^\star_1, \widehat{V}^\star_2)$. This is only for the ease of visualization.

The standard absorbing MDP technique [61, 68] leverages a set of absorbing MDPs to cover the range of value functions (following the standard covering principle) to make sure $\widehat{V}^\star$ is close to one of the element (absorbing MDP) in the set (Figure 4.2,4.3). The size of the covering set (*i.e.* the covering number) grows exponentially in H 4.3 in the finite horizon setting and this is due to the fact that there are $\widehat{V}^\star_1, \widehat{V}^\star_2, \ldots, \widehat{V}^\star_H$ quantities to cover. This results in the metric entropy (the log of the covering number) to blow up by a factor of H and incurs sub-optimality. On the other hand, by the nifty chosen singleton absorbing MDP $\widehat{V}^\star_{h,u^\star}$ (Figure 4.4,4.5), we completely get rid of the covering issue. To cover the H-dimensional space requires exponential H in size, maintain the independence, and control the error propagation (i.e. $\left\|\widehat{V}^\star - \widehat{V}^\star_{u^\star}\right\|_\infty$ is sufficiently small).

4.8 Extension to Linear MDP with Anchor Representations

The principle of our *Singleton absorbing MDP* technique (with model-based construction) in decoupling the dependence between $\widehat{P}_{s,a}$ and $\widehat{V}^\star$ is not confined to tabular MDPs and therefore it is natural to generalize such idea for the episodic stationary transition setting for other problems. As an example, we further present a sharp result for the setting of finite horizon linear MDP with anchor points. We narrate by assuming a generative oracle (that allows sampling from $s' \sim P(\cdot|s,a)$) for the ease of exposition.

Definition 4.8.1 (Linear MDP with anchor points [90, 68])**.** *Let $\mathcal{S}$ be the exponential large space and $\mathcal{A}$ be the infinite (or even continuous) spaces. Assume there is feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ (where $K \ll |\mathcal{S}|$), i.e. $\phi(s,a) = [\phi_1(s,a), \ldots, \phi_K(s,a)]$. Transition P admits a linear representation:*

$$P(s'|s,a) = \sum_{k\in[K]} \phi_k(s,a)\psi_k(s')$$

where $\psi_1(\cdot), \ldots, \psi_K(\cdot)$ are unknowns. We further assume there exists a set of anchor state-action pairs $\mathcal{K}$ such that any (s,a) can be represented as a convex combination of the anchors $\{(s_k,a_k)|k \in \mathcal{K}\}$:

$$\exists \left\{\lambda_k^{s,a}\right\} : \phi(s,a) = \sum_{k\in\mathcal{K}} \lambda_k^{s,a}\phi\left(s_k,a_k\right), \sum_{k\in\mathcal{K}} \lambda_k^{s,a} = 1, \lambda_k \geq 0, \forall k \in \mathcal{K}, (s,a) \in (\mathcal{S},\mathcal{A}).$$

Under the definition, denote N be the number of samples at each anchor pairs. Then we have the following (see Appendix C for the proof):

Theorem 4.8.1 (Optimal sample complexity)**.** *Under Definition 4.8.1, let $\widehat{\pi}^\star = \operatorname{argmax}_\pi \widehat{V}_1^\pi$. Then if $N \geq cH^2|S|\log(KH/\delta)$, we have with probability $1-\delta$, $||Q_1^\star - Q_1^{\widehat{\pi}^\star}||_\infty \leq \widetilde{O}(\sqrt{H^3/N})$.*

Comparing to Theorem 4 of [68], Theorem 4.8.1 removes the additional dependence $\min\{|S|, K, H\}$. In term of the total sample complexity, Theorem 4.8.1 gives $\tilde{O}(KH^3/\epsilon^2)$ while [68] has $\widetilde{O}(KH^4/\epsilon^2)$

(see their Section 7, first bullet point). Our result again reveals the model-based method is statistically optimal for the current setting.

Remark 8. *The rate $\tilde{O}(KH^3/\epsilon^2)$ with anchor point assumption has the linear dependence on K and for the standard linear bandit [91] $\Omega(\sqrt{d^2T})$ or the linear (mixture) MDP [60, 92] $\Omega(\sqrt{d^2H^2T})$ the lower bound dependence on the feature dimension d is quadratic. We believe one reason for this to happen is that the anchor representations assumption is somewhat strong as it abstracts the whole state action space by only finite points (via convex combination).*

4.9 Conclusion

This work [93] studies the uniform convergence problems for offline policy evaluation (OPE) and provides complete answers for their optimality behaviors. We achieve the optimal sample complexity for stationary-transition case using a novel adaptation of the absorbing MDP trick, which is more generally applicable to the new offline task-agnostic and reward-free settings combined with the model-based approach and we hope it can be applied to a broader range of future problems. We end the section by two future directions.

On the higher order error term. Our main result (Theorem 4.6.1) has an additional $\sqrt{HS}$ dependence in the higher order error term and we cannot further remove it based on our current technique. Nevertheless, this is already among the best higher order results to our knowledge. In fact, most state-of-the-art works (*e.g.* [39, 94, 95]) have additional S dependence in the higher order and [96] has only extra $\sqrt{S}$ in the higher order term but it also has additional $\sqrt{A}$ (see Table 1 of [95] for a clear reference). How to obtain optimality not only for the main term but also for the higher order error terms remains elusive for the community.

Uniform OPE and beyond. The current study of uniform OPE derives results with expression using parameter dependence and deriving instance-dependent uniform convergence result will draw a clearer picture on the individual behaviors for each policy. Besides, this work

concentrates on Tabular MDPs and generalizing uniform convergence to more practical settings like linear MDPs, game environments and multi-agent settings are promising future directions. Specifically, general complexity measure (mirroring VC-dimensions and Rademacher complexities for statistical learning problems) that precisely captures local and global uniform convergence would be of great interest.

Chapter 5

Conclusions and Future Directions

5.1 Conclusions and Summary

In this thesis, we analyzed the *Marginalized Importance Sampling* (MIS) estimator for Offline Policy Evaluation (Chapter 1), proposed the uniform convergence problem in such OPE (Chapter 2), and obtained the near-optimal sample complexity (Chapter 2, Chapter 3) in the time-homogeneous and time-inhomogeneous settings respectively. In (Chapter 4, we also studied a variety of topics that are not covered in the previous chapters.

- **Offline Policy Learning.** For the policy learning task, we propose the *Double Variance Reduction* algorithm (DVR)[7] for the tabular reinforcement learning, which attains the near-optimal minimax sample complexity guarantees for finite-horizon time-homogenuous, time-inhomogenuous, and infinite horizon discounted settings respectively. Furthermore, we improve the previous worst-case guarantees to the instance-adaptive guarantee [97], which subsumes nearly all the previous optimality results. Later, we consider the linear function approximation [98, 99] and the differentiable parametric function approximation [100] for offline RL.

- **Stochastic Shortest Path.** We initiated the stochastic shortest path setting in the offline regime under the tabular setting (there are finite number of states and actions) [101]. We consider both the offline policy learning and the offline policy evaluation tasks for this goal-orientated setting.

- **Low-switching RL.** In many real-world reinforcement learning (RL) tasks, it is costly to run fully adaptive algorithms that update the exploration policy frequently. Instead, collecting data in large batches using the current policy deployment is usually cheaper. Those problems can be cast as the low-switching RL problem, and [102] first achieves the $\log \log T$ switching cost with $\sqrt{T}$ regret.

- **Deep Reinforcement Learning.** We design the Closed-Form Policy Improvement (CFPI) [103] operator for tackling the locomotion tasks. We initiate offline RL algorithms with our novel policy improvement operators and empirically demonstrate their effectiveness over state-of-the-art algorithms on the standard D4RL benchmark [104].

5.2 Future Directions

Some Open problems:

As noted earlier, the conjecture posed in [21] remains unsolved. This has to do with the infinite $\mathcal{A}$ case, where we can never observe any (s, a) pair more than once, hence not able to estimate the transition dynamics or the expected reward. The minimax lower bound in [16] (for the contextual bandit setting) already establishes that the Cramer-Rao lower bound is not achievable in this setting even if $H = 1$ and $S = 1$. It remains open question as to whether H^3 is required. Another problem concerns better dependence for stationary transition case. We conjecture the dependence of H^2 can be further reduced in the stationary transition case. Our current analysis cannot further reduce the dependence for stationary transition setting.

Uniform OPE that depends on π. In this work, we primarily considered obtaining uniform bound independent to π; however, given a logging policy μ, it is often easier to evaluate certain policies than others, as revealed in the pointwise OPE bound of [57]. Specifically, obtaining a high probability bound of the form $\sup_\pi \frac{\sqrt{n}|\hat{v}^\pi - v^\pi|}{\gamma(\pi,\mu,M,\delta)} \leq C$ for some function γ and constant C would be of great interest. We could already get such a bound by applying union bound to the data-dependent high probability pointwise convergence of either [57] or [64] but it comes with an additional $O(S)$ factor. Characterizing the optimal per-instance OPE bound is an interesting future direction.

This thesis focuses on the tabular Reinforcement Learning (RL) with discrete states and actions. Recently, there has been a surge of studies in RL with general function approximation [105, 106, 107] that goes beyond the tabular setting and under a wide range of problem classes. However, while these algorithms have nice statistical guarantees, most of the algorithms are computationally inefficient. In the future, it would be exciting to study RL with general function approximation classes that can bridge the gap between the theory and practice (say via neural network approximations).

Appendix A

Supplementary Material to Chapter 2

A.1 Related settings to OPE

Markov Decision Processes have a long history of associated research [34, 35], but many theoretical problems in the basic tabular setting remain an active area of research as of today. We briefly review the other settings and connect them to our results.

Regret bound and sample complexity in the online setting. The bulk of existing work focuses on online learning, where the agent interacts with the MDP with the interests of identifying the optimal policy or minimizing the regret against the optimal policy. The optimal regret is obtained by [39] using a model-based approach which translates into a sample complexity bound of $O(H^3SA/\epsilon^2)$, which matches the lower bound of $\Omega(H^3SA/\epsilon^2)$[43]. The method is however not "uniform PAC" where the state of the art sample complexity remains $O(H^4SA/\epsilon^2)$ [108]. Model-free approaches that require a space constraint of $O(HSA)$ were studied by [96] which implies a sample complexity bound of $O(H^4SA/\epsilon^2)$.

Sample complexity with a generative model. Another sequence of work assumes access to a generative model where one can sample from s_{t+1} and r_t given any s_t, a_t in time $O(1)$ [109]. [42] is the first that establishes the optimal sample complexity of $\tilde{\Theta}(H^3SA/\epsilon^2)$ under this setting (counting H generative model calls as one episode). [110] establishes a similar results by estimating the parameters of the MDP model using maximum-likelihood estimation.

A.2 Proof of the main Theorem 2.5.1

To analyze the MSE upper bound $\mathbb{E}_\mu[(\widehat{v}^\pi_{\text{TMIS}} - v^\pi)^2]$, we create a fictitious surrogate $\widetilde{v}^\pi_{\text{TMIS}}$, which is an unbiased version of $\widehat{v}^\pi_{\text{TMIS}}$. A few auxiliary lemmas are first presented and Bellman equations are used for deriving variance decomposition in a recursive way. Second order moment of marginalized state distribution $\widetilde{d}^\pi_t$ can then be bounded by analyzing its variance.

A.2.1 Fictitious tabular MIS estimator.

The fictitious estimator[1] $\widetilde{v}^{\pi}$ fills in the gap of state-action location (s_t, a_t) of the true estimator $\widehat{v}^{\pi}$ where $n_{s_t,a_t} = 0$. Specifically, it replaces every component in $\widehat{v}^{\pi}$ with a fictitious counterpart, *i.e.* $\widetilde{v}^{\pi} := \sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi}, \widetilde{r}_t^{\pi} \rangle$, with $\widetilde{d}_t^{\pi} = \widetilde{P}_t^{\pi} \widetilde{d}_{t-1}^{\pi}$ and $\widetilde{P}_t^{\pi}(s_t|s_{t-1}) = \sum_{a_{t-1}} \widetilde{P}_t(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|s_{t-1})$, $\widetilde{r}_t^{\pi}(s_t) = \sum_{a_t} \widetilde{r}_t(s_t, a_t)\pi(a_t|s_t)$. In particular, let E_t denotes the event $\{n_{s_t,a_t} \geq n d_t^{\mu}(s_t, a_t)(1-\theta)\}$[2], then

$$\widetilde{r}_t(s_t, a_t) = \widehat{r}_t(s_t, a_t)\mathbf{1}(E_t) + r_t(s_t, a_t)\mathbf{1}(E_t^c)$$
$$\widetilde{P}_{t+1}(\cdot|s_t, a_t) = \widehat{P}_{t+1}(\cdot|s_t, a_t)\mathbf{1}(E_t) + P_{t+1}(\cdot|s_t, a_t)\mathbf{1}(E_t^c).$$

where $0 < \theta < 1$ is a parameter that we will choose later.

The name "fictitious" comes from the fact that $\widetilde{v}^{\pi}$ is *not implementable* using the data[3], but it creates a bridge between $\widehat{v}^{\pi}$ and v^{π} because of its unbiasedness, see Lemma A.2.5. Also, for simplicity of the proof, throughout the rest of the paper we denote: $\mathcal{D}_t := \left\{ s_{1:t}^{(i)}, a_{1:t}^{(i)}, r_{1:t-1}^{(i)} \right\}_{i=1}^{n}$. Also, in the base case, we denote $\mathcal{D}_1 := \left\{ s_1^{(i)}, a_1^{(i)} \right\}_{i=1}^{n}$ and that $r_t^{\pi}(s_t) := \mathbb{E}_{\pi}[r_t^{(1)}|s_t^{(1)} = s_t] = \sum_{a_t} \mathbb{E}[r_t^{(1)}|s_t^{(1)} = s_t, a_t^{(1)} = a_t]\pi(a_t|s_t) := \sum_{a_t} r_t(s_t, a_t)\pi(a_t|s_t)$. Then we have the following preliminary auxiliary lemmas.

Lemma A.2.1. *$\widetilde{d}_t^{\pi}$ and $\widetilde{r}_{t-1}^{\pi}$ are deterministic given $\mathcal{D}_t$. Moreover, given $\mathcal{D}_t$, $\widetilde{P}_{t+1,t}^{\pi}$ is unbiased of $P_{t+1,t}^{\pi}$ and $\widetilde{r}_t^{\pi}$ is unbiased of r_t^{π}.*

Proof: [Proof of Lemma A.2.1] By construction of the estimator, $\widetilde{d}_t^{\pi}$ and $\widetilde{r}_{t-1}^{\pi}$ only depend on $\mathcal{D}_t$, therefore $\widetilde{d}_t^{\pi}$ and $\widetilde{r}_{t-1}^{\pi}$ given $\mathcal{D}_t$ are constants. For the second argument, we have $\forall s_t, s_{t+1}$,

$$\begin{aligned}
&\mathbb{E}[\widetilde{P}_{t+1,t}^{\pi}(s_{t+1}|s_t)|\mathcal{D}_t] = \sum_{a_t} \mathbb{E}[\widetilde{P}_{t+1,t}(s_{t+1}|s_t, a_t)|\mathcal{D}_t]\pi(a_t|s_t) \\
&= \sum_{a_t} \left(\mathbf{1}(E_t)\mathbb{E}[\widehat{P}_{t+1,t}(s_{t+1}|s_t, a_t)|\mathcal{D}_t] + \mathbf{1}(E_t^c)P_{t+1,t}(s_{t+1}|s_t, a_t) \right)\pi(a_t|s_t) \\
&= \sum_{a_t} \left(\mathbf{1}(E_t)P_{t+1,t}(s_{t+1}|s_t, a_t) + \mathbf{1}(E_t^c)P_{t+1,t}(s_{t+1}|s_t, a_t) \right)\pi(a_t|s_t) \\
&= \sum_{a_t} P_{t+1,t}(s_{t+1}|s_t, a_t)\pi(a_t|s_t) = P_{t+1,t}^{\pi}(s_{t+1}|s_t),
\end{aligned}$$

[1] We replcace the notation of $\widetilde{v}_{\text{TMIS}}^{\pi}$ with just $\widetilde{v}^{\pi}$ throughout the proof. $\widetilde{v}^{\pi}$ always denotes fictitious tabular MIS estimator.

[2] More rigorously, E_t depends on the specific pair s_t, a_t and should be written as $E_t(s_t, a_t)$. However, for brevity we just use E_t and this notation should be clear in each context.

[3] It depends on unknown information such as d_t^{μ}, $P_{t,t-1}^{\pi}$, exact conditional expectation of the reward r_t^{π} and so on.

where the third equal sign comes from the fact that conditional on E_t, $\hat{P}(s_{t+1}|s_t, a_t)$ — the empirical mean — is unbiased. The result about $\tilde{r}_t^\pi$ can be derived using a similar fashion.

Using Lemma A.2.1, we can derive the following recursions for expectation and variance:

Lemma A.2.2. *For $h = 1, ..., H$, we have*

$$\mathbb{E}\left[\langle \widetilde{d}_h^\pi, V_h^\pi\rangle + \sum_{t=1}^{h-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle \middle| \mathcal{D}_{h-1}\right] = \langle \widetilde{d}_{h-1}^\pi, V_{h-1}^\pi\rangle + \sum_{t=1}^{h-2}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle, \tag{A.1}$$

$$\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \sum_{t=1}^{h}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] = \mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle \widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle \middle| \mathcal{D}_h\right]\right] + \mathrm{Var}\left[\langle \widetilde{d}_h^\pi, V_h^\pi\rangle + \sum_{t=1}^{h-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] \tag{A.2}$$

Proof: The proof of Lemma A.2.2 can be found in Lemma B.2 and Lemma 4.1 in xie2019towards by coupling the standard Bellman equation:

$$V_h^\pi = r_h^\pi + [P_{h+1,h}^\pi]^T V_{h+1}^\pi \tag{A.3}$$

with the total law of expectations and the total law of variances.

Lemma A.2.3 (Boundedness of Tabular MIS estimators)**.** $0 \le \widehat{v}^\pi \le H R_{\max}$, $0 \le \widetilde{v}^\pi \le H R_{\max}$.

Proof: we show $\widehat{P}_t^\pi(\cdot|s_{t-1})$ is a (degenerated) probability distribution for all t, s_{t-1}.

$$\begin{aligned}
\sum_{s_t}\widehat{P}_t^\pi(s_t|s_{t-1}) &= \sum_{s_t}\sum_{a_{t-1}}\widehat{P}_t(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|s_{t-1}) \\
&= \sum_{a_{t-1}}\sum_{s_t}\widehat{P}_t(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|s_{t-1}) \qquad \text{This is since } |\mathcal{A}|, |\mathcal{S}| < \infty \\
&= \sum_{a_{t-1}}\sum_{s_t}\frac{n_{s_t,s_{t-1},a_{t-1}}}{n_{s_{t-1},a_{t-1}}}\pi(a_{t-1}|s_{t-1}) \\
&\le \sum_{a_{t-1}}\pi(a_{t-1}|s_{t-1}) = 1
\end{aligned} \tag{A.4}$$

The last line is inequality since $\widehat{P}_t(s_t|s_{t-1}, a_{t-1}) = 0$ when $n_{s_{t-1},a_{t-1}} = 0$. Following the same logic, it is easy to show $\widetilde{P}_t^\pi(\cdot|s_{t-1})$ is a non-degenerated probability distribution.

Next note $\sum_{s_1}\widehat{d}_1^\pi(s_1) = \sum_{s_1}\widehat{d}_1^\mu(s_1) = \sum_{s_1}\frac{n_{s_1}}{n} = 1$. Suppose $\widehat{d}_{t-1}^\pi(\cdot)$ is a (degenerated) probability distribution, then from $\widehat{d}_t^\pi = \widehat{P}_t^\pi\widehat{d}_{t-1}^\pi$ and (A.4), by induction we know $\widehat{d}_t^\pi(\cdot)$ is a (degenerated) probability distribution for all t.

Using Assumption 3.3.1, it is easy to show $\hat{r}_t^\pi(s_t) \le R_{\max}$ for all s_t, then combining all results above we have $\widehat{v}^\pi := \sum_{t=1}^H\langle \widehat{d}_t^\pi, \widehat{r}_t^\pi\rangle \le H R_{\max}$. Similarly, $\widetilde{v}^\pi \le H R_{\max}$.

The boundedness of Tabular-MIS estimator cannot be inherited by the State-MIS estimator since $\widehat{v}_{\mathrm{SMIS}}^\pi$ explicitly uses importance weights and there is no reason for it to be less than $H R_{\max}$.

As a result, we do not need an extra projection step for our estimation to be valid. Thanks to the following lemma, throughout the rest of the analysis we only need to consider $\widetilde{v}^{\pi}$.

Lemma A.2.4. *Let $\widehat{v}^{\pi}$ be the Tabular-MIS estimator and $\widetilde{v}^{\pi}$ be the fictitious version of TMIS we described above with parameter θ. Then the MSE of the TMIS and fictitious TMIS satisfies*

$$\mathbb{E}[(\widehat{v}^{\pi} - v^{\pi})^2] \leq \mathbb{E}[(\widetilde{v}^{\pi} - v^{\pi})^2] + 3H^3SAR_{\max}^2 e^{-\frac{\theta^2 n \min_{t,s_t,a_t} d_t^{\mu}(s_t,a_t)}{2}}$$

Proof: [Proof of Lemma A.2.4] Define $E := \{\exists t, s_t, a_t \text{ s.t. } n_{s_t,a_t} < n d_t^{\mu}(s_t,a_t)(1-\theta)\}$. Similarly to Lemma B.1 in the appendix of [21], we have

$$\begin{aligned}
&\mathbb{E}[(\widehat{v}^{\pi} - v^{\pi})^2] \leq \mathbb{E}[(\widehat{v}^{\pi} - v^{\pi})^2] = \mathbb{E}[(\widehat{v}^{\pi} - \widetilde{v}^{\pi})^2] + 2\mathbb{E}[(\widehat{v}^{\pi} - \widetilde{v}^{\pi})(\widetilde{v}^{\pi} - v^{\pi})] + \mathbb{E}[(\widetilde{v}^{\pi} - v^{\pi})^2] \\
=&\mathbb{P}[E]\mathbb{E}\left[(\widehat{v}^{\pi} - \widetilde{v}^{\pi})^2 + 2(\widehat{v}^{\pi} - \widetilde{v}^{\pi})(\widetilde{v}^{\pi} - v^{\pi})\middle| E\right] + \mathbb{P}[E^c]\cdot 0 + \mathbb{E}[(\widetilde{v}^{\pi} - v^{\pi})^2] \\
\leq&3\mathbb{P}[E]H^2R_{\max}^2 + \mathbb{E}[(\widetilde{v}^{\pi} - v^{\pi})^2],
\end{aligned}$$

where the last inequality uses Lemma A.2.3. Then combining the multiplicative Chernoff bound and a union bound over each t,s_t and a_t, we get that

$$\mathbb{P}[E] \leq \sum_t \sum_{s_t} \sum_{a_t} \mathbb{P}[n_{s_t,a_t} < n d_t^{\mu}(s_t,a_t)(1-\theta)] \leq HSAe^{-\frac{\theta^2 n \min_{t,s_t,a_t} d_t^{\mu}(s_t,a_t)}{2}},$$

which provides the stated result.

Lemma A.2.4 tells that MSE of two TMISs differs by a quantity $3H^3SAR_{\max}^2 e^{-\frac{\theta^2 n \min_{t,s_t,a_t} d_t^{\mu}(s_t,a_t)}{2}}$ and this illustrates that the gap between two MSE's can be sufficiently small as long as $n \geq \frac{\text{polylog}(S,A,H,n)}{\min_{t,s_t,a_t} d_t^{\mu}(s_t,a_t)}$.

A.2.2 Variance and Bias of Fictitious tabular MIS estimator.

Lemma A.2.5 ([21] Lemma B.2)**.** *Tabular-MIS estimator is unbiased: $\mathbb{E}[\widetilde{v}^{\pi}] = v^{\pi}$ for all $\theta < 1$.*

Lemma A.2.6 (Variance decomposition)**.**

$$\begin{aligned}
\text{Var}[\widetilde{v}^{\pi}] =& \frac{\text{Var}[V_1^{\pi}(s_1^{(1)})]}{n} \\
&+ \sum_{h=1}^{H} \sum_{s_h} \sum_{a_h} \mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h,a_h}} \mathbf{1}(E_h)\right] \pi(a_h|s_h)^2 \text{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right].
\end{aligned} \tag{A.5}$$

where $V_t^{\pi}(s_t)$ denotes the value function under π which satisfies the Bellman equation

$$V_t^{\pi}(s_t) = r_t^{\pi}(s_t) + \sum_{s_{t+1}} P_t^{\pi}(s_{t+1}|s_t) V_{t+1}^{\pi}(s_{t+1}).$$

Remark 9. *Note even though the construction of TMIS and SMIS are different, both fictitious estimators are unbiased for v^π. Therefore the MSE of MIS estimators are dominated by the variance of the fictitious estimators. Comparing Lemma A.2.6 with Lemma 4.1 in [21] we can see our Tabular-MIS estimator achieves a lower bound, and it is essentially asymptotic optimal, as explained by Remark 1.*

Proof: [Proof of Lemma A.2.6] The proof relies on applying Lemma A.2.2 in a recursive way. To begin with, we use the following variance decomposition, which applies (A.2) recursively.

$$
\begin{aligned}
\mathrm{Var}[\widetilde{v}^\pi] =& \mathbb{E}\mathrm{Var}[\widetilde{v}^\pi|\mathcal{D}_H] + \mathrm{Var}[\mathbb{E}[\widetilde{v}^\pi|\mathcal{D}_H]] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \mathrm{Var}[\mathbb{E}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H] + \sum_{t=1}^{H-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \mathrm{Var}[\langle \widetilde{d}_H^\pi, r_H^\pi\rangle + \sum_{t=1}^{H-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \mathrm{Var}[\langle \widetilde{d}_H^\pi, V_H^\pi\rangle + \sum_{t=1}^{H-1}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle] \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_H^\pi, V_H^\pi\rangle + \langle \widetilde{d}_{H-1}^\pi, \widetilde{r}_{H-1}^\pi\rangle\middle|\mathcal{D}_{H-1}\right]\right] \\
&+ \mathrm{Var}\left[\langle \widetilde{d}_{H-1}^\pi, V_{H-1}^\pi\rangle + \sum_{t=1}^{H-2}\langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi\rangle\right] = \ldots \\
=& \mathbb{E}\left[\mathrm{Var}[\langle \widetilde{d}_H^\pi, \widetilde{r}_H^\pi\rangle|\mathcal{D}_H]\right] + \sum_{h=1}^{H-1}\mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle \widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle\middle|\mathcal{D}_h\right]\right] + \mathrm{Var}\left[\langle \widetilde{d}_1^\pi, V_1^\pi\rangle\right]
\end{aligned}
$$

Now let us analyze $\mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^\pi, V_{h+1}^\pi\rangle + \langle \widetilde{d}_h^\pi, \widetilde{r}_h^\pi\rangle\middle|\mathcal{D}_h\right]\right]$. Note $\widetilde{P}_{h+1,h}^\pi(\cdot, s_h)$ and $\widetilde{r}_h^\pi(s_h)$ for each s_h are conditionally independent given $\mathcal{D}_h$, since $\mathcal{D}_h$ partitions the n episodes into S disjoint sets according to the states $s_h^{(i)}$ at time h. Similarly, $\tilde{P}_{h+1}(\cdot|s_h, a_h)$ and $\tilde{r}_h^\pi(s_h, a_h)$ for each (s_h, a_h) are also conditionally independent given $\mathcal{D}_h$. These observations imply:

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{Var}\left[\langle \widetilde{d}_{h+1}^{\pi}, V_{h+1}^{\pi}\rangle + \langle \widetilde{d}_h^{\pi}, \widetilde{r}_h^{\pi}\rangle \middle| \mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h} \mathrm{Var}\left[\widetilde{d}_h^{\pi}(s_h)\langle \widetilde{P}_{h+1,h}^{\pi}(\cdot, s_h), V_{h+1}^{\pi}\rangle + \widetilde{d}_h^{\pi}(s_h)\cdot \widetilde{r}_h^{\pi}(s_h)\middle| \mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h} \widetilde{d}_h^{\pi 2}(s_h)\mathrm{Var}\left[\sum_{a_h}\langle \widetilde{P}_{h+1}(\cdot|s_h,a_h)\cdot\pi(a_h|s_h), V_{h+1}^{\pi}\rangle + \sum_{a_h}\widetilde{r}_h(s_h,a_h)\cdot\pi(a_h|s_h)\middle| \mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h} \widetilde{d}_h^{\pi}(s_h)^2\sum_{a_h}\pi(a_h|s_h)^2\mathrm{Var}\left[\langle \widetilde{P}_{h+1}(\cdot|s_h,a_h), V_{h+1}^{\pi}\rangle + \widetilde{r}_h(s_h,a_h)\middle| \mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h} \widetilde{d}_h^{\pi}(s_h)^2\sum_{a_h}\pi(a_h|s_h)^2\mathbf{1}(E_t)\mathrm{Var}\left[\frac{1}{n_{s_h,a_h}}\sum_{i|s_h^{(i)}=s_h,a_h^{(i)}=a_h}(V_{h+1}^{\pi}(s_{h+1}^{(i)}) + r_h^{(i)})\middle| \mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\sum_{s_h} \widetilde{d}_h^{\pi}(s_h)^2\sum_{a_h}\pi(a_h|s_h)^2\cdot\frac{\mathbf{1}(E_t)}{n_{s_h,a_h}}\cdot\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(i)}) + r_h^{(i)})\middle| s_h^{(i)}=s_h, a_h^{(i)}=a_h\right]\right] \\
=&\sum_{s_h}\sum_{a_h}\pi(a_h|s_h)^2\cdot\mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h,a_h}}\cdot\mathbf{1}(E_t)\right]\cdot\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(i)}) + r_h^{(i)})\middle| s_h^{(i)}=s_h, a_h^{(i)}=a_h\right].
\end{aligned}
\tag{A.6}
$$

The second line and the fourth line use the conditional independence for s_t and (s_t, a_t) respectively. The fifth line uses that when $n_{s_h,a_h} < nd_h^{\mu}(s_h,a_h)(1-\theta)$, the conditional variance is 0. The sixth line uses the fact that episodes are iid.

Plug (A.6) into the above variance decomposition and uses $V_{H+1} = 0$, we finally get

$$
\begin{aligned}
\mathrm{Var}[\widetilde{v}^{\pi}] =& \frac{\mathrm{Var}[V_1^{\pi}(s_1^{(1)})]}{n} \\
&+ \sum_{h=1}^{H}\sum_{s_h}\sum_{a_h}\mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h,a_h}}\mathbf{1}(E_h)\right]\pi(a_h|s_h)^2\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)}=s_h, a_h^{(1)}=a_h\right].
\end{aligned}
$$

A.2.3 Bounding the variance of $\widetilde{d}_h^{\pi}(s_h)$.

Applying the definition of variance, we directly have

$$
\mathbb{E}\left[\frac{\widetilde{d}_h^{\pi}(s_h)^2}{n_{s_h,a_h}}\mathbf{1}(E_h)\right] \le \frac{(1-\theta)^{-1}}{nd_h^{\mu}(s_h,a_h)}\mathbb{E}\left[\widetilde{d}_h^{\pi}(s_h)^2\right] = \frac{(1-\theta)^{-1}}{nd_h^{\mu}(s_h,a_h)}(d_h^{\pi}(s_h)^2 + \mathrm{Var}[\widetilde{d}_h^{\pi}(s_h)]), \tag{A.7}
$$

where we use the fact that $\widetilde{d}_h^{\pi}(s_h)$ is unbiased (which can be proved by induction through applying total law of expectations and the recursive relationship $\widetilde{d}_t^{\pi} = \widetilde{P}_t^{\pi}\widetilde{d}_{t-1}^{\pi}$). Therefore the only

thing left is to bound the the variance of $\widetilde{d}_h^\pi(s_h)$. To tackle it, we consider bounding the covariance matrix of $\widetilde{d}_h^\pi(s_h)$. As we shall see in Lemma A.2.7, fortunately, we are able to derive an identical result of Lemma B.4 in xie2019towards for our Tabular-MIS estimator, which helps greatly in bounding the the variance of $\widetilde{d}_h^\pi(s_h)$.

Lemma A.2.7 (Covariance of $\widetilde{d}_h^\pi$ with TMIS)**.**

$$\begin{aligned}\mathrm{Cov}(\widetilde{d}_h^\pi) \preceq & \frac{(1-\theta)^{-1}}{n}\sum_{t=1}^{h-1}\mathbb{P}^\pi_{h+1,t+1}\mathrm{diag}\left[\sum_{s_t,a_t}\frac{d_t^\pi(s_t)^2+\mathrm{Var}(\widetilde{d}_t^\pi(s_t))}{d_t^\mu(s_t)}\frac{\pi(a_t|s_t)^2}{\mu(a_h|s_t)}\mathbb{P}_{t+1,t}(\cdot|s_t,a_t)\right]\left[\mathbb{P}^\pi_{h+1,t+1}\right]^T \\ & + \frac{1}{n}\mathbb{P}^\pi_{h,1}\mathrm{diag}\left[d_1^\pi\right][\mathbb{P}^\pi_{h,1}]^T.\end{aligned}$$

where $\mathbb{P}^\pi_{h,t} = \mathbb{P}^\pi_{h,h-1}\cdot\mathbb{P}^\pi_{h-1,h-2}\cdot\ldots\cdot\mathbb{P}^\pi_{t+1,t}$ *— the transition matrices under policy* π *from time* t *to* h *(define* $\mathbb{P}^\pi_{h,h} := I$*).*

Proof: [Proof of Lemma A.2.7] We start by applying the law of total variance to obtain the following recursive equation

$$\mathrm{Cov}[\widetilde{d}_h^\pi] = \mathbb{E}\left[\mathrm{Cov}\left[\widetilde{\mathbb{P}}^\pi_{h,h-1}\widetilde{d}^\pi_{h-1}\middle|\mathcal{D}_{h-1}\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\widetilde{\mathbb{P}}^\pi_{h,h-1}\widetilde{d}^\pi_{h-1}\middle|\mathcal{D}_{h-1}\right]\right] \tag{A.8}$$

$$= \mathbb{E}\left[\mathrm{Cov}\left[\sum_{s_{h-1}}\widetilde{\mathbb{P}}^\pi_{h,h-1}(\cdot|s_{h-1})\widetilde{d}^\pi_{h-1}(s_{h-1})\middle|\mathcal{D}_{h-1}\right]\right] + \mathrm{Cov}\left[\mathbb{E}\left[\widetilde{\mathbb{P}}^\pi_{h,h-1}\widetilde{d}^\pi_{h-1}\middle|\mathcal{D}_{h-1}\right]\right] \tag{A.9}$$

$$= \underbrace{\mathbb{E}\left[\sum_{s_{h-1}}\mathrm{Cov}\left[\widetilde{\mathbb{P}}^\pi_{h,h-1}(\cdot|s_{h-1})\middle|\mathcal{D}_{h-1}\right]\widetilde{d}^\pi_{h-1}(s_{h-1})^2\right]}_{(*)} + \mathbb{P}^\pi_{h,h-1}\mathrm{Cov}[\widetilde{d}^\pi_{h-1}][\mathbb{P}^\pi_{h,h-1}]^T. \tag{A.10}$$

The decomposition of the covariance in the third line uses that $\mathrm{Cov}(X+Y) = \mathrm{Cov}(X)+\mathrm{Cov}(Y)$ when X and Y are statistically independent and the columns of $\widetilde{\mathbb{P}}_{h,h-1}$ are independent when conditioning on $\mathcal{D}_{h-1}$.

$$(*) = \mathbb{E}\left[\sum_{s_{h-1}}\sum_{a_{h-1}} \pi(a_{h-1}|s_{h-1})^2 \mathrm{Cov}\left[\widetilde{\mathbb{P}}_h(\cdot|s_{h-1},a_{h-1})\middle|\mathrm{Data}_{h-1}\right] \widetilde{d}^{\pi}_{h-1}(s_{h-1})^2\right] \tag{A.11}$$

$$= \mathbb{E}\left[\sum_{s_{h-1}}\sum_{a_{h-1}} \pi(a_{h-1}|s_{h-1})^2 \mathbf{1}(E_{h-1}) \mathrm{Cov}\left[\widehat{\mathbb{P}}_h(\cdot|s_{h-1},a_{h-1})\middle|\mathrm{Data}_{h-1}\right] \widetilde{d}^{\pi}_{h-1}(s_{h-1})^2\right] \tag{A.12}$$

$$= \mathbb{E}\left[\sum_{s_{h-1}}\sum_{a_{h-1}} \pi(a_{h-1}|s_{h-1})^2 \frac{\mathbf{1}(E_{h-1})}{n_{s_{h-1},a_{h-1}}} \mathrm{Cov}\left[\mathbf{e}_{s_h^{(1)}}\middle|s_{h-1}^{(1)}=s_{h-1}, a_{h-1}^{(1)}=a_{h-1}\right] \widetilde{d}^{\pi}_{h-1}(s_{h-1})^2\right] \tag{A.13}$$

$$= \sum_{s_{h-1},a_{h-1}} \pi(a_{h-1}|s_{h-1})^2 \mathbb{E}\left[\frac{\widetilde{d}^{\pi}_{h-1}(s_{h-1})^2}{n_{s_{h-1},a_{h-1}}}\mathbf{1}(E_{h-1})\right]\Big[\mathrm{diag}[\mathbb{P}_h(\cdot|s_{h-1},a_{h-1})] \tag{A.14}$$

$$- \mathbb{P}_h(\cdot|s_{h-1},a_{h-1})\cdot\mathbb{P}_h(\cdot|s_{h-1},a_{h-1})^T\Big] \tag{A.15}$$

$$\prec \sum_{s_{h-1}}\sum_{a_{h-1}} \left\{\frac{d^{\pi}_{h-1}(s_{h-1})^2 + \mathrm{Var}[\widetilde{d}^{\pi}_{h-1}(s_{h-1})]}{n d^{\mu}_{h-1}(s_{h-1})(1-\theta)} \frac{\pi(a_{h-1}|s_{h-1})^2}{\mu(a_{h-1}|s_{h-1})} \mathrm{diag}[\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})]\right\} \tag{A.16}$$

The second line uses the fact that conditional on E^c_{h-1}, the variance of $\widetilde{\mathbb{P}}(\cdot|s_{h-1},a_{h-1})$ is zero given Data_h. The third line uses the basic property of empirical average, and the fourth line comes from the fact

$$\begin{aligned}
&\mathrm{Cov}\left[\mathbf{e}_{s_h^{(1)}}\middle|s_{h-1}^{(1)}=s_{h-1}, a_{h-1}^{(1)}=a_{h-1}\right]\\
=&\mathbb{E}\left[\mathbf{e}_{s_h^{(1)}}\cdot\mathbf{e}^T_{s_h^{(1)}}\middle|s_{h-1}^{(1)}=s_{h-1}, a_{h-1}^{(1)}=a_{h-1}\right]\\
&-\mathbb{E}\left[\mathbf{e}_{s_h^{(1)}}\middle|s_{h-1}^{(1)}=s_{h-1}, a_{h-1}^{(1)}=a_{h-1}\right]\cdot\mathbb{E}\left[\mathbf{e}_{s_h^{(1)}}\middle|s_{h-1}^{(1)}=s_{h-1}, a_{h-1}^{(1)}=a_{h-1}\right]^T\\
=&\mathrm{diag}(\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})) - \mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})[\mathbb{P}_{h,h-1}(\cdot|s_{h-1},a_{h-1})]^T
\end{aligned}$$

The last line (A.16) uses the fact that $\mathbb{P}^{\pi}_{h,h-1}(\cdot|s_{h-1})[\mathbb{P}^{\pi}_{h,h-1}(\cdot|s_{h-1})]^T$ is positive semidefinite, $n_{s_{h-1},a_{h-1}} \geq n d^{\mu}_{h-1}(s_{h-1},a_{h-1})(1-\theta)$ and the definition of variance for $\widetilde{d}^{\pi}_{h-1}(s_{h-1})$. Combining (A.10) and (A.16) and by recursively apply them, we get the stated results.

Benefitting from the identical semidefinite ordering bound on $\mathrm{Cov}(\widetilde{d}^{\pi}_h)$ for TMIS and SMIS, we can borrow the following results from [21] for our Tabular-MIS estimator.

Lemma A.2.8 (Corollary 2 of [21]). *For $h = 1$, we have* $\mathrm{Var}[\widetilde{d}^{\pi}_1(s_1)] = \frac{1}{n}(d^{\pi}_h(s_1) - d^{\pi}_h(s_1)^2)$,

and for $h = 2, 3, ..., H$*, we have:*

$$\mathrm{Var}[\widetilde{d}_h^{\pi}(s_h)] \leq \frac{(1-\theta)^{-1}}{n} \sum_{t=2}^{h} \sum_{s_t} \mathbb{P}_{h,t}^{\pi}(s_h|s_t)^2 \varrho(s_t) + \frac{1}{n} \sum_{s_1} \mathbb{P}_{h,1}^{\pi}(s_h|s_1)^2 d_1(s_1)$$

where $\varrho(s_t) := \sum_{s_{t-1}} \left(\frac{d_{t-1}^{\pi}(s_{t-1})^2 + \mathrm{Var}(\widetilde{d}_{t-1}^{\pi}(s_{t-1}))}{d_{t-1}^{\mu}(s_{t-1})} \sum_{a_{t-1}} \frac{\pi(a_{t-1}|s_{t-1})^2}{\mu(a_{t-1}|s_{t-1})} \mathbb{P}_{t,t-1}(s_t|s_{t-1}, a_{t-1}) \right)$.

Lemma A.2.9 (Error propagation: Theorem B.1 of [21])**.** *Let* $\tau_a := \max_{t,s_t,a_t} \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$ *and* $\tau_s := \max_{t,s_t} \frac{d_t^{\pi}(s_t)}{d_t^{\mu}(s_t)}$*. If* $n \geq \frac{2(1-\theta)^{-1} t \tau_a \tau_s}{\max\{d_t^{\pi}(s_t), d_t^{\mu}(s_t)\}}$ *for all* $t = 2, ..., H$*, then for all* $h = 1, 2, ..., H$ *and* s_h*, we have that:*

$$\mathrm{Var}[\widetilde{d}_h^{\pi}(s_h)] \leq \frac{2(1-\theta)^{-1} h \tau_a \tau_s}{n} d_h^{\pi}(s_h).$$

Before giving the proof of Theorem 2.5.1, we first prove Lemma B.4.4.

Proof: [Proof of Lemma B.4.4] Let value function $V_h^{\pi}(s_h) = \mathbb{E}_{\pi}[\sum_{t=h}^{H} r_t^{(1)} | s_h^{(1)} = s_h]$ and Q-function $Q_h^{\pi}(s_h, a_h) = \mathbb{E}_{\pi}[\sum_{t=h}^{H} r_t^{(1)} | s_h^{(1)} = s_h, a_h^{(1)} = a_h]$, then by total law of variance we

obtain (let's suppress the policy π for simplicity):

$$
\begin{aligned}
&\operatorname{Var}\left[\sum_{t=1}^{h} r_t^{(1)} + V_{h+1}(s_{h+1}^{(1)})\right] \\
=&\mathbb{E}\left[\operatorname{Var}\left[\sum_{t=1}^{h} r_t^{(1)} + V_{h+1}(s_{h+1}^{(1)})\middle|\mathcal{D}_h\right]\right] + \operatorname{Var}\left[\mathbb{E}\left[\sum_{t=1}^{h} r_t^{(1)} + V_{h+1}(s_{h+1}^{(1)})\middle|\mathcal{D}_h\right]\right] \\
=&\mathbb{E}\left[\operatorname{Var}\left[r_h^{(1)} + V_{h+1}(s_{h+1}^{(1)})\middle|s_h^{(1)}, a_h^{(1)}\right]\right] + \operatorname{Var}\left[\sum_{t=1}^{h-1} r_t^{(1)} + \mathbb{E}\left[V_{h+1}(s_{h+1}^{(1)}) + r_h^{(1)}\middle|s_h^{(1)}, a_h^{(1)}\right]\right] \\
=&\mathbb{E}\left[\operatorname{Var}\left[r_h^{(1)} + V_{h+1}(s_{h+1}^{(1)})\middle|s_h^{(1)}, a_h^{(1)}\right]\right] + \operatorname{Var}\left[\sum_{t=1}^{h-1} r_t^{(1)} + Q_h(s_h^{(1)}, a_h^{(1)})\right] \\
=&\mathbb{E}\left[\operatorname{Var}\left[r_h^{(1)} + V_{h+1}(s_{h+1}^{(1)})\middle|s_h^{(1)}, a_h^{(1)}\right]\right] + \mathbb{E}\left[\operatorname{Var}\left[\sum_{t=1}^{h-1} r_t^{(1)} + Q_h(s_h^{(1)}, a_h^{(1)})\middle|s_h^{(1)}, r_{1:h-1}^{(1)}\right]\right] \\
+&\operatorname{Var}\left[\mathbb{E}\left[\sum_{t=1}^{h-1} r_t^{(1)} + Q_h(s_h^{(1)}, a_h^{(1)})\middle|s_h^{(1)}, r_{1:h-1}^{(1)}\right]\right] \\
=&\mathbb{E}\left[\operatorname{Var}\left[r_h^{(1)} + V_{h+1}(s_{h+1}^{(1)})\middle|s_h^{(1)}, a_h^{(1)}\right]\right] + \mathbb{E}\left[\operatorname{Var}\left[Q_h(s_h^{(1)}, a_h^{(1)})\middle|s_h^{(1)}, r_{1:h-1}^{(1)}\right]\right] \\
+&\operatorname{Var}\left[\sum_{t=1}^{h-1} r_t^{(1)} + \mathbb{E}\left[Q_h(s_h^{(1)}, a_h^{(1)})\middle|s_h^{(1)}\right]\right] \\
=&\mathbb{E}\left[\operatorname{Var}\left[r_h^{(1)} + V_{h+1}(s_{h+1}^{(1)})\middle|s_h^{(1)}, a_h^{(1)}\right]\right] + \mathbb{E}\left[\operatorname{Var}\left[Q_h(s_h^{(1)}, a_h^{(1)})\middle|s_h^{(1)}\right]\right] + \operatorname{Var}\left[\sum_{t=1}^{h-1} r_t^{(1)} + V_h(s_h^{(1)})\right],
\end{aligned}
\tag{A.17}
$$

where we use Markovian property that $(V_{h+1}(s_{h+1}^{(1)})|\mathcal{D}_h)$ equals $(V_{h+1}(s_{h+1}^{(1)})|s_h^{(1)}, a_h^{(1)})$ in distribution and $\mathbb{E}\left[V_{h+1}(s_{h+1}^{(1)}) + r_h^{(1)}\middle|s_h^{(1)}, a_h^{(1)}\right] = Q_h(s_h^{(1)}, a_h^{(1)})$. Then by applying (A.17) recursively and letting $h = H$, we get the stated result.

Remark 10. *A straight forward implication of Lemma B.4.4 is the following:*

$$
\sum_{t=1}^{H} \mathbb{E}_\pi\left[\operatorname{Var}\left[V_{t+1}^{\pi}(s_{t+1}^{(1)}) + r_t^{(1)}\middle|s_t^{(1)}, a_t^{(1)}\right]\right] \leq H^2 R_{\max}^2.
$$

Combing Lemma A.2.6 and A.2.9, we are now ready to prove the main Theorem 2.5.1.

Proof: [Proof of Theorem 2.5.1] Plug the result of Lemma A.2.9 into Lemma A.2.6 and

uses the unbiasedness of $\widetilde{v}^{\pi}_{\text{TMIS}}$ (Lemma A.2.5) we obtain $\forall\ 0 < \theta < 1$:

$$\begin{aligned}
&\mathbb{E}[(\widetilde{v}^{\pi}_{\text{TMIS}} - v^{\pi})^2] \\
\leq &\frac{\text{Var}[V_1^{\pi}(s_1^{(1)})]}{n} + \sum_{h=1}^{H}\sum_{s_h,a_h}\frac{(1-\theta)^{-1}}{n d_h^{\mu}(s_h,a_h)} d_h^{\pi}(s_h)^2\pi(a_h|s_h)^2\text{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\Big| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right]. \\
&+ \frac{(1-\theta)^{-1}}{n}\sum_{h=1}^{H}\sum_{s_h,a_h}\frac{2(1-\theta)^{-1}h\tau_a\tau_s}{n}\frac{d_h^{\pi}(s_h)}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\text{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\Big| s_h^{(1)} = s_h\right]
\end{aligned} \tag{A.18}$$

Choose $\theta = \sqrt{4\log(n)/(n\min_{t,s_t,a_t} d_t^{\mu}(s_t,a_t))}$. Then by assumption $n > \frac{16\log n}{\min_{t,s_t,a_t} d_t^{\mu}(s_t,a_t)}$ we have $\theta < 1/2$, which allows us to write $(1-\theta)^{-1} \leq (1+2\theta)$ in the leading term and $(1-\theta)^{-1} \leq 2$ in the subsequent terms. The condition of Lemma A.2.9 is satisfied by The second assumption on n. Then, combining (A.18) with Lemma A.2.4 we get:

$$\begin{aligned}
&\mathbb{E}[(\widehat{v}^{\pi}_{\text{TMIS}} - v^{\pi})^2] \leq \frac{1}{n}\sum_{h=0}^{H}\sum_{s_h,a_h}\frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\text{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\Big| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right] \\
&\cdot\left(1 + \sqrt{\frac{16\log n}{n\min_{t,s_t} d_t^{\mu}(s_t)}}\right) + \frac{3}{n^2}H^3SAR_{\max}^2 \\
&+\frac{8\tau_a\tau_s}{n^2}\sum_{h=1}^{H}\sum_{s_h,a_h}\frac{h\cdot d_h^{\pi}(s_h)}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\cdot\text{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\Big| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right],
\end{aligned} \tag{A.19}$$

now use Lemma B.4.4, we can bound the last term in (A.19) by

$$\frac{8\tau_a^2\tau_s H}{n^2\cdot d_m}\sum_{t=1}^{H}\mathbb{E}_{\pi}\left[\text{Var}\left[V_{t+1}^{\pi}(s_{t+1}^{(1)}) + r_t^{(1)}\Big| s_t^{(1)}, a_t^{(1)}\right]\right] \leq \frac{8\tau_a^2\tau_s H^3R_{\max}^2}{n^2\cdot d_m},$$

Combine this term with $\frac{3}{n^2}H^3SAR_{\max}^2$ we obtain the higher order term $O(\frac{\tau_a^2\tau_s H^3R_{\max}^2}{n^2\cdot d_m})$, where we use that pigeonhole principle implies that $S < \tau_s, A < \tau_a$.

This completes the proof.

A.3 Proofs of data splitting Tabular-MIS estimator.

We define the fictitious data splitting Tabular-MIS estimator as:

$$\widetilde{v}^{\pi}_{\text{split}} = \frac{1}{N}\sum_{i=1}^{N}\widetilde{v}^{\pi}_{(i)},$$

where each $\widetilde{v}^{\pi}_{(i)}$ is the fictitious Tabular-MIS estimator of $\widehat{v}^{\pi}_{(i)}$. Moreover, we set all $\widetilde{v}^{\pi}_{(1)}, \widetilde{v}^{\pi}_{(2)}, ..., \widetilde{v}^{\pi}_{(N)}$ jointly share the same fictitious parameter θ_M.

Proof: [Proof of Theorem 2.7.1] Let $E' := \{\exists\ \widetilde{v}^{\pi}_{(i)} : s.t. \widetilde{v}^{\pi}_{(i)} \neq \widehat{v}^{\pi}_{(i)}\}$, then an argument similar to Lemma A.2.4 can be derived:

$$\mathbb{E}[(\widehat{v}^{\pi}_{\text{split}} - v^{\pi})^2] \leq 3\mathbb{P}[E']H^2R^2_{\max} + \mathbb{E}[(\widetilde{v}^{\pi}_{\text{split}} - v^{\pi})^2],$$

and

$$\mathbb{P}[E'] \leq N \sum_t \sum_{s_t} \sum_{a_t} \mathbb{P}[n_{s_t,a_t} < M \cdot d^{\mu}_t(s_t,a_t)(1-\theta_M)] \leq NHSAe^{-\frac{\theta^2_M M \min_{t,s_t,a_t} d^{\mu}_t(s_t,a_t)}{2}},$$

therefore $\mathbb{P}[E']$ will be sufficiently small if $M \geq O(\text{Polylog}(H,S,A,n)/\min_{t,s_t,a_t} d^{\mu}_t(s_t,a_t))$. By near-uniformity we $M \geq O(\text{Polylog}(H,S,A,n)SA) \geq O(\text{Polylog}(H,S,A,n)/\min_{t,s_t,a_t} d^{\mu}_t(s_t,a_t))$.

Moreover, by i.i.d and unbiasedness of $\widetilde{v}^{\pi}_{(i)}$, we have

$$\mathbb{E}[(\widetilde{v}^{\pi}_{\text{split}} - v^{\pi})^2] = \frac{1}{N}\mathbb{E}[(\widetilde{v}^{\pi}_{(1)} - v^{\pi})^2] \leq \frac{1}{N} \cdot O(\frac{H^2SA}{M}) = O(\frac{H^2SA}{n}),$$

by the second assumption on M and Theorem 2.5.1.

We now proof Lemma 2.7.1, since it will be used to as the intermediate step for proving Theorem 2.7.2.

Proof: [Proof of Lemma 2.7.1] Note that

$$\begin{aligned}
&\mathbb{P}\left[\left\{\exists \pi \in \prod\ s.t.\ \widetilde{v}^{\pi}_{\text{split}} \neq \widehat{v}^{\pi}_{\text{split}}\right\}\right] \leq N \cdot \mathbb{P}\left[\left\{\exists \pi \in \prod,\ s.t.\ \widetilde{v}^{\pi}_{(1)} \neq \widehat{v}^{\pi}_{(1)}\right\}\right] \\
&\leq N \cdot \mathbb{P}\left[\{\exists t, s_t, a_t \text{ s.t. } n^{(1)}_{s_t,a_t} < nd^{\mu}_t(s_t,a_t)(1-\theta_M)\}\right] \\
&\leq NHSAe^{-\frac{\theta^2_M M \min_{t,s_t,a_t} d^{\mu}_t(s_t,a_t)}{2}},
\end{aligned}$$

therefore by near-uniformity $M > \max\left[O(SA \cdot \text{Polylog}(S,H,A,N,1/\delta)), O(H\tau_a\tau_s)\right]$ is sufficient to guarantee the stated result.

Now we can prove Theorem 2.7.2.

Proof: [Proof of Theorem 2.7.2] First of all, we have

$$\mathbb{P}\left(|\widehat{v}^{\pi}_{\text{split}} - v^{\pi}| > \epsilon\right) \leq \mathbb{P}\left(|\widehat{v}^{\pi}_{\text{split}} - \widetilde{v}^{\pi}_{\text{split}}| > 0\right) + \mathbb{P}\left(|\widetilde{v}^{\pi}_{\text{split}} - v^{\pi}| > \epsilon\right), \tag{A.20}$$

Now by Bernstein inequality we have

$$\mathbb{P}\left(|\widetilde{v}^{\pi}_{\text{split}} - v^{\pi}| > \epsilon\right) = \mathbb{P}\left(|\frac{1}{N}\sum_{i=1}^{N}(\widetilde{v}^{\pi}_{(i)} - v^{\pi})| \geq \epsilon\right) \leq \exp\left(-\frac{N\epsilon^2}{2\text{Var}(\widetilde{v}^{\pi}_{(1)}) + 2HR_{\max}\epsilon/3}\right) := \delta/2. \tag{A.21}$$

Solving (A.21) and apply Theorem 2.5.1, we obtain

$$\epsilon \leq \sqrt{\frac{2\mathrm{Var}(\widetilde{v}^{\pi}_{(1)})\log(2/\delta)}{N}} + \frac{2HR_{\max}\log(2/\delta)}{3N} \leq \widetilde{O}(\sqrt{\frac{H^2SA\log(2/\delta)}{M\cdot N}}) + \frac{2HR_{\max}\log(2/\delta)}{3N}. \tag{A.22}$$

As N goes large, the square root term in (A.22) will dominate and it seems we only need to consider the square root term in N and treat the second term as the higher order term. However, since $M > \max\left[O(SA\cdot \mathrm{Polylog}(S,H,A,N,1/\delta)), O(H\tau_a\tau_s)\right]$, N cannot be arbitrary large given n. An example is: when $N = n$, then $M = n/N = 1$ does not satisfy the condition. Therefore to make the square root term dominates we need

$$\sqrt{\frac{H^2SA\log(2/\delta)}{M\cdot N}} \geq O(\frac{HR_{\max}\log(2/\delta)}{N}).$$

This translates to

$$M \leq \widetilde{O}(\sqrt{nSA}), \tag{A.23}$$

where $\widetilde{O}$ absorbs all the Polylog terms.

Therefore under the condition (A.23), we can really absorb the second term in (A.22) (as higher order term) and combine it with Lemma 2.7.1 to get that with probability $1-\delta$,

$$|\widehat{v}^{\pi}_{\mathrm{split}} - v^{\pi}| \leq 0 + \widetilde{O}(\sqrt{\frac{H^2SA}{M\cdot N}}) = \widetilde{O}(\sqrt{\frac{H^2SA}{n}}).$$

Proof: [Proof of Theorem 2.9.1] The non-uniform result of Theorem 2.7.2 gives:

$$|\widehat{v}^{\pi}_{\mathrm{split}} - v^{\pi}| \leq \widetilde{O}(\sqrt{\frac{H^2SA}{n}})$$

Note that all nonstationary deterministic polices class have cardinality $|\prod| = A^{HS}$, which implies $\log|\prod| = HS\log A$, therefore combine Lemma 2.7.1 with a direct union bound and Multiplicative Chernoff bound we obtain

$$\sup_{\pi\in\prod} |\widehat{v}^{\pi}_{\mathrm{split}} - v^{\pi}| \leq \widetilde{O}(\sqrt{\frac{H^3S^2A}{n}})$$

Algorithm 2 Data Splitting Tabular MIS OPE

Input: Logging data $\mathcal{D} = \{\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{t=1}^{H}\}_{i=1}^{n}$ from the behavior policy μ. A target policy π which we want to evaluate its cumulative reward. Splitting data size M.

1: Randomly splitting the data $\mathcal{D}$ evenly into N folds, with each fold $|\mathcal{D}^{(i)}| = M$.
2: **for** $i = 1, 2, \ldots, N$ **do**
3: Use Algorithm 3 to estimate $\widehat{v}_{(i)}^{\pi}$ with data $\mathcal{D}^{(i)}$.
4: **end for**
5: Use the mean of $\widehat{v}_{(1)}^{\pi}, \widehat{v}_{(2)}^{\pi}, ..., \widehat{v}_{(N)}^{\pi}$ as the final estimation of v^{π}.

A.4 More details about Empirical Results.

Restate Time-varying, non-mixing Tabular MDP in Section 2.8.

There are two states s_0 and s_1 and two actions a_1 and a_2. State s_0 always has probability 1 going back to itself, regardless of the actions, *i.e.* $P_t(s_0|s_0, a_1) = 1$ and $P_t(s_0|s_0, a_2) = 1$. For state s_1, at each time step there is one action (we call it a) that has probability $2/H$ going to s_0 and the other action (we call it a') has probability 1 going back to s_1,

$$P_t(s|s_1, a) = \begin{cases} \frac{2}{H} & \text{if } s = s_0; \\ 1 - \frac{2}{H} & \text{if } s = s_1. \end{cases} \quad P_t(s|s_1, a') = \begin{cases} 0 & \text{if } s = s_0; \\ 1 & \text{if } s = s_1. \end{cases}$$

and which action will make state s_1 go to state s_0 with probability $2/H$ is decided by a random parameter p_t uniform sampled in $[0, 1]$. If $p_t < 0.5$, $a = a_1$ and if $p_t \geq 0.5$, $a = a_2$. These $p_1, ..., p_H$ are generated by a sequence of pseudo-random numbers. Moreover, one can receive reward 1 at each time step if $t > H/2$ and is in state s_0, and will receive reward 0 otherwise. Lastly, for logging policy, we define it to be uniform:

$$\mu(\cdot|s_0) = \begin{cases} \frac{1}{2} & \text{if } \cdot = a_1; \\ \frac{1}{2} & \text{if } \cdot = a_2. \end{cases} \quad \text{and} \quad \mu(\cdot|s_1) = \begin{cases} \frac{1}{2} & \text{if } \cdot = a_1; \\ \frac{1}{2} & \text{if } \cdot = a_2. \end{cases}$$

For target policy π, we define it as:

$$\pi(\cdot|s_0) = \begin{cases} \frac{1}{2} & \text{if } \cdot = a_1; \\ \frac{1}{2} & \text{if } \cdot = a_2. \end{cases} \quad \text{and} \quad \pi(\cdot|s_1) = \begin{cases} \frac{1}{4} & \text{if } \cdot = a_1; \\ \frac{3}{4} & \text{if } \cdot = a_2. \end{cases}$$

We run this non-stationary MDP model in the `Python` environment and pseudo-random numbers p_t's are generated by keeping `numpy.random.seed(100)`.

We run each methods under $K = 100$ macro-replications with data $\mathcal{D}_{(k)} = \left\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)})\right\}_{(k)}^{i\in[n],t\in[H]}$, and use each $\mathcal{D}_{(k)}$ $(k = 1, ..., K)$ to construct a estimator $\widehat{v}_{[k]}^{\pi}$, then the (empirical) RMSE is com-

puted as:

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{K}(\widehat{v}_{[k]}^{\pi} - v_{\text{true}}^{\pi})^2}{K}},$$

where v_{true}^{π} is obtained by calculating $P_{t+1,t}^{\pi}(s'|s) = \sum_a P_{t+1,t}(s'|s,a)\pi_t(a|s)$, the marginal state distribution $d_t^{\pi} = P_{t,t-1}^{\pi} d_{t-1}^{\pi}$, $r_t^{\pi}(s_t) = \sum_{a_t} r_t(s_t,a_t)\pi_t(a_t|s_t)$ and $v_{\text{true}}^{\pi} = \sum_{t=1}^{H}\sum_{s_t} d_t^{\pi}(s_t) r_t^{\pi}(s_t)$. Then Relative-RMSE equals to $\text{RMSE}/v_{\text{true}}^{\pi}$.

Other generic IS-based estimators. There are other Importance Sampling based estimators including *weighted importance sampling* (WIS) and *importance sampling with stationary state distribution* (SSD-IS, [29]). The empirical comparisons including these methods are well-demonstrated in [21] and it was empirically shown that they are worse than SMIS. Because of that, we only focus on comparing SMIS and TMIS in our simulation study.

Appendix B

Supplementary Material to Chapter 3

B.1 On error metric for OPE

In this section, we discuss the metric considered in this work. Traditionally, most works directly use *Mean Square Error* (MSE) $\mathbb{E}[(\widehat{v}^{\pi}-v^{\pi})^2]$ as the criterion for measuring OPE methods *e.g.* [19, 111, 23, 56], or equivalently, by proposing unbiased estimators and discussing its variance *e.g.* [18]. Alternately, one can consider bounding the absolute difference between v^{π} and $\widehat{v}^{\pi}$ with high probability (*e.g.* [64]), *i.e.* $|\widehat{v}^{\pi} - v^{\pi}| \leq \epsilon_{\text{prob}}$ *w.h.p.* Generally speaking, high probability bound can be seen as a stricter criterion compared to MSE since

$$\begin{aligned}\mathbb{E}[(\widehat{v}^{\pi} - v^{\pi})^2] &= \mathbb{E}[(\widehat{v}^{\pi} - v^{\pi})^2 \mathbf{1}_E] + \mathbb{E}[(\widehat{v}^{\pi} - v^{\pi})^2 \mathbf{1}_{E^c}] \\ &\leq \epsilon_{\text{prob}}(\delta)^2 \cdot (1-\delta) + H^2 \cdot \delta,\end{aligned} \tag{B.1}$$

where E is the event that ϵ_{prob} error holds and δ is the failure probability. As a result, if both δ and $\epsilon_{\text{prob}}(\delta)$ can be controlled small, then the high probability bound implies a result for MSE bound. This is realistic, since δ mostly appears inside the logarithmic term of $\epsilon_{\text{prob}}(\delta)$ so the second term can be scaled to sufficiently small without affecting the polynomial dependence for the first term.

B.2 Some preparations

In this section we present some results that are critical for proving the main theorems.

Lemma B.2.1. *For any $0 < \delta < 1$, there exists an absolute constant c_1 such that when total episode $n > c_1 \cdot 1/d_m \cdot \log(HSA/\delta)$, then with probability $1-\delta$,*

$$n_{s_t,a_t} \geq n \cdot d_t^{\mu}(s_t, a_t)/2, \quad \forall\ s_t, a_t.$$

If state s_t is not accessible, then $n_{s_t,a_t} = d_t^\mu(s_t,a_t) = 0$ so the lemma holds trivially.[1]

Proof: [Proof of Lemma B.2.1] Define $E := \{\exists t, s_t, a_t \text{ s.t. } n_{s_t,a_t} < nd_t^\mu(s_t,a_t)/2\}$. Then combining the multiplicative Chernoff bound and a union bound over each t,s_t and a_t, we obtain

$$\begin{aligned}\mathbb{P}[E] &\leq \sum_t \sum_{s_t} \sum_{a_t} \mathbb{P}[n_{s_t,a_t} < nd_t^\mu(s_t,a_t)/2] \\ &\leq HSA \cdot e^{-\frac{n\cdot\min_{t,s_t,a_t} d_t^\mu(s_t,a_t)}{8}} = HSA \cdot e^{-\frac{n\cdot d_m}{8}} := \delta\end{aligned}$$

solving this for n then provides the stated result.

Now we define: $N := \min_{t,s_t,a_t} n_{s_t,a_t}$, then above implies $N \geq nd_m/2$ (recall d_m in Assumption 4.4.1). Now we aggregate only the first N pieces of data in each state-action (s_t,a_t)[2] of off-policy data $\mathcal{D}$ and they consist of a new dataset $\mathcal{D}' = \{(s_t,a_t,s_{t+1}^{(i)},r_t^{(i)}) : i = 1,...,N; t \in [H]; s_t \in \mathcal{S}, a_t \in \mathcal{A}\}$, and is a subset of $\mathcal{D}$. For the rest of paper, we will use either $\mathcal{D}'$ or the original $\mathcal{D}$ to create OPEMA $\widehat{v}^\pi$ (only for theoretical analysis purpose). Whether $\mathcal{D}$ or $\mathcal{D}'$ is used will be stated clearly in each context.

Remark 11. *It is worth mentioning that when use $\mathcal{D}'$ to construct $\widehat{v}^\pi$, $n_{s_t,a_t}^{\mathcal{D}'} = N$ for all s_t,a_t. Also, $N := \min n_{s_t,a_t}^{\mathcal{D}}$ (note $n_{s_t,a_t}^{\mathcal{D}}$ is the count from $\mathcal{D}$) itself is a random variable and in the extreme case we could have $N = 0$ and if that happens $\widehat{v}^\pi = 0$ (since in that case $\widehat{P}_t \equiv 0$ and $\widehat{d}_t^\pi$ is degenerated). However, there is only tiny probability N will be small, as guaranteed by Lemma B.2.1.*

We wanted to point out that this technique of dropping certain amount of data, is not uncommon for analyzing model-based method in RL: e.g. Rmax exploration [112] for online episodic setting (see [[41], Notes on Rmax exploration] Section 2 Algorithm for tabular MDP. The data they use is the "known set" K with parameter m, in step3 data pairs observed more than m times are not recorded).

B.2.1 Fictitious OPEMA estimator.

Similar to [21, 57], we introduce an unbiased version of $\widehat{v}^\pi$ to fill in the gap at (s_t,a_t) where n_{s_t,a_t} is small. Concretely, every component in $\widehat{v}^\pi$ is substituted by the fictitious counterpart, *i.e.* $\widetilde{v}^\pi := \sum_{t=1}^H \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle$, with $\widetilde{d}_t^\pi = \widetilde{P}_t^\pi \widetilde{d}_{t-1}^\pi$ and $\widetilde{P}_t^\pi(s_t|s_{t-1}) = \sum_{a_{t-1}} \widetilde{P}_t(s_t|s_{t-1},a_{t-1})\pi(a_{t-1}|s_{t-1})$. In particular, consider the high probability event in Lemma B.2.1, *i.e.* let E_t denotes the event

[1] In general, non-accessible state will not affect our results so to make our presentation succinct we will not mention non-accessible state for the rest of paper unless necessary.

[2] Note we can do this since by definition $N \leq n_{s_t,a_t}$ for all s_t,a_t.

$\{n_{s_t,a_t} \geq nd_t^{\mu}(s_t,a_t)/2\}$[3], then we define

$$\widetilde{r}_t(s_t,a_t) = \widehat{r}_t(s_t,a_t)\mathbf{1}(E_t) + r_t(s_t,a_t)\mathbf{1}(E_t^c)$$
$$\widetilde{P}_{t+1}(\cdot|s_t,a_t) = \widehat{P}_{t+1}(\cdot|s_t,a_t)\mathbf{1}(E_t) + P_{t+1}(\cdot|s_t,a_t)\mathbf{1}(E_t^c).$$

Similarly, for the OPEMA estimator uses data $\mathcal{D}'$, the fictitious estimator is set to be

$$\widetilde{r}_t(s_t,a_t) = \widehat{r}_t(s_t,a_t)\mathbf{1}(E) + r_t(s_t,a_t)\mathbf{1}(E^c)$$
$$\widetilde{P}_{t+1}(\cdot|s_t,a_t) = \widehat{P}_{t+1}(\cdot|s_t,a_t)\mathbf{1}(E) + P_{t+1}(\cdot|s_t,a_t)\mathbf{1}(E^c)$$

where E denote the event $\{N \geq nd_m/2\}$.

$\widetilde{v}^{\pi}$ creates a bridge between $\widehat{v}^{\pi}$ and v^{π} because of its unbiasedness and it is also bounded by H (see Lemma B.3 and Lemma B.5 in [57] for those preliminary results). Also, $\widetilde{v}^{\pi}$ is identical to $\widehat{v}^{\pi}$ with high probability, as stated by the following lemma.

Lemma B.2.2. *For any $0 < \delta < 1$, there exists an absolute constant c_1 such that when total episode $n > c_1 d_m \cdot \log(HSA/\delta)$, then with probability $1-\delta$,*

$$\sup_{\pi\in\Pi} |\widehat{v}^{\pi} - \widetilde{v}^{\pi}| = 0.$$

Proof: This Lemma is a direct corollary of Lemma B.2.1 by considering the event $E_1 := \{\exists t, s_t, a_t \text{ s.t. } n_{s_t,a_t} < nd_t^{\mu}(s_t,a_t)/2\}$ or $\{N < nd_m/2\}$ since $\widehat{v}^{\pi}$ and $\widetilde{v}^{\pi}$ are identical on E_1^c.

Note $\widehat{v}^{\pi}$ and $\widetilde{v}^{\pi}$ even equal to each other uniformly over all π in Π. This is not surprising since only logging policy μ will decide if they are equal or not. This lemma shows how close $\widehat{v}^{\pi}$ and $\widetilde{v}^{\pi}$ are. Therefore in the following it suffices to consider the uniform convergence of $\sup_{\pi\in\Pi} |\widetilde{v}^{\pi} - v^{\pi}|$.

Next by using a fictitious analogy of state-action expression as in equation (3.1), we have:

$$\begin{aligned}
\sup_{\pi\in\Pi} |\widetilde{v}^{\pi} - v^{\pi}| &= \sup_{\pi\in\Pi} |\sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi}, \widetilde{r}_t\rangle - \sum_{t=1}^{H} \langle d_t^{\pi}, r_t\rangle| \\
&= \sup_{\pi\in\Pi} |\sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi}, \widetilde{r}_t\rangle - \sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi}, r_t\rangle + \sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi}, r_t\rangle - \sum_{t=1}^{H} \langle d_t^{\pi}, r_t\rangle| \\
&\leq \underbrace{\sup_{\pi\in\Pi} |\sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle|}_{(*)} + \underbrace{\sup_{\pi\in\Pi} |\sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi}, \widetilde{r}_t - r_t\rangle|}_{(**)}
\end{aligned} \tag{B.2}$$

We first deal with $(**)$ by the following lemma.

[3]More rigorously, E_t depends on the specific pair s_t, a_t and should be written as $E_t(s_t,a_t)$. However, for brevity we just use E_t and this notation should be clear in each context.

Lemma B.2.3. *We have with probability* $1-\delta$*:*

$$\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\leq O(\sqrt{\frac{H^2\log(HSA/\delta)}{n\cdot d_m}})$$

Proof: [Proof of Lemma B.2.3]
Since $|\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\leq||\widetilde{d}_t^{\pi}||_1\cdot||\widetilde{r}_t-r_t||_\infty$, we obtain

$$|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\leq\sum_{t=1}^{H}||\widetilde{d}_t^{\pi}||_1\cdot||\widetilde{r}_t-r_t||_\infty=\sum_{t=1}^{H}||\widetilde{r}_t-r_t||_\infty,$$

where we used $\widetilde{d}_t^{\pi}(\cdot)$ is a probability distribution. Therefore above expression further indicates $\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\leq\sum_{t=1}^{H}||\widetilde{r}_t-r_t||_\infty$. Now by a union bound and Hoeffding inequality,

$$\begin{aligned}
\mathbb{P}(\sup_t||\widetilde{r}_t-r_t||_\infty>\epsilon)&=\mathbb{P}(\sup_{t,s_t,a_t}|\widetilde{r}_t(s_t,a_t)-r_t(s_t,a_t)|>\epsilon)\\
&\leq HSA\cdot\mathbb{P}(|\widetilde{r}_t(s_t,a_t)-r_t(s_t,a_t)|>\epsilon)\\
&=HSA\cdot\mathbb{P}(|\widehat{r}_t(s_t,a_t)-r_t(s_t,a_t)|\mathbf{1}(E_t)>\epsilon)\\
&\leq 2HSA\cdot\mathbb{E}[\mathbb{E}[e^{-2n_{s_t,a_t}\epsilon^2}|E_t]]\\
&\leq 2HSA\cdot\mathbb{E}[\mathbb{E}[e^{-nd_m\epsilon^2}|E_t]]=2HSA\cdot e^{-nd_m\epsilon^2}:=\frac{\delta}{2}.
\end{aligned}$$

where we use $\mathbb{P}(A)=\mathbb{E}[\mathbf{1}_A]=\mathbb{E}[\mathbb{E}[\mathbf{1}_A|X]]$. Solving for ϵ, then it follows:

$$\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi},\widetilde{r}_t-r_t\rangle|\leq\sum_{t=1}^{H}||\widetilde{r}_t-r_t||_\infty\leq O(\sqrt{\frac{H^2\log(HSA/\delta)}{n\cdot d_m}})$$

with probability $1-\delta$. The case for $E=\{N\geq nd_m/2\}$ can be proved easily in a similar way.

Note that in order to measure the randomness in reward, sample complexity n only has dependence of order H^2, this result implies random reward will only cause error of lower order dependence in H. Therefore, in many RL literature deterministic reward is directly assumed. Next we consider $(*)$ in (B.2) by decomposing $\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle$ into a martingale type representation. This is the key for our proof since with it we can use either uniform concentration inequalities or martingale concentration inequalities to prove efficiency.

B.2.2 Decomposition of $\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle$

Let $\widetilde{d}_t^{\pi}\in\mathbb{R}^{S\cdot A}$ denote the marginal state-action probability vector, $\pi_t\in\mathbb{R}^{(S\cdot A)\times S}$ is the policy matrix with $(\pi_t)_{(s_t,a_t),s_t}=\pi_t(a_t|s_t)$ and $(\pi_t)_{(s_t,a_t),s}=0$ for $s\neq s_t$. Moreover, let state-

action transition matrix $T_t \in \mathbb{R}^{S\times(S\cdot A)}$ to be $(T_t)_{s_t,(s_{t-1},a_{t-1})} = P_t(s_t|s_{t-1},a_{t-1})$, then we have

$$\widetilde{d}_t^{\pi} = \pi_t \widetilde{T}_t \widetilde{d}_{t-1}^{\pi} \tag{B.3}$$
$$d_t^{\pi} = \pi_t T_t d_{t-1}^{\pi}. \tag{B.4}$$

Therefore we have

$$\widetilde{d}_t^{\pi} - d_t^{\pi} = \pi_t(\widetilde{T}_t - T_t)\widetilde{d}_{t-1}^{\pi} + \pi_t T_t(\widetilde{d}_{t-1}^{\pi} - d_{t-1}^{\pi}) \tag{B.5}$$

recursively apply this formula, we have

$$\widetilde{d}_t^{\pi} - d_t^{\pi} = \sum_{h=2}^{t} \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi} + \Gamma_{1:t}(\widetilde{d}_1^{\pi} - d_1^{\pi}) \tag{B.6}$$

where $\Gamma_{h:t} = \prod_{v=h}^{t} \pi_v T_v$ and $\Gamma_{t+1:t} := 1$. Now let $X = \sum_{t=1}^{H}\langle r_t, \widetilde{d}_t^{\pi} - d_t^{\pi}\rangle$, then we have the following:

Theorem B.2.1 (martingale decomposition of X: Restate of the fictitious version of Lemma 3.5.1)**.** *We have:*

$$X = \sum_{h=2}^{H}\langle V_h^{\pi}(s), ((\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi})(s)\rangle + \langle V_1^{\pi}(s), (\widetilde{d}_1^{\pi} - d_1^{\pi})(s)\rangle,$$

where the inner product is taken w.r.t states.

Proof: [Proof of Theorem B.2.1] By applying (B.6) and the change of summation, we have

$$\begin{aligned}
X &= \sum_{t=1}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi}\rangle + \langle r_t, \Gamma_{1:t}(\widetilde{d}_1^{\pi} - d_1^{\pi})\rangle\right) \\
&= \sum_{t=1}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi}\rangle\right) + \sum_{h=1}^{H}\langle r_t, \Gamma_{1:t}(\widetilde{d}_1^{\pi} - d_1^{\pi})\rangle \\
&= \sum_{t=2}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi}\rangle\right) + \sum_{h=1}^{H}\langle r_t, \Gamma_{1:t}(\widetilde{d}_1^{\pi} - d_1^{\pi})\rangle \\
&= \sum_{h=2}^{H}\left(\sum_{t=h}^{H}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi}\rangle\right) + \sum_{h=1}^{H}\langle (\pi_1^T\Gamma_{1:t}^T r_t)(s), (\widetilde{d}_1^{\pi} - d_1^{\pi})(s)\rangle \\
&= \sum_{h=2}^{H}\left(\langle \underbrace{\sum_{t=h}^{H}\pi_h^T\Gamma_{h+1:t}^T r_t}_{V_h^{\pi}(s)}, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi}\rangle\right) + \langle \underbrace{(\sum_{h=1}^{H}\pi_1^T\Gamma_{1:t}^T r_t)(s)}_{V_1^{\pi}(s)}, (\widetilde{d}_1^{\pi} - d_1^{\pi})(s)\rangle
\end{aligned}$$

B.3 Proof of uniform convergence in OPE with full policies using standard uniform concentration tools: Theorem 3.5.1

As a reminder for the reader, the OPEMA estimator used in this section is with data subset $\mathcal{D}'$. Also, by Lemma B.2.3 we only need to consider $\sup_{\pi\in\Pi} |\sum_{t=1}^{H}\langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle|$.

Theorem B.3.1. *There exists an absolute constant c such that if* $n > c\cdot\frac{1}{d_m}\cdot\log(HSA/\delta)$*, then with probability* $1-\delta$*, we have:*

$$\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle\right| \le O\left(\sqrt{\frac{H^4\log(HSA/\delta)}{nd_m}}\right) + \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle\right|\right]$$

Proof: [Proof of Theorem B.3.1] Note in data $\mathcal{D}' = \{(s_t, a_t, s_{t+1}^{(i)}) : i = 1, ..., N; t = 1, ..., H; s_t \in \mathcal{S}, a_t \in \mathcal{A}\}$[4], not only $s_{t+1}^{(i)}$ but also N are random variables.

We first conditional on N, then $(s_t, a_t, s_{t+1}^{(i)})$'s are independent samples for all i, s_t, a_t since any sample will not contain information about other samples. Therefore we can regroup $\mathcal{D}'$ into N independent samples with $\mathcal{D}' = \{X^{(i)} : i = 1, ..., N\}$ where $X^{(i)} = \{(s_t, a_t, s_{t+1}^{(i)}), t = 1, ..., H; s_t \in \mathcal{S}, a_t \in \mathcal{A}\}$. Now for any i_0, change $X^{(i_0)}$ to $X'^{(i_0)} = \{(s_t, a_t, s_{t+1}^{\prime(i_0)}), t = 1, ..., H; s_t \in \mathcal{S}, a_t \in \mathcal{A}\}$ and keep the rest $N-1$ data the same, use this data to create new estimator with state-action transition $\widetilde{d}'^{\pi}$, then we have

$$\begin{aligned}
&\left|\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle| - \sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle \widetilde{d}_t'^{\pi} - d_t^{\pi}, r_t\rangle|\right| \\
\le &\sup_{\pi\in\Pi}\left||\sum_{t=1}^{H}\langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle| - |\sum_{t=1}^{H}\langle \widetilde{d}_t'^{\pi} - d_t^{\pi}, r_t\rangle|\right| \\
\le &\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle - \sum_{t=1}^{H}\langle \widetilde{d}_t'^{\pi} - d_t^{\pi}, r_t\rangle\right| \\
= &\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle \widetilde{d}_t^{\pi} - \widetilde{d}_t'^{\pi}, r_t\rangle\right| \\
= &\sup_{\pi\in\Pi}\left|\sum_{h=2}^{H}\langle \widetilde{V}_h'^{\pi}, (\widetilde{T}_h - \widetilde{T}_h')\widetilde{d}_{h-1}^{\pi}\rangle + \langle \widetilde{V}_1'^{\pi}, \widetilde{d}_1^{\pi} - \widetilde{d}_1'^{\pi}\rangle\right|,
\end{aligned}$$

where the last equation comes from the trick that substitutes d_t^{π} by $\widetilde{d}_t'^{\pi}$ in Theorem B.2.1. By

[4]Here we do not include $r_t^{(i)}$ since the quantity $\sup_{\pi\in\Pi}|\sum_{t=1}^{H}\langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle|$ only contains the mean reward function r_t.

definition, the above equals to

$$
\begin{aligned}
&= \sup_{\pi\in\Pi} \left| \sum_{h=2}^{H} \langle \widehat{V}_h'^{\pi}, (\widehat{T}_h - \widehat{T}_h')\widehat{d}_{h-1}^{\pi} \rangle + \langle \widehat{V}_1'^{\pi}, \widehat{d}_1^{\pi} - \widehat{d}_1'^{\pi} \rangle \right| \cdot \mathbf{1}(E) \\
&\leq \sup_{\pi\in\Pi} \left(\sum_{h=2}^{H} ||(\widehat{T}_h - \widehat{T}_h')^T \widehat{V}_h'^{\pi}||_\infty ||\widehat{d}_{h-1}^{\pi}||_1 + |\langle \widehat{V}_1'^{\pi}, \widehat{d}_1^{\pi} - \widehat{d}_1'^{\pi} \rangle| \right) \cdot \mathbf{1}(E) \\
&\leq \sup_{\pi\in\Pi} \left(\sum_{h=2}^{H} ||(\widehat{T}_h - \widehat{T}_h')^T \widehat{V}_h'^{\pi}||_\infty + |\langle \widehat{V}_1'^{\pi}, \widehat{d}_1^{\pi} - \widehat{d}_1'^{\pi} \rangle| \right) \cdot \mathbf{1}(E)
\end{aligned}
$$

Note the change of a single $X^{(i_0)}$ will only change two entries of each row of $(\widehat{T}_h - \widehat{T}_h')^T$ by $1/N$ since with data $\mathcal{D}'$, $n_{s_t,a_t} = N$ for all s_t, a_t. Or in other words, given E,

$$
\widehat{T}_h^T - \widehat{T}_h'^T = \begin{bmatrix}
0 & ... & 0 & \frac{1}{N} & 0 & ... & -\frac{1}{N} & ... & 0 \\
0 & \frac{1}{N} & 0 & ... & -\frac{1}{N} & ... & ... & ... & 0 \\
..... & & & & & & & & \\
-\frac{1}{N} & 0 & ... & 0 & ... & ... & 0 & ... & \frac{1}{N}
\end{bmatrix},
$$

where the locations of $1/N, -1/N$ in each row are random as it depends on how different is $X'^{(i_0)}$ from $X^{(i_0)}$. However, based on this fact, it is enough for us to guarantee

$$
||(\widehat{T}_h - \widehat{T}_h')^T \widehat{V}_h'^{\pi}||_\infty \leq \frac{2}{N} ||\widehat{V}_h'^{\pi}||_\infty \leq \frac{2}{N}(H - h + 1) \leq \frac{2}{N} H
$$

and same result holds for $|\langle \widehat{V}_1'^{\pi}, \widehat{d}_1^{\pi} - \widehat{d}_1'^{\pi} \rangle| \leq 2H/N$ given N.

Combine all the results above, for a single change of $X^{(i_0)}$ we have

$$
\left| \sup_{\pi\in\Pi} | \sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t \rangle| - \sup_{\pi\in\Pi} | \sum_{t=1}^{H} \langle \widetilde{d}_t'^{\pi} - d_t^{\pi}, r_t \rangle| \right| \leq 2\frac{H^2}{N} \mathbf{1}(E) \leq 2\frac{H^2}{N}
$$

for any fixed N. If we let $Z = S(X^{(1)}, ..., X^{(N)}) = \sup_{\pi\in\Pi} | \sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi} - d_t^{\pi}, r_t \rangle|$, then for a given N by independence and above bounded difference condition we can apply Mcdiarmid inequality (where $\xi_i = 2H^2/N$) to obtain

$$
\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \epsilon | N) \leq 2e^{-\frac{N\epsilon^2}{2H^4}} := \frac{\delta}{2} \tag{B.7}
$$

Now note when $n > O(\frac{1}{d_m} \cdot \log(HSA/\delta))$, by Lemma B.2.1 we can obtain $N > nd_m/2$ with

probability $1-\delta/2$, combining this result and solving ϵ in (B.7), we have

$$\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle\right|\leq O(\sqrt{\frac{H^4\log(HSA/\delta)}{n\cdot d_m}})+\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle\right|\right]$$

with probability $1-\delta$.

Next before bounding $\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi}-d_t^{\pi},r_t\rangle\right|\right]$, we first give a useful lemma.

Let $\gamma\in(0,1)$ to be any threshold parameter. Then we first have the following lemma:

Lemma B.3.1. *Recall by definition $P_h(s_h,|s_{h-1},a_{h-1})=T_h(s_h,|s_{h-1},a_{h-1})$. It holds that with probability $1-\delta$, for all $t,s_t,a_t\in[H],\mathcal{S},\mathcal{A}$: if $P_h(s_h|s_{h-1},a_{h-1})\leq\gamma$, then*

$$\left|\widetilde{T}_h(s_h|s_{h-1},a_{h-1})-T_h(s_h|s_{h-1},a_{h-1})\right|\leq\sqrt{\frac{\gamma\log(HSA/\delta)}{2nd_m}}+\frac{2\log(HSA/\delta)}{3nd_m};$$

if $P_h(s_h,|s_{h-1},a_{h-1})>\gamma$, then

$$\left|\frac{\widetilde{T}_h(s_h|s_{h-1},a_{h-1})-T_h(s_h|s_{h-1},a_{h-1})}{T_h(s_h|s_{h-1},a_{h-1})}\right|\leq\sqrt{\frac{\log(HSA/\delta)}{2nd_m\gamma}}+\frac{2\log(HSA/\delta)}{3nd_m\gamma};$$

Proof: First consider the case where $P_h(s_h|s_{h-1},a_{h-1})\leq\gamma$.

$$\widetilde{T}_h(s_h|s_{h-1},a_{h-1})-T_h(s_h|s_{h-1},a_{h-1})=\frac{1}{n_{s_{h-1},a_{h-1}}}\sum_{i=1}^{n_{s_{h-1},a_{h-1}}}\left(\mathbf{1}[s_h^{(i)}]-T_h(s_h|s_{h-1},a_{h-1})\right)\mathbf{1}(E_h),$$

since $\mathrm{Var}[\mathbf{1}[s_h^{(i)}]|s_{h-1},a_{h-1}]=P_h(s_h|s_{h-1},a_{h-1})(1-P_h(s_h|s_{h-1},a_{h-1}))\leq P_h(s_h|s_{h-1},a_{h-1})\leq\gamma$, therefore by Bernstein Inequality,

$$\left|\widetilde{T}_h(s_h|s_{h-1},a_{h-1})-T_h(s_h|s_{h-1},a_{h-1})\right|\leq\mathbf{1}(E_h)\left(\sqrt{\frac{\gamma\log(1/\delta)}{n_{s_{h-1},a_{h-1}}}}+\frac{2\log(1/\delta)}{n_{s_{h-1},a_{h-1}}}\right)\leq\sqrt{\frac{\gamma\log(1/\delta)}{2nd_m}}+\frac{2\log(1/\delta)}{3nd_m};$$

Second, when $P_h(s_h|s_{h-1},a_{h-1})>\gamma$.

$$\frac{\widetilde{T}_h(s_h|s_{h-1},a_{h-1})-T_h(s_h|s_{h-1},a_{h-1})}{T_h(s_h|s_{h-1},a_{h-1})}=\frac{1}{n_{s_{h-1},a_{h-1}}}\sum_{i=1}^{n_{s_{h-1},a_{h-1}}}\left(\frac{\mathbf{1}[s_h^{(i)}]}{T_h(s_h|s_{h-1},a_{h-1})}-1\right)\mathbf{1}(E_h),$$

since

$$\mathrm{Var}\left[\frac{\mathbf{1}[s_h^{(i)}]}{T_h(s_h|s_{h-1},a_{h-1})}\middle|s_{h-1},a_{h-1}\right]\leq\frac{1}{T_h(s_h|s_{h-1},a_{h-1})^2}\mathrm{Var}\left[\mathbf{1}[s_h^{(i)}]\middle|s_{h-1},a_{h-1}\right]\leq\frac{1}{T_h(s_h|s_{h-1},a_{h-1})}\leq\frac{1}{\gamma},$$

and since $\frac{\mathbf{1}[s_h^{(i)}]}{T_h(s_h|s_{h-1},a_{h-1})} \le 1/\gamma$, again be Bernstein inequality we have

$$\left|\frac{\widetilde{T}_h(s_h|s_{h-1},a_{h-1}) - T_h(s_h|s_{h-1},a_{h-1})}{T_h(s_h|s_{h-1},a_{h-1})}\right| \le \sqrt{\frac{\log(1/\delta)}{2nd_m\gamma}} + \frac{2\log(1/\delta)}{3nd_m\gamma};$$

apply the union bound over t, s_t, a_t we obtain the stated result.

Bounding $\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^H \langle \widetilde{d}_t^\pi - d_t^\pi, r_t\rangle\right|\right]$. First note by Theorem B.2.1:

$$\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^H \langle \widetilde{d}_t^\pi - d_t^\pi, r_t\rangle\right|\right] \le \sum_{h=2}^H \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle v_h^\pi(s), ((\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right] + \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_1^\pi(s), (\widetilde{d}_1^\pi - d_1^\pi)(s)\rangle\right|\right],$$

so it suffices to bound each $\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_h^\pi(s), ((\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right]$. First of all,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_h^\pi(s), ((\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right] \\
=&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)(\widetilde{T}_h - T_h)(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\right|\right] \\
\le&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)(\widetilde{T}_h - T_h)(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\right| \cdot \mathbf{1}[T_h(s_h|s_{h-1},a_{h-1}) > \gamma]\right] \\
+&\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)(\widetilde{T}_h - T_h)(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\right| \cdot \mathbf{1}[T_h(s_h|s_{h-1},a_{h-1}) \le \gamma]\right] \\
=&\underbrace{\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)T_h(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\frac{\widetilde{T}_h - T_h}{T_h}(s_h|s_{h-1},a_{h-1})\right| \cdot \mathbf{1}[T_h > \gamma]\right]}_{(a)} \\
+&\underbrace{\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})(\widetilde{T}_h - T_h)(s_h|s_{h-1},a_{h-1})\right| \cdot \mathbf{1}[T_h(s_h|s_{h-1},a_{h-1}) \le \gamma]\right]}_{(b)},
\end{aligned}$$

Apply Lemma B.3.1 with $\delta'/2$ where $\delta' = \delta/H$, then

$$
\begin{aligned}
(a) \leq & \sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)T_h(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\left(\sqrt{\frac{\log(2HSA/\delta')}{2nd_m\gamma}}+\frac{2\log(2HSA/\delta')}{3nd_m\gamma}\right)\right|(1-\frac{\delta'}{2}) \\
& +H\delta'/2 \\
\leq & \sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)T_h(s_h|s_{h-1},a_{h-1})\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\left(\sqrt{\frac{\log(2H^2SA/\delta)}{2nd_m\gamma}}+\frac{2\log(2H^2SA/\delta)}{3nd_m\gamma}\right)\right| \\
& +\delta/2 \\
\leq & \sup_{\pi\in\Pi}\left|H\left(\sqrt{\frac{2\log(H^2SA/\delta)}{2nd_m\gamma}}+\frac{2\log(2H^2SA/\delta)}{3nd_m\gamma}\right)\right|+\delta/2 = H\left(\sqrt{\frac{\log(2H^2SA/\delta)}{2nd_m\gamma}}+\frac{2\log(2H^2SA/\delta)}{3nd_m\gamma}\right)+\delta/2,
\end{aligned}
$$

$$
\begin{aligned}
(b) \leq & \sup_{\pi\in\Pi}\left|\sum_{s_h,s_{h-1},a_{h-1}} V_h^\pi(s_h)\widetilde{d}_{h-1}^\pi(s_{h-1},a_{h-1})\left(\sqrt{\frac{\gamma\log(2HSA/\delta)}{2nd_m}}+\frac{2\log(2HSA/\delta)}{3nd_m}\right)\right|(1-\frac{\delta'}{2})+H\frac{\delta'}{2} \\
\leq & \sup_{\pi\in\Pi}\left|HS\left(\sqrt{\frac{\gamma\log(2H^2SA/\delta)}{2nd_m}}+\frac{2\log(2H^2SA/\delta)}{3nd_m}\right)\right|+\frac{\delta}{2} \\
= & HS\left(\sqrt{\frac{\gamma\log(2H^2SA/\delta)}{2nd_m}}+\frac{2\log(2H^2SA/\delta)}{3nd_m}\right)+\frac{\delta}{2},
\end{aligned}
$$

Hence we have for any γ,

$$
\begin{aligned}
& \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_h^\pi(s),((\widetilde{T}_h-T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right] \\
\leq & H\left(\sqrt{\frac{\log(2H^2SA/\delta)}{2nd_m\gamma}}+\frac{2\log(2H^2SA/\delta)}{3nd_m\gamma}\right)+HS\left(\sqrt{\frac{\gamma\log(2H^2SA/\delta)}{2nd_m}}+\frac{2\log(2H^2SA/\delta)}{3nd_m}\right)+\delta
\end{aligned}
$$

In particular, choose $\gamma = 1/S < 1$, then above becomes

$$
\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle V_h^\pi(s),((\widetilde{T}_h-T_h)\widetilde{d}_{h-1}^\pi)(s)\rangle\right|\right] \leq \sqrt{\frac{2H^2S\log(2H^2SA/\delta)}{nd_m}}+\frac{4HS\log(2H^2SA/\delta)}{3nd_m}+\delta
$$

Critically, above holds for any $\forall 1 > \delta > 0$. Based on theorem condition $n > c \cdot 1/d_m \log(HSA/\theta) > c \cdot 1/d_m$[5], choose $\delta = \frac{c}{nd_m}$, then above is further less equal to

$$\sqrt{\frac{2H^2S\log(2nH^2SA)}{nd_m}} + \frac{4HS\log(2nH^2SA)}{3nd_m} + \frac{c}{nd_m} \leq \sqrt{\frac{2H^2S\log(2nH^2SA)}{nd_m}} + C \cdot \frac{HS\log(2nH^2SA)}{3nd_m}$$

where C is a new constant absorbs $1/nd_m$. If we further reducing it to

Finally, summing over all H, and again using new constant C' to absorb higher order term, we obtain

$$\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle\right|\right] \leq C'\sqrt{\frac{H^4S\log(nHSA)}{nd_m}}$$

Combing this with Theorem B.3.1 and Lemma B.2.3, we have proved Theorem 3.5.1.

Remark 12. *The key for proving this uniform convergence bound is that applying concentration inequality only to terms that are independent of the policies, i.e.* $\widetilde{T}_h(s_h|s_{h-1}, a_{h-1}) - T_h(s_h|s_{h-1}, a_{h-1})$*. Therefore when taking supremum over policies, high probability event holds with same probability without decay.*

B.4 Proof of uniform convergence in OPE with deterministic policies using martingale concentration inequalities: Theorem 3.5.2

A reminder that all results in this section use data $\mathcal{D}$ for OPEMA estimator $\widehat{v}^{\pi}$.

B.4.1 Martingale concentration result on $\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle$.

Let $X = \sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle$ and $\mathcal{D}_h := \{s_t^{(i)}, a_t^{(i)} : t = 1, ..., h\}_{i=1}^{n}$. Since $\mathcal{D}_h$ forms a filtration, then by law of total expectation we have $X_t = \mathbb{E}[X|\mathcal{D}_t]$ is martingale. Moreover, we have

Lemma B.4.1.

$$X_t := \mathbb{E}[X|\mathcal{D}_t] = \sum_{h=2}^{t}\langle V_h^{\pi}, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi}\rangle + \langle V_1^{\pi}, \widetilde{d}_1^{\pi} - d_1^{\pi}\rangle.$$

Proof: [Proof of Lemma B.4.1] By martingale decomposition Theorem B.2.1 and note

[5]Note the θ in $\log(HSA/\theta)$ is identical to the failure probability in Theorem B.3.1

$\widetilde{T}_i, \widetilde{d}_i^\pi$ are measurable *w.r.t.* $\mathcal{D}_t$ for $i = 1, ..., t$, so we have

$$\mathbb{E}[X|\mathcal{D}_t] = \sum_{h=t+1}^{H} \mathbb{E}\left[\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle \middle| \mathcal{D}_t\right] + \sum_{h=2}^{t} \langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle + \langle V_1^\pi, (\widetilde{d}_1^\pi - d_1^\pi)\rangle.$$

Note for $h \geq t+1$, $\mathcal{D}_t \subset \mathcal{D}_{h-1}$, so by total law of expectation (tower property) we have

$$\begin{aligned}
&\mathbb{E}\left[\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle \middle| \mathcal{D}_t\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle \middle| \mathcal{D}_{h-1}\right] \middle| \mathcal{D}_t\right] \\
=&\mathbb{E}\left[\langle V_h^\pi, \mathbb{E}\left[(\widetilde{T}_h - T_h) \middle| \mathcal{D}_{h-1}\right] \widetilde{d}_{h-1}^\pi\rangle \middle| \mathcal{D}_t\right] = 0
\end{aligned}$$

where the last equality uses $\widetilde{T}_h$ is unbiased of T_h given $\mathcal{D}_{h-1}$. This gives the desired result.

Next we show martingale difference $|X_t - X_{t-1}|$ is bounded with high probability.

Lemma B.4.2. *With probability* $1 - \delta$*,*

$$\sup_t |X_t - X_{t-1}| \leq O(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot d_m}}).$$

Proof:

$$|X_t - X_{t-1}| = \langle V_t^\pi, (\widetilde{T}_t - T_t)\widetilde{d}_{t-1}^\pi\rangle \leq ||(\widetilde{T}_t - T_t)^T V_t^\pi||_\infty ||\widetilde{d}_{t-1}^\pi||_1 = ||(\widetilde{T}_t - T_t)^T V_t^\pi||_\infty.$$

For any fixed pair (s_t, a_t), we have

$$
\begin{aligned}
&((\widetilde{T}_t - T_t)^T V_t^\pi)(s_{t-1}, a_{t-1}) \\
=&\mathbf{1}(E_{t-1}) \cdot ((\widehat{T}_t - T_t)^T V_t^\pi)(s_{t-1}, a_{t-1}) \\
=&\mathbf{1}(E_{t-1}) \cdot \sum_{s_t} V_t^\pi(s_t)(\widehat{T}_t - T_t)(s_t|s_{t-1}, a_{t-1}) \\
=&\mathbf{1}(E_{t-1}) \cdot \left(\sum_{s_t} V_t^\pi(s_t)\widehat{T}_t(s_t|s_{t-1}, a_{t-1}) - \mathbb{E}[V_t^\pi|s_{t-1}, a_{t-1}] \right) \\
=&\mathbf{1}(E_{t-1}) \cdot \left(\sum_{s_t} V_t^\pi(s_t) \frac{1}{n_{s_{t-1},a_{t-1}}} \sum_{i=1}^{n} \mathbf{1}(s_t^{(i)} = s_t, s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)} = a_{t-1}) - \mathbb{E}[V_t^\pi|s_{t-1}, a_{t-1}] \right) \\
=&\mathbf{1}(E_{t-1}) \left(\frac{1}{n_{s_{t-1},a_{t-1}}} \sum_{i=1}^{n} V_t^\pi(s_t^{(i)}) \mathbf{1}(s_t^{(i)} = s_t, s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)} = a_{t-1}) - \mathbb{E}[V_t^\pi|s_{t-1}, a_{t-1}] \right) \\
=&\mathbf{1}(E_{t-1}) \left(\frac{1}{n_{s_{t-1},a_{t-1}}} \sum_{i\,:\,s_{t-1}^{(i)}=s_{t-1}, a_{t-1}^{(i)}=a_{t-1}} V_t^\pi(s_t^{(i)}) - \mathbb{E}[V_t^\pi|s_{t-1}, a_{t-1}] \right),
\end{aligned}
$$

where the fourth line uses the definition of $\widehat{T}_t$ and the fifth line uses the fact $\sum_{s_t} V_t^\pi(s_t)\mathbf{1}(s_t^{(i)} = s_t, s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)} = a_{t-1}) = V_t^\pi(s_t^{(i)})\mathbf{1}(s_t^{(i)} = s_t, s_{t-1}^{(i)} = s_{t-1}, a_{t-1}^{(i)} = a_{t-1})$.

Note $||V_t^\pi(\cdot)||_\infty \le H$ and also conditional on E_t, $n_{s_t,a_t} \ge n d_t^\mu(s_t, a_t)/2$, therefore by Hoeffding's inequality and a Union bound we obtain with probability $1 - \delta$

$$
\sup_t |X_t - X_{t-1}| \le O\left(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot \min_{t,s_t,a_t} d_t^\mu(s_t, a_t)}}\right) = O\left(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot d_m}}\right).
$$

Next we calculate the conditional variance of $\mathrm{Var}[X_{t+1}|\mathcal{D}_t]$.

Lemma B.4.3. *We have the following decomposition of conditional variance:*

$$
\mathrm{Var}[X_{t+1}|\mathcal{D}_t] = \sum_{s_t,a_t} \frac{\widetilde{d}_t^\pi(s_t, a_t)^2 \cdot \mathbf{1}(E_t)}{n_{s_t,a_t}} \cdot \mathrm{Var}[V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)} = s_t, a_t^{(1)} = a_t]
$$

Proof: Indeed,

$$\begin{aligned}
\mathrm{Var}[X_{t+1}|\mathcal{D}_t] &= \mathrm{Var}\left[\sum_{s_t,a_t}\sum_{s_{t+1}} V^{\pi}_{t+1}(s_{t+1})(\widetilde{T}-T)(s_{t+1}|s_t,a_t)\widetilde{d}^{\pi}_t(s_t,a_t)\middle|\mathcal{D}_t\right] \\
&= \sum_{s_t,a_t}\mathrm{Var}\left[\sum_{s_{t+1}} V^{\pi}_{t+1}(s_{t+1})(\widetilde{T}-T)(s_{t+1}|s_t,a_t)\middle|\mathcal{D}_t\right]\widetilde{d}^{\pi}_t(s_t,a_t)^2 \\
&= \sum_{s_t,a_t}\mathbf{1}(E_t)\cdot\mathrm{Var}\left[\sum_{s_{t+1}} V^{\pi}_{t+1}(s_{t+1})\widehat{T}(s_{t+1}|s_t,a_t)\middle|\mathcal{D}_t\right]\widetilde{d}^{\pi}_t(s_t,a_t)^2 \\
&= \sum_{s_t,a_t}\mathbf{1}(E_t)\cdot\mathrm{Var}\left[\sum_{s_{t+1}} V^{\pi}_{t+1}(s_{t+1})\frac{1}{n_{s_t,a_t}}\sum_{i=1}^{n}\mathbf{1}(s^{(i)}_{t+1}=s_{t+1},s^{(i)}_t=s_t,a^{(i)}_t=a_t)\middle|\mathcal{D}_t\right]\widetilde{d}^{\pi}_t(s_t,a_t)^2 \\
&= \sum_{s_t,a_t}\frac{\mathbf{1}(E_t)}{n^2_{s_t,a_t}}\cdot\mathrm{Var}\left[\sum_{i:s^{(i)}_t=s_t,a^{(i)}_t=a_t} V^{\pi}_{t+1}(s^{(i)}_{t+1})\middle|\mathcal{D}_t\right]\widetilde{d}^{\pi}_t(s_t,a_t)^2 \\
&= \sum_{s_t,a_t}\frac{\widetilde{d}^{\pi}_t(s_t,a_t)^2\cdot\mathbf{1}(E_t)}{n_{s_t,a_t}}\cdot\mathrm{Var}[V^{\pi}_{t+1}(s^{(1)}_{t+1})|s^{(1)}_t=s_t,a^{(1)}_t=a_t]
\end{aligned} \tag{B.8}$$

where the second equal sign comes from the fact that when conditional on $\mathcal{D}_t$, we can separate n episodes into SA groups and episodes from different groups are independent of each other. The third uses $\mathbf{1}(E_t)$ is measurable w.r.t $\mathcal{D}_t$. Similarly, the last equal sign again uses n_{s_t,a_t} episodes are independent given $\mathcal{D}_t$.

Lemma B.4.4 ([57] Lemma 3.4)**.** *For any policy π and any MDP.*

$$\begin{aligned}
\mathrm{Var}_{\pi}\left[\sum_{t=1}^{H} r^{(1)}_t\right] = \sum_{t=1}^{H}\Big(&\mathbb{E}_{\pi}\left[\mathrm{Var}\left[r^{(1)}_t + V^{\pi}_{t+1}(s^{(1)}_{t+1})\middle|s^{(1)}_t,a^{(1)}_t\right]\right] \\
&+\mathbb{E}_{\pi}\left[\mathrm{Var}\left[\mathbb{E}[r^{(1)}_t + V^{\pi}_{t+1}(s^{(1)}_{t+1})|s^{(1)}_t,a^{(1)}_t]\middle|s^{(1)}_t\right]\right]\Big).
\end{aligned}$$

This Lemma suggests if we can bound $\widetilde{d}^{\pi}_t$ by $O(d^{\pi}_t)$ with high probability, then by Lemma B.4.3 we have w.h.p

$$\sum_{t=1}^{H}\mathrm{Var}[X_{t+1}|\mathcal{D}_t] \le O(\frac{1}{nd_m}\cdot\sum_{t=1}^{H}\mathbb{E}[\mathrm{Var}[V^{\pi}_{t+1}(s^{(1)}_{t+1})|s^{(1)}_t,a^{(1)}_t]]) \le O(\frac{H^2}{nd_m})$$

Note this gives only H^2 dependence for $\sum_{t=1}^{H}\mathrm{Var}[X_{t+1}|\mathcal{D}_t]$ instead of a naive bound with H^3 and helps us to save a H factor.

Next we show how to bound $\widetilde{d}^{\pi}_t$.

B.4.2 Bounding $\widetilde{d}_t^{\pi}(s_t, a_t) - d_t^{\pi}(s_t, a_t)$

Our analysis is based on using martingale structure to derive bound on $\widetilde{d}_t^{\pi}(s_t, a_t) - d_t^{\pi}(s_t, a_t)$ for fixed t, s_t, a_t with probability $1 - \delta/HSA$, then use a union bound to get a bound for all $\widetilde{d}_t^{\pi}(s_t, a_t) - d_t^{\pi}(s_t, a_t)$ with probability $1 - \delta$.

Concretely, in (B.6) if we only extract the specific (s_t, a_t), then we have

$$\widetilde{d}_t^{\pi}(s_t, a_t) - d_t^{\pi}(s_t, a_t) = \sum_{h=2}^{t} (\Gamma_{h+1:t} \pi_h (\widetilde{T}_h - T_h) \widetilde{d}_{h-1}^{\pi})(s_t, a_t) + (\Gamma_{1:t} (\widetilde{d}_1^{\pi} - d_1^{\pi}))(s_t, a_t),$$

here $\widetilde{d}_t^{\pi}(s_t, a_t) - d_t^{\pi}(s_t, a_t)$ already forms a martingale with filtration $\mathcal{F}_t = \sigma(\mathcal{D}_t)$ and $(\Gamma_{h+1:t} \pi_h (\widetilde{T}_h - T_h) \widetilde{d}_{h-1}^{\pi})(s_t, a_t)$ is the corresponding martingale difference since

$$\mathbb{E}[(\Gamma_{h+1:t} \pi_h (\widetilde{T}_h - T_h) \widetilde{d}_{h-1}^{\pi})(s_t, a_t) | \mathcal{F}_{h-1}] = (\Gamma_{h+1:t} \pi_h \mathbb{E}[(\widetilde{T}_h - T_h) | \mathcal{F}_{h-1}] \widetilde{d}_{h-1}^{\pi})(s_t, a_t) = 0.$$

Now we fix specific (s_t, a_t). Then denote $(\Gamma_{h+1:t} \pi_h)(s_t, a_t) := \Gamma'_{h:t} \in \mathbb{R}^{1 \times S}$, then we have

$$|(\Gamma_{h+1:t} \pi_h (\widetilde{T}_h - T_h) \widetilde{d}_{h-1}^{\pi})(s_t, a_t)| = |\Gamma'_{h:t} (\widetilde{T}_h - T_h) \widetilde{d}_{h-1}^{\pi}| = |\langle (\widetilde{T}_h - T_h)^T \Gamma'^T_{h:t}, \widetilde{d}_{h-1}^{\pi} \rangle| \leq ||\Gamma'_{h:t} (\widetilde{T}_h - T_h)||_{\infty} \cdot 1.$$

Note here $\Gamma'_{h:t} (\widetilde{T}_h - T_h)$ is a row vector with dimension SA.

Bounding $||\Gamma'_{h:t} (\widetilde{T}_h - T_h)||_{\infty}$.

In fact, for any given (s_{h-1}, a_{h-1}), we have

$$\begin{aligned}
&\Gamma'_{h:t} (\widetilde{T}_h - T_h)(s_{h-1}, s_{h-1}) = \mathbf{1}(E_t) \cdot \Gamma'_{h:t} (\widehat{T}_h - T_h)(s_{h-1}, a_{h-1}) \\
=&\mathbf{1}(E_t) \cdot \Gamma'_{h:t} \left(\frac{1}{n_{s_{t-1}, a_{t-1}}} \sum_{i: s_{h-1}^{(i)} = s_{h-1}, a_{h-1}^{(i)} = a_{h-1}} \mathbf{e}_{s_h^{(i)}} - \mathbb{E}[\mathbf{e}_{s_h^{(1)}} | s_{h-1}^{(1)} = s_{h-1}, a_{h-1}^{(1)} = a_{h-1}] \right) \\
=&\mathbf{1}(E_t) \left(\frac{1}{n_{s_{t-1}, a_{t-1}}} \sum_{i: s_{h-1}^{(i)} = s_{h-1}, a_{h-1}^{(i)} = a_{h-1}} \Gamma'_{h:t} (s_h^{(i)}) - \mathbb{E}[\Gamma'_{h:t} (s_h^{(1)}) | s_{h-1}^{(1)} = s_{h-1}, a_{h-1}^{(1)} = a_{h-1}] \right)
\end{aligned}$$

Note by definition $\Gamma'_{h:t} (s_h^{(i)}) \leq 1$, since $(\Gamma_{h+1:t} \pi_h)(s_t, a_t) := \Gamma'_{h:t} \in \mathbb{R}^{1 \times S}$ and $\Gamma_{h+1:t}, \pi_h$ are just probability transitions. Therefore by Hoeffding's inequality and law of total expectation, we have

$$\mathbb{P}\left(|\Gamma'_{h:t}(\widetilde{T}_h - T_h)(s_{h-1}, a_{h-1})| > \epsilon\right) = \mathbb{P}\left(|\Gamma'_{h:t}(\widehat{T}_h - T_h)(s_{h-1}, a_{h-1})| > \epsilon \middle| E_t\right)$$
$$\leq \mathbb{E}\left[\exp(-\frac{2n_{s_{h-1},a_{h-1}} \cdot \epsilon^2}{1}) \middle| E_t\right] \leq \exp(-\frac{nd^{\mu}_{h-1}(s_{h-1}, a_{h-1}) \cdot \epsilon^2}{1})$$

and apply a union bound to get

$$\begin{aligned} &P(\sup_h ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty > \epsilon) \leq H \cdot \sup_h P(||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty > \epsilon) \\ \leq &HSA \cdot \sup_{h, s_{h-1}, a_{h-1}} \mathbb{P}\left(|\Gamma'_{h:t}(\widetilde{T}_h - T_h)(s_{h-1}, a_{h-1})| > \epsilon\right) \\ \leq &HSA \cdot \exp(-\frac{n \min d^{\mu}_{h-1}(s_{h-1}, a_{h-1}) \cdot \epsilon^2}{1}) := \frac{\delta}{HSA} \end{aligned} \tag{B.9}$$

Let the right hand side of (B.9) to be δ/HSA, then we have w.p. $1 - \delta/HSA$,

$$\sup_h ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty \leq O(\sqrt{\frac{1}{n \cdot d_m} \log \frac{H^2S^2A^2}{\delta}}). \tag{B.10}$$

Go back to bounding $\widetilde{d}^{\pi}_t(s_t, a_t) - d^{\pi}_t(s_t, a_t)$. By Azuma-Hoeffding's inequality (Lemma D.0.5), we have[6]

$$\mathbb{P}(|\widetilde{d}^{\pi}_t(s_t, a_t) - d^{\pi}_t(s_t, a_t)| > \epsilon) \leq \exp(-\frac{\epsilon^2}{\sum_{i=1}^t (\sup_h ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty)^2}) := \delta/HSA,$$

where $\sum_{i=1}^t (\sup_h ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty)^2$ is the sum of bounded square differences in Azuma-Hoeffding's inequality. Therefore we have w.p. $1 - \delta/HSA$,

$$|\widetilde{d}^{\pi}_t(s_t, a_t) - d^{\pi}_t(s_t, a_t)| \leq O(\sqrt{t \cdot (\sup_h ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty)^2 \log \frac{HSA}{\delta}}), \tag{B.11}$$

combining (B.10) with above we further have that w.p. $1 - 2\delta/HSA$,

$$|\widetilde{d}^{\pi}_t(s_t, a_t) - d^{\pi}_t(s_t, a_t)| \leq O(\sqrt{\frac{t}{nd_m} \log \frac{H^2S^2A^2}{\delta} \log \frac{HSA}{\delta}})$$

[6]To be more precise here we actually use a weaker version of Azuma-Hoeffding's inequality, see Remark 13.

Lastly, by a union bound and simple scaling (from 2δ to δ) we have w.p. $1-\delta$

$$\sup_t ||\widetilde{d}_t^{\pi} - d_t^{\pi}||_\infty \leq O(\sqrt{\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}}).$$

This implies that w.p. $1-\delta$, $\forall t, s_t, a_t$,

$$\widetilde{d}_t^{\pi}(s_t,a_t)^2 \leq 2d_t^{\pi}(s_t,a_t)^2 + O(\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}). \tag{B.12}$$

Combining (B.12) with Lemma B.4.4 and Lemma B.4.3, we obtain:

Lemma B.4.5. *With probability* $1-\delta$*,*

$$\sum_{t=1}^{H}\mathrm{Var}[X_{t+1}|\mathcal{D}_t] \leq O(\frac{H^2}{nd_m}) + O(\frac{H^4SA}{n^2d_m^2}\cdot\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}) \tag{B.13}$$

Proof: [Proof of Lemma B.4.5] By (B.12) and Lemma B.4.3, we have $\forall t$, with probability ay least $1-\delta$,

$$\begin{aligned}
&\mathrm{Var}[X_{t+1}|\mathcal{D}_t] \leq \sum_{s_t,a_t} O(\frac{\widetilde{d}_t^{\pi}(s_t,a_t)^2\cdot}{nd_m})\cdot\mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)}=s_t,a_t^{(1)}=a_t]\\
&\leq \sum_{s_t,a_t} O(\frac{1}{nd_m})\left(2d_t^{\pi}(s_t,a_t)^2 + O(\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta})\right)\cdot\mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)}=s_t,a_t^{(1)}=a_t]\\
&\leq \sum_{s_t,a_t} O(\frac{1}{nd_m})\left(2d_t^{\pi}(s_t,a_t) + O(\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta})\right)\cdot\mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)}=s_t,a_t^{(1)}=a_t]\\
&\leq O(\frac{1}{nd_m})\mathbb{E}\left[\mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)},a_t^{(1)}]\right] + O(\frac{1}{nd_m}\cdot\frac{H}{nd_m}\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta}\cdot H^2SA)\\
&= O(\frac{1}{nd_m})\mathbb{E}\left[\mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)},a_t^{(1)}]\right] + O(\frac{H^3SA}{n^2d_m^2}\cdot\log\frac{H^2S^2A^2}{\delta}\log\frac{HSA}{\delta})
\end{aligned}$$

then sum over t and apply Lemma B.4.4 gives the stated result.

Combining all the results, we are able to prove:

Theorem B.4.1. *With probability* $1-\delta$*, we have*

$$\left|\sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle\right| \leq O(\sqrt{\frac{H^2\log(HSA/\delta)}{nd_m}} + \sqrt{\frac{H^4SA\cdot\log(H^2S^2A^2/\delta)\log(HSA/\delta)}{n^2d_m^2}})$$

where $O(\cdot)$ *absorbs only the absolute constants.*

Proof: [Proof of Theorem B.4.1] Recall $X = \sum_{t=1}^{H}\langle\widetilde{d}_t^{\pi} - d_t^{\pi}, r_t\rangle$ and by law of total expecta-

tion it is easy to show $E[X] = 0$. Next denote $\sigma^2 = O(\frac{H^2}{nd_m}) + O(\frac{H^4SA}{n^2d_m^2} \cdot \log \frac{H^2S^2A^2}{\delta} \log \frac{HSA}{\delta})$ as in Lemma B.4.5 and also let $M = \sup_t |X_t - X_{t-1}|$. Then by Freedman inequality (Lemma D.0.6), we have with probability $1 - \delta/3$,

$$|X - \mathbb{E}[X]| \leq \sqrt{8\sigma^2 \cdot \log(3/\delta)} + \frac{2M}{3} \cdot \log(3/\delta), \quad \text{Or} \quad W \geq \sigma^2.$$

where $W = \sum_{t=1}^{H} \mathrm{Var}[X_{t+1}|\mathcal{D}_t]$. Next by Lemma B.4.5, we have $\mathbb{P}(W \geq \sigma^2) \leq 1/3\delta$, this implies with probability $1 - 2\delta/3$,

$$|X - \mathbb{E}[X]| \leq \sqrt{8\sigma^2 \cdot \log(3/\delta)} + \frac{2M}{3} \cdot \log(3/\delta).$$

Finally, by Lemma B.4.2, we have $\mathbb{P}(M \geq O(\sqrt{\frac{H^2 \log(HSA/\delta)}{nd_m}})) \leq \delta/3$. Also use $\mathbb{E}[X] = 0$, we have with probability $1 - \delta$,

$$|X| \leq \sqrt{8\sigma^2 \cdot \log(3/\delta)} + O(\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{nd_m}} \log(3/\delta)).$$

Plugging back the expression of $\sigma^2 = O(\frac{H^2}{nd_m}) + O(\frac{H^4SA}{n^2d_m^2} \cdot \log \frac{H^2S^2A^2}{\delta} \log \frac{HSA}{\delta})$ and assimilating the same order terms give the desired result.

Remark 13. *Rigorously, standard Azuma-Hoeffding's inequality does not apply to* (B.11) *since* $\sup_h ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty$ *is not a deterministic upper bound, we only have the difference bound with high probability sense, see* (B.10). *Therefore, strictly speaking, we need to apply Theorem 32 in [113] which is a weaker Azuma-Hoeffding's inequality allowing bounded difference with high probability. The same logic applies for a weaker freedman's inequality consisting of Theorem 34 and Theorem 37 in [113] since our martingale difference* $M = \sup_t |X_t - X_{t-1}|$ *in the proof of Theorem B.4.1 is bounded with high probability. We avoid explicitly using them in order to make our proofs more readable for our readers.*

We end this section by giving the proofs of Theorem 3.5.2 and Theorem 3.5.2.

Proof: [Proof of Lemma 3.5.2 and Theorem 3.5.2] The proof of Lemma 3.5.2 comes from Lemma B.2.2, Lemma B.2.3 and Theorem B.4.1. The proof of Theorem 3.5.2 relies on applying a union bound over Π in Theorem 3.5.2 (recall all non-stationary deterministic policies have $|\Pi| = A^{HS}$), then extra dependence of $\sqrt{\log(|\Pi|)} = \sqrt{HS \log(A)}$ pops out. Note that the higher order term has two trailing log terms (see the right hand side of (B.13)), so when replacing δ by $\delta/|\Pi|$ with a union bound, both terms will give extra $\sqrt{HS}$ dependence so in higher order term we have extra HS dependence but not just $\sqrt{HS}$.

B.5 Proof of uniform convergence problem with local policy class.

In this section, we consider using OPEMA estimator with data $\mathcal{D}'$. Also, WLOG we only consider deterministic reward (as implied by Lemma B.2.3 random reward only causes lower order dependence). Also, we fix $N > 0$ for the moment. First recall for all $t = 1, ..., H$

$$V_t^\pi(s_t) = \mathbb{E}_\pi \left[\sum_{t'=t}^{H} r_{t'}(s_{t'}^{(1)}, a_t^{(1)}) \middle| s_t^{(1)} = s_t \right]$$
$$Q_t^\pi(s_t, a_t) = \mathbb{E}_\pi \left[\sum_{t'=t}^{H} r_{t'}(s_{t'}^{(1)}, a_t^{(1)}) \middle| s_t^{(1)} = s_t, a_t^{(1)} = a_t \right]$$

where $r_t(s, a)$ are deterministic rewards and $s_t^{(1)}, a_t^{(1)}$ are random variables. Consider V_t^π, Q_t^π as vectors, then by standard Bellman equations we have for all $t = 1, ..., H$ (define $V_{H+1} = Q_{H+1} = 0$)

$$Q_t^\pi = r_t + P_{t+1}^\pi Q_{t+1}^\pi = r_t + P_{t+1} V_{t+1}^\pi, \tag{B.14}$$

where $P_t^\pi \in \mathbb{R}^{(SA)\times(SA)}$ is the state-action transition and $P_t(\cdot|\cdot,\cdot) \in \mathbb{R}^{(SA)\times S}$ is the transition probabilities. Also, we have bellman optimality equations:

$$Q_t^\star = r_t + P_{t+1} V_{t+1}^\star, \qquad V_t^\star(s_t) := \max_{a_t} Q_t^\star(s_t, a_t), \quad \pi_t^\star(s_t) := \operatorname*{argmax}_{a_t} Q_t^\star(s_t, a_t) \ \ \forall s_t \tag{B.15}$$

where $\pi^\star$ is one optimal deterministic policy. The corresponding Bellman equations and Bellman optimality equations for empirical MDP $\widehat{M}$ are defined similarly. Since we consider deterministic rewards, by Bellman equations we have

$$\widehat{Q}_t^\pi - Q_t^\pi = \widehat{P}_{t+1}^\pi \widehat{Q}_{t+1}^\pi - P_{t+1}^\pi Q_{t+1}^\pi = (\widehat{P}_{t+1}^\pi - P_{t+1}^\pi)\widehat{Q}_{t+1}^\pi + P_{t+1}^\pi(\widehat{Q}_{t+1}^\pi - Q_{t+1}^\pi)$$

for $t = 1, ..., H$. By writing it recursively, we have $\forall t = 1, ..., H-1$

$$\begin{aligned}\widehat{Q}_t^\pi - Q_t^\pi &= \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi (\widehat{P}_h^\pi - P_h^\pi)\widehat{Q}_h^\pi \\ &= \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^\pi (\widehat{P}_h - P_h)\widehat{V}_h^\pi\end{aligned}$$

where $\Gamma_{t:h}^\pi = \prod_{i=t}^{h} P_i^\pi$ is the multi-step state-action transition and $\Gamma_{t+1:t}^\pi := I$.

Note $\widehat{\pi}^*$ to be the empirical optimal policy over $\widehat{M}$, we are interested in how to obtain uniform convergence for any policy π that is close to $\widehat{\pi}^*$. More precisely, in this section we

consider the policy class Π_1 to be:

$$\Pi_1 := \{\pi : s.t. \ ||\widehat{V}_t^{\pi} - \widehat{V}_t^{\widehat{\pi}^\star}||_\infty \leq \epsilon_{\text{opt}}, \ \forall t = 1, ..., H\}$$

where $\epsilon_{\text{opt}} \geq 0$ is a parameter decides how large the policy class is. We now assume $\widehat{\pi}$ to be any policy within Π_1 throughout this section. **Also, $\widehat{\pi}$ may be a policy learned from a learning algorithm using the data $\mathcal{D}$. In this case, $\widehat{\pi}$ may not be independent of $\widehat{P}$.**

We start with the following simple calculation:[7]

$$\begin{aligned}\left|\widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}}\right| &\leq \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left|(\widehat{P}_h - P_h)\widehat{V}_h^{\widehat{\pi}}\right| \\ &\leq \underbrace{\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left|(\widehat{P}_h - P_h)\widehat{V}_h^{\widehat{\pi}^\star}\right|}_{(***)} + \underbrace{\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left|(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})\right|}_{(****)}\end{aligned} \tag{B.16}$$

We now analyze $(***)$ and $(****)$.

B.5.1 Analyzing $\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left|(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})\right|$

First, by vector induced matrix norm[8] we have

$$\begin{aligned}\left\|\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\widehat{\pi}} \cdot \left|(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})\right|\right\|_\infty &\leq H \cdot \sup_h \left\|\Gamma_{t+1:h-1}^{\widehat{\pi}}\right\|_\infty \left\||(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})|\right\|_\infty \\ &\leq H \cdot \sup_h \left\||(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})|\right\|_\infty\end{aligned}$$

where the last equal sign uses multi-step transition $\Gamma_{t+1:h-1}^{\pi}$ is row-stochastic. Note given N, $\widehat{P}_t(\cdot|\cdot,\cdot)$ all have N in the denominator. Therefore, by Hoeffding inequality and a union bound we have with probability $1-\delta$,

$$\sup_{t,s_t,s_{t-1},a_{t-1}} |\widehat{P}_t(s_t|s_{t-1},a_{t-1}) - P_t(s_t|s_{t-1},a_{t-1})| \leq O(\sqrt{\frac{\log(HSA/\delta)}{N}}),$$

this indicates

$$\sup_h \left\||(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})|\right\|_\infty \leq \epsilon_{\text{opt}} \cdot \sup_h \left\||\widehat{P}_h - P_h| \cdot \mathbf{1}\right\|_\infty \leq \epsilon_{\text{opt}} \cdot O(S\sqrt{\frac{\log(HSA/\delta)}{N}}),$$

where $\mathbf{1} \in \mathbb{R}^S$ is all-one vector. To sum up, we have

[7] Since all quantities in the calculation are vectors, so the absolute value $|\cdot|$ used is point-wise operator.

[8] For A a matrix and x a vector we have $\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$.

Lemma B.5.1. *Fix $N > 0$, we have with probability $1-\delta$, for all $t = 1, ..., H-1$*

$$\sum_{h=t+1}^{H} \Gamma^{\hat{\pi}}_{t+1:h-1} \left| (\widehat{P}_h - P_h)(\widehat{V}_h^{\hat{\pi}^\star} - \widehat{V}_h^{\hat{\pi}}) \right| \le \epsilon_{\text{opt}} \cdot O\left(\sqrt{\frac{H^2 S^2 \log(HSA/\delta)}{N}} \cdot \mathbf{1} \right)$$

Now we consider (∗∗∗).

B.5.2 Analyzing $\sum_{h=t+1}^{H} \Gamma^{\hat{\pi}}_{t+1:h-1} \left| (\widehat{P}_h - P_h)\widehat{V}_h^{\hat{\pi}^\star} \right|$.

Lemma B.5.2. *Given N, we have with probability $1-\delta$, $\forall t = 1, ..., H-1$*

$$\sum_{h=t+1}^{H} \Gamma^{\hat{\pi}}_{t+1:h-1} \left| (\widehat{P}_h - P_h)\widehat{V}_h^{\hat{\pi}^\star} \right| \le \sum_{h=t+1}^{H} \Gamma^{\hat{\pi}}_{t+1:h-1} \left(4\sqrt{\frac{\log(HSA/\delta)}{N}} \sqrt{\text{Var}(\widehat{V}_h^{\hat{\pi}^\star})} + \frac{4(H-t)}{3N} \log(\frac{HSA}{\delta}) \cdot \mathbf{1} \right)$$

where $\text{Var}(v_t^\pi) \in \mathbb{R}^{SA}$ and $\text{Var}(V_t^\pi)(s_{t-1}, a_{t-1}) = \text{Var}_{s_t}[V_t^\pi(\cdot)|s_{t-1}, a_{t-1}]$ and $|\cdot|, \sqrt{\cdot}$ are pointwise operator.

Proof: [Proof of Lemma B.5.2] The key point is to guarantee $\widehat{P}_h$ is independent of $\widehat{V}_h^{\hat{\pi}^\star}$ so that we can apply Bernstein inequality w.r.t the randomness in $\widehat{P}_h$. In fact, note given N all data pairs in $\mathcal{D}'$ are independent of each other, and $\widehat{P}_h$ only uses data from $h-1$ to h. Moreover, $\widehat{V}_h^{\hat{\pi}^\star}$ only uses data from time h to H since $\widehat{V}_h^{\pi}$ uses data from h to H by bellman equation (B.14) for any π and optimal policy $\hat{\pi}^\star_{h:H}$ also only uses data from h to H by bellman optimality equation (B.15).

Then by Bernstein inequality, with probability $1-\delta$

$$\left| (\widehat{P}_h - P_h)\widehat{V}_h^{\hat{\pi}^\star} \right| (s_{t-1}, a_{t-1}) \le 4\sqrt{\frac{\log(1/\delta)}{N}} \sqrt{\text{Var}(\widehat{V}_h^{\hat{\pi}^\star})(s_{t-1}, a_{t-1})} + \frac{4(H-t)}{3N} \log(\frac{1}{\delta})$$

apply a union bound and take the sum we get the stated result.

Now combine Lemma C.1.5 and Lemma B.5.2 we obtain with probability $1-\delta$, for all $t = 1, ..., H-1$

$$\begin{aligned}
\left| \widehat{Q}_t^{\hat{\pi}} - Q_t^{\hat{\pi}} \right| &\le \sum_{h=t+1}^{H} \Gamma^{\hat{\pi}}_{t+1:h-1} \left(4\sqrt{\frac{\log(HSA/\delta)}{N}} \sqrt{\text{Var}(\widehat{V}_h^{\hat{\pi}^\star})} + \frac{4(H-t)}{3N} \log(\frac{HSA}{\delta}) \cdot \mathbf{1} \right) \\
&\quad + c_1 \epsilon_{\text{opt}} \cdot \sqrt{\frac{H^2 S^2 \log(HSA/\delta)}{N}} \cdot \mathbf{1} \\
&\le 4\sqrt{\frac{\log(HSA/\delta)}{N}} \sum_{h=t+1}^{H} \Gamma^{\hat{\pi}}_{t+1:h-1} \sqrt{\text{Var}(\widehat{V}_h^{\hat{\pi}^\star})} + \frac{4H^2}{3N} \log(\frac{HSA}{\delta}) \cdot \mathbf{1} \\
&\quad + c_1 \epsilon_{\text{opt}} \cdot \sqrt{\frac{H^2 S^2 \log(HSA/\delta)}{N}} \cdot \mathbf{1},
\end{aligned} \tag{B.17}$$

Next note $\sqrt{\mathrm{Var}(\cdot)}$ is a norm, therefore by norm triangle inequality we have

$$\begin{aligned}\sqrt{\mathrm{Var}(\widehat{V}_h^{\hat{\pi}^\star})} &\le \sqrt{\mathrm{Var}(\widehat{V}_h^{\hat{\pi}^\star} - \widehat{V}_h^{\hat{\pi}})} + \sqrt{\mathrm{Var}(\widehat{V}_h^{\hat{\pi}} - V_h^{\hat{\pi}})} + \sqrt{\mathrm{Var}(V_h^{\hat{\pi}})} \\ &\le \left\|\widehat{V}_h^{\hat{\pi}^\star} - \widehat{V}_h^{\hat{\pi}}\right\|_\infty \cdot \mathbf{1} + \left\|\widehat{V}_h^{\hat{\pi}} - V_h^{\hat{\pi}}\right\|_\infty \cdot \mathbf{1} + \sqrt{\mathrm{Var}(V_h^{\hat{\pi}})} \\ &\le \epsilon_{\mathrm{opt}} \cdot \mathbf{1} + \left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty \cdot \mathbf{1} + \sqrt{\mathrm{Var}(V_h^{\hat{\pi}})}\end{aligned}$$

Plug this into (B.17) to obtain

$$\begin{aligned}\left|\widehat{Q}_t^{\hat{\pi}} - Q_t^{\hat{\pi}}\right| \le 4\sqrt{\frac{\log(HSA/\delta)}{N}} \sum_{h=t+1}^{H} \left(\Gamma_{t+1:h-1}^{\hat{\pi}} \sqrt{\mathrm{Var}(V_h^{\hat{\pi}})} + \left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty \cdot \mathbf{1}\right) + \frac{4H^2}{3N} \log(\frac{HSA}{\delta}) \cdot \mathbf{1} \\ + c_2 \epsilon_{\mathrm{opt}} \cdot \sqrt{\frac{H^2 S^2 \log(HSA/\delta)}{N}} \cdot \mathbf{1}.\end{aligned} \tag{B.18}$$

Next lemma helps us to bound $\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\hat{\pi}} \sqrt{\mathrm{Var}(V_h^{\hat{\pi}})}$.

Lemma B.5.3. *A conditional version of Lemma B.4.4 holds:*

$$\begin{aligned}\mathrm{Var}_\pi \left[\sum_{t=h}^{H} r_t^{(1)} \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right] = \sum_{t=h}^{H} \Bigg(\mathbb{E}_\pi \left[\mathrm{Var}\left[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)}) \middle| s_t^{(1)}, a_t^{(1)}\right] \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right] \\ + \mathbb{E}_\pi \left[\mathrm{Var}\left[\mathbb{E}[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)}) | s_t^{(1)}, a_t^{(1)}] \middle| s_t^{(1)}\right] \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right] \Bigg).\end{aligned} \tag{B.19}$$

and by using (B.19) *we can show*

$$\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\hat{\pi}} \sqrt{\mathrm{Var}(V_h^{\hat{\pi}})} \le \sqrt{(H-t)^3} \cdot \mathbf{1}.$$

Proof: The proof of (B.19) uses the identical trick as Lemma B.4.4 except the total law of variance is replaced by the total law of conditional variance.

Moreover, recall $\Gamma_{t+1:h-1}^{\hat{\pi}} = \prod_{i=t+1}^{h-1} P_i^{\hat{\pi}}$ is the multi-step transition, so for any pair (s_t, a_t),

$$
\begin{aligned}
&\sum_{h=t+1}^{H}\left(\Gamma^{\hat{\pi}}_{t+1:h-1}\sqrt{\mathrm{Var}(V_h^{\hat{\pi}})}\right)(s_t,a_t)\\
=&\sum_{h=t+1}^{H}\sum_{s_{h-1},a_{h-1}}\sqrt{\mathrm{Var}[V_h^{\hat{\pi}}|s_{h-1},a_{h-1}]}d_t^{\hat{\pi}}(s_{h-1},a_{h-1}|s_t,a_t)\\
=&\sum_{h=t+1}^{H}\sum_{s_{h-1},a_{h-1}}\sqrt{\mathrm{Var}[V_h^{\hat{\pi}}|s_{h-1},a_{h-1}]d_t^{\hat{\pi}}(s_{h-1},a_{h-1}|s_t,a_t)}\cdot\sqrt{d_t^{\hat{\pi}}(s_{h-1},a_{h-1}|s_t,a_t)}\\
\leq&\sum_{h=t+1}^{H}\sqrt{\sum_{s_{h-1},a_{h-1}}\mathrm{Var}[V_h^{\hat{\pi}}|s_{h-1},a_{h-1}]d_t^{\hat{\pi}}(s_{h-1},a_{h-1}|s_t,a_t)\cdot\sum_{s_{h-1},a_{h-1}}d_t^{\hat{\pi}}(s_{h-1},a_{h-1}|s_t,a_t)}\\
=&\sum_{h=t+1}^{H}\sqrt{\sum_{s_{h-1},a_{h-1}}\mathrm{Var}[V_h^{\hat{\pi}}|s_{h-1},a_{h-1}]d_t^{\hat{\pi}}(s_{h-1},a_{h-1}|s_t,a_t)}\\
=&\sum_{h=t+1}^{H}\sqrt{\mathbb{E}_{\hat{\pi}}\left[\mathrm{Var}[V_h^{\hat{\pi}}|s_{h-1}^{(1)},a_{h-1}^{(1)}]\middle|s_t,a_t\right]}\\
=&\sum_{h=t+1}^{H}\sqrt{1}\cdot\sqrt{\mathbb{E}_{\hat{\pi}}\left[\mathrm{Var}[V_h^{\hat{\pi}}|s_{h-1}^{(1)},a_{h-1}^{(1)}]\middle|s_t,a_t\right]}\\
\leq&\sqrt{(H-t)\sum_{h=t+1}^{H}\mathbb{E}_{\hat{\pi}}\left[\mathrm{Var}[V_h^{\hat{\pi}}|s_{h-1}^{(1)},a_{h-1}^{(1)}]\middle|s_t,a_t\right]}\\
\leq&\sqrt{(H-t)\cdot\mathrm{Var}_{\hat{\pi}}\left[\sum_{h=t+1}^{H}r_h^{(1)}\middle|s_t^{(1)}=s_t,a_t^{(1)}=a_t\right]}\leq\sqrt{(H-t)^3}
\end{aligned}
$$

where all the inequalities are Cauchy-Schwarz inequalities.

Apply Lemma B.5.3 to bound (B.18), and use ∞ norm on both sides, we obtain

Theorem B.5.1. *Conditional on $N > 0$, then with probability $1-\delta$, we have for all $t = 1, ..., H-1$*

$$
\begin{aligned}
\left\|\widehat{Q}_t^{\hat{\pi}}-Q_t^{\hat{\pi}}\right\|_\infty \leq 4\sqrt{\frac{H^3\log(HSA/\delta)}{N}}+4\sqrt{\frac{\log(HSA/\delta)}{N}}\sum_{h=t+1}^{H}\left\|\widehat{Q}_h^{\hat{\pi}}-Q_h^{\hat{\pi}}\right\|_\infty+\frac{4H^2}{3N}\log(\frac{HSA}{\delta})\\
+c_2\epsilon_{opt}\cdot\sqrt{\frac{H^2S^2\log(HSA/\delta)}{N}}.
\end{aligned}
$$

Then by using backward induction and Theorem B.5.1, we have the following:

Theorem B.5.2. *Suppose $N \geq 64H^2 \cdot \log(HSA/\delta)$ and $\epsilon_{opt} \leq \sqrt{H}/S$, then we have with probability $1-\delta$,*

$$\left\|\widehat{Q}_1^{\hat{\pi}} - Q_1^{\hat{\pi}}\right\|_\infty \leq 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}$$

where c_2 is the same constant in Theorem B.5.1.

Proof: Under the condition, by Theorem B.5.1 it is easy to check for all $t = 1, ..., H-1$ with probability $1-\delta$,

$$\left\|\widehat{Q}_t^{\hat{\pi}} - Q_t^{\hat{\pi}}\right\|_\infty \leq (5+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{\log(HSA/\delta)}{N}}\sum_{h=t+1}^{H}\left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty,$$

which we conditional on.

For $t = H-1$, we have

$$\begin{aligned}\left\|\widehat{Q}_{H-1}^{\hat{\pi}} - Q_{H-1}^{\hat{\pi}}\right\|_\infty \leq& (5+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{\log(HSA/\delta)}{N}}\left\|\widehat{Q}_H^{\hat{\pi}} - Q_H^{\hat{\pi}}\right\|_\infty \\ \leq& (5+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{H^2\log(HSA/\delta)}{N}} \\ \leq& (9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}\end{aligned}$$

Suppose $\left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty \leq 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}$ holds for all $h = t+1, ..., H$, then for $h = t$, we have

$$\begin{aligned}\left\|\widehat{Q}_t^{\hat{\pi}} - Q_t^{\hat{\pi}}\right\|_\infty &\leq (5+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{\log(HSA/\delta)}{N}}\sum_{h=t+1}^{H}\left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty \\ &\leq (9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + 4\sqrt{\frac{(H-1)^2\log(HSA/\delta)}{N}}\cdot 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}} \\ &\leq 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}\end{aligned}$$

where the last line uses the condition $N \geq 64H^2 \cdot \log(HSA/\delta)$. By induction, we have the result.

Proof: [Proof of Theorem 3.5.3] By Theorem B.5.2 we have for $N \geq c \cdot H^2 \cdot \log(HSA/\delta)$,

$$\mathbb{P}\left(\left\|\widehat{Q}_1^{\hat{\pi}} - Q_1^{\hat{\pi}}\right\|_\infty \geq 2(9+c_2)\sqrt{\frac{H^3\log(HSA/\delta)}{N}}\middle| N\right) \leq \delta$$

The only thing left is to use Lemma B.2.1 to bound the event that $\{N < nd_m/2\}$ has small

probability.

Last but not least, the condition $n > c_1 H^2 \log(HSA/\delta)/d_m$ is sufficient for applying Lemma B.2.1 and it also implies $N \geq c \cdot H^2 \cdot \log(HSA/\delta)$ (the condition of Theorem B.5.2) when $N \geq nd_m/2$ since:

$$n > c_1 H^2 \log(HSA/\delta)/d_m \Rightarrow nd_m/2 \geq c_2 H^2 \log(HSA/\delta)$$

which implies $N \geq c_2 \cdot H^2 \cdot \log(HSA/\delta)$ when $N \geq nd_m/2$.

B.6 Proof of uniform convergence lower bound.

In this section we prove a uniform convergence OPE lower bound of $\Omega(H^3/d_m\epsilon^2)$. Conceptually, uniform convergence lower bound can be derived by a reduction to the lower bound of identifying the ϵ-optimal policy. There are quite a few literature that provide information theoretical lower bounds in different setting, *e.g.* [114, 65, 115, 96, 42]. However, to the best of our knowledge, there is no result proven for the non-stationary transition finite horizon episodic setting with bounded rewards. For example, [42] prove the result sample complexity lower bound of $\Omega(H^3SA/\epsilon^2)$ with stationary MDP and their proof cannot be directly applied to non-stationary setting as they reduce the problem to infinite horizon discounted setting which always has stationary transitions. [114] prove the episode complexity of $\widetilde{\Omega}(H^2SA/\epsilon^2)$ for the stationary transition setting. [96] prove the $\Omega(\sqrt{H^2SAT})$ regret lower bound for non-stationary finite horizon online setting but it is not clear how to translate the regret to PAC-learning setting by keeping the same sample complexity optimality. [65] prove the $\Omega(HSA/\epsilon^2)$ lower bound for the non-stationary finite horizon offline episodic setting where they assume $\sum_{i=1}^{H} r_i \leq 1$ and this is also different from our setting since we have $0 \leq r_t \leq 1$ for each time step.

Our proof consists of three steps. **1.** We will first show a minimax lower bound (**over all MDP instances**) for learning ϵ-optimal policy is $\Omega(H^3SA/\epsilon^2)$; **2.** Based on 1, we can further show a minimax lower bound (**over problem class $\mathcal{M}_{d_m}$**) for learning ϵ-optimal policy is $\Omega(H^3/d_m\epsilon^2)$; **3.** prove the uniform convergence OPE lower bound of the same rate.

B.6.1 Information theoretical lower sample complexity bound over all MDP instances for identifying ϵ-optimal policy.

In fact, a modified construction of Theorem 5 in [65] is our tool for obtaining $\Omega(H^3SA/\epsilon^2)$ lower bound. We can get the additional H^2 factor by using $\sum_{i=1}^{H} r_i$ can be of order $O(H)$.

Theorem B.6.1. *Given $H \geq 2$, $A \geq 2$, $0 < \epsilon < \frac{1}{48\sqrt{8}}$ and $S \geq c_1$ where c_1 is a universal constant. Then there exists another universal constant c such that for any algorithm and any $n \leq cH^3SA/\epsilon^2$, there exists a non-stationary H horizon MDP with probability at least $1/12$, the algorithm outputs a policy $\widehat{\pi}$ with $v^\star - v^{\widehat{\pi}} \geq \epsilon$.*

Like in [65], the proof relies on embedding $\Theta(HS)$ independent multi-arm bandit problems into a hard-to-learn MDP so that any algorithm that wants to output a near-optimal policy needs

to identify the best action in $\Omega(HS)$ problems. However, in our construction we make a further modification of [65] so that there is **no** waiting states, which is crucial for the reduction from offline family. We also double the length of the hard-to-learn MDP instance so that the latter half uses a "naive" copy construction which is uninformative. The uninformative extension will help to produce the additional H^2 factor.

Proof: [Proof of Theorem B.6.2] We construct a non-stationary MDP with S states per level, A actions per state and has horizon $2H$. At each time step, states are categorized into four types with two special states g_h, b_h and the remaining $S-2$ "bandit" states denoted by $s_{h,i}$, $i \in [S-2]$. Each bandit state has an unknown best action $a^\star_{h,i}$ that provides the highest expected reward comparing to other actions.

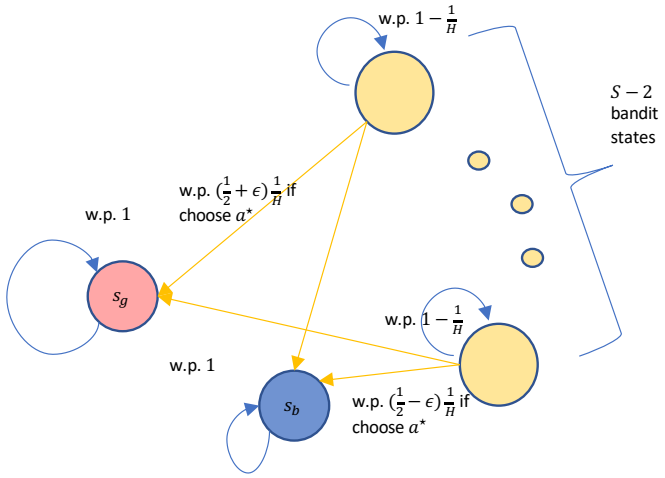

Figure B.1: An illustration of State-space transition diagram

The transition dynamics are defined as follows:

- for $h = 1, ..., H-1$,
  - For bandit states $b_{h,i}$, there is probability $1-\frac{1}{H}$ to transition to $b_{h+1,i}$ regardless of the action chosen. For the rest of $\frac{1}{H}$ probability, optimal action $a^\star_{h,i}$ will have probability $\frac{1}{2}+\tau$ or $\frac{1}{2}-\tau$ transition to g_{h+1} or b_{h+1} and all other actions a will have equal probability $\frac{1}{2}$ for either g_{h+1} or b_{h+1}, where τ is a parameter will be decided later. Or equivalently,

$$\mathbb{P}(\cdot|s_{h,i}, a^\star_{h,i}) = \begin{cases} 1-\frac{1}{H} & \text{if } \cdot = s_{h+1,i} \\ (\frac{1}{2}+\tau)\cdot\frac{1}{H} & \text{if } \cdot = g_{h+1} \\ (\frac{1}{2}-\tau)\cdot\frac{1}{H} & \text{if } \cdot = b_{h+1} \end{cases} \quad \mathbb{P}(\cdot|s_{h,i}, a) = \begin{cases} 1-\frac{1}{H} & \text{if } \cdot = s_{h+1,i} \\ \frac{1}{2}\cdot\frac{1}{H} & \text{if } \cdot = g_{h+1} \\ \frac{1}{2}\cdot\frac{1}{H} & \text{if } \cdot = b_{h+1} \end{cases}$$

– g_h always transitions to g_{h+1} and b_h always transitions to b_{h+1}, *i.e.* for all $a \in \mathcal{A}$, we have

$$\mathbb{P}(g_{h+1}|g_h, a) = 1, \quad \mathbb{P}(b_{h+1}|b_h, a) = 1.$$

We will determine parameter τ at the end of the proof.

- for $h = H, ..., 2H-1$, all states will always transition to the same type of states for the next step, *i.e.* $\forall a \in \mathcal{A}$,

$$\mathbb{P}(g_{h+1}|g_h, a) = \mathbb{P}(b_{h+1}|b_h, a) = \mathbb{P}(s_{h+1,i}|s_{h,i}, a) = 1, \ \forall i \in [S-2]. \tag{B.20}$$

- The initial distribution is decided by:

$$\mathbb{P}(s_{1,i}) = \frac{1}{S}, \ \forall i \in [S-2], \ \mathbb{P}(g_1) = \frac{1}{S}, \ \ \mathbb{P}(b_1) = \frac{1}{S} \tag{B.21}$$

- State s will receives reward 1 if and only if $s = g_h$ and $h \geq H$. The reward at all other states is zero.

By this construction the optimal policy must take $a^\star_{h,i}$ for each bandit state $s_{h,i}$ for at least the first half of the MDP, *i.e.* need to take $a^\star_{h,i}$ for $h \leq H$. In other words, this construction embeds at least $H(S-2)$ independent best arm identification problems that are identical to the stochastic multi-arm bandit problem in Lemma D.0.7 into the MDP. **Note the key innovation here is that we can remove the waiting states used in jiang2017contextual but still keep the multi-arm bandit problem independent!**[9]

Notice in our construction, for any bandit state $s_{h,i}$ with $h \leq H$, the difference of the expected reward between optimal action $a^\star_{h,i}$ and other actions is:

$$\begin{aligned}
&(\frac{1}{2}+\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):2H}|g_{h+1}] + (\frac{1}{2}-\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):2H}|b_{h+1}] + (1-\frac{1}{H})\cdot\mathbb{E}[r_{(h+1):2H}|s_{h+1,i}] \\
&\quad - \frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):2H}|g_{h+1}] - \frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):2H}|b_{h+1}] - (1-\frac{1}{H})\cdot\mathbb{E}[r_{(h+1):2H}|s_{h+1,i}] \\
=&(\frac{1}{2}+\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):2H}|g_{h+1}] + (\frac{1}{2}-\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):2H}|b_{h+1}] \\
&\quad - \frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):2H}|g_{h+1}] - \frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):2H}|b_{h+1}] \\
=&(\frac{1}{2}+\tau)\frac{1}{H}\cdot H + (\frac{1}{2}-\tau)\frac{1}{H}\cdot 0 - \frac{1}{2H}\cdot H + \frac{1}{2H}\cdot 0 = \tau
\end{aligned} \tag{B.22}$$

so it seems by Lemma D.0.7 one suffices to use the least possible $\frac{A}{72(\tau)^2}$ samples to identify the best action $a^\star_{h,i}$. However, note the construction of the latter half of the MDP (B.20) uses mindless reproduction of previous steps and therefore provides no additional information about the best action once the state at time H is known. In other words, observing $\sum_{t=1}^{2H} r_t = H$ is

[9]Here independence means solving one bandit problem provides no information on other bandit problems.

equivalent as observing $\sum_{t=1}^{H} r_t = 1$. Therefore, for the bandit states in the first half the samples that provide information for identifying the best arm is up to time H. As a result, the difference of the expected reward between optimal action $a^\star_{h,i}$ and other action for identifying the best arm should be corrected as:

$$
\begin{aligned}
&(\frac{1}{2}+\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):H}|g_{h+1}]+(\frac{1}{2}-\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):H}|b_{h+1}]+(1-\frac{1}{H})\cdot\mathbb{E}[r_{(h+1):H}|s_{h+1,i}]\\
&-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):H}|g_{h+1}]-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):H}|b_{h+1}]-(1-\frac{1}{H})\cdot\mathbb{E}[r_{(h+1):H}|s_{h+1,i}]\\
=&(\frac{1}{2}+\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):H}|g_{h+1}]+(\frac{1}{2}-\tau)\cdot\frac{1}{H}\cdot\mathbb{E}[r_{(h+1):H}|b_{h+1}]\\
&-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):H}|g_{h+1}]-\frac{1}{2H}\cdot\mathbb{E}[r_{(h+1):H}|b_{h+1}]\\
=&(\frac{1}{2}+\tau)\frac{1}{H}\cdot 1+(\frac{1}{2}-\tau)\frac{1}{H}\cdot 0-\frac{1}{2H}\cdot 1+\frac{1}{2H}\cdot 0=\frac{\tau}{H}
\end{aligned}
$$

Now by Lemma D.0.7, for each bandit state $s_{h,i}$ satisfying $h \le H$, unless $\frac{A}{72(\tau/H)^2}$ samples are collected from that state, the learning algorithm fails to identify the optimal action $a^\star_{h,i}$ with probability at least $1/3$.

After running any algorithm, let C be the set of (h, s) pairs for which the algorithm identifies the correct action. Let D be the set of (h, s) pairs for which the algorithm collects fewer than $\frac{A}{72(\tau/H)^2}$ samples. Then by Lemma D.0.7 we have

$$
\begin{aligned}
\mathbb{E}[|C|] &= \mathbb{E}\left[\sum_{(h,s)}\mathbf{1}[a_{h,s}=a^\star_{h,s}]\right] \le ((S-2)H-|D|)+\mathbb{E}\left[\sum_{(h,s)\in D}\mathbf{1}[a_{h,s}=a^\star_{h,s}]\right]\\
&\le ((S-2)H-|D|)+\frac{2}{3}|D| = (S-2)H-\frac{1}{3}|D|.
\end{aligned}
$$

If we have $n \le \frac{H(S-2)}{2}\times\frac{A}{72(\tau/H)^2}$, by pigeonhole principle the algorithm can collect $\frac{A}{72(\tau/H)^2}$ samples for at most half of the bandit problems, *i.e.* $|D| \ge H(S-2)/2$. Therefore we have

$$
\mathbb{E}[|C|] \le (S-2)H-\frac{1}{3}|D| \le \frac{5}{6}(S-2)H.
$$

Then by Markov inequality

$$
\mathbb{P}\left[|C| \ge \frac{11}{12}H(S-2)\right] \le \frac{5/6}{11/12} = \frac{10}{11}
$$

so the algorithm failed to identify the optimal action on 1/12 fraction of the bandit problems with probability at least $1/11$. Note for each failure in identification, the reward is differ by τ (see (B.22)), therefore under the event $\{|C'| \ge \frac{1}{12}H(S-2)\}$, following the similar calculation of [65] the suboptimality of the policy produced by the algorithm is

$$\begin{aligned}\epsilon := v^\star - v^{\hat{\pi}} &= \mathbb{P}[\text{visit } C'] \times \tau + \mathbb{P}[\text{visit } C] \times 0 = \mathbb{P}[\bigcup_{(h,i)\in C'} \text{visit}(h,i)] \times \tau \\ &= \sum_{(h,i)\in C'} \mathbb{P}[\text{visit}(h,i)] \times \tau = \sum_{(h,i)\in C'} \frac{1}{HS}(1-1/H)^{h-1}\tau \\ &\geq \sum_{(h,i)\in C'} \frac{1}{HS}(1-1/H)^{H}\tau \geq \sum_{(h,i)\in C'} \frac{1}{HS}\frac{1}{4}\tau \\ &\geq \frac{H(S-2)}{12}\frac{1}{HS}\frac{1}{4}\tau = c_1\frac{\tau}{48}.\end{aligned}$$

where the third equal sign uses all best arm identification problems are independent. Now we set $\tau = \min(\sqrt{1/8}, 48\epsilon/c_1)$ and under condition $n \leq cH^3SA/\epsilon^2$, we have

$$n \leq cH^3SA/\epsilon^2 \leq c48^2H^3SA/\tau^2 = c48^2 \cdot 72HS\cdot\frac{A}{72(\tau/H)^2} := c'HS\cdot\frac{A}{72\tau^2} \leq \frac{H(S-2)}{2}\cdot\frac{A}{72\tau^2},$$

the last inequality holds as long as $S \geq 2/(1-2c')$. Therefore in this situation, with probability at least $1/11$, $v^\star - v^{\hat{\pi}} \geq \epsilon$. Finally, we can use scaling to reduce the horizon from $2H$ to H.

B.6.2 Information theoretical lower sample complexity bound over problems in $\mathcal{M}_{d_m}$ for identifying ϵ-optimal policy.

For all $0 < d_m \leq \frac{1}{SA}$, let the class of problems be

$$\mathcal{M}_{d_m} := \left\{(\mu, M) \;\middle|\; \min_{t,s_t,a_t} d_t^\mu(s_t,a_t) \geq d_m\right\},$$

now we consider deriving minimax lower bound over this class.

Theorem B.6.2. *Under the same condition of Theorem B.6.1. In addition assume* $0 < d_m \leq \frac{1}{SA}$. *There exists another universal constant c such that when* $n \leq cH^3/d_m\epsilon^2$, *we always have*

$$\inf_{v^{\pi_{alg}}} \sup_{(\mu,M)\in\mathcal{M}_{d_m}} \mathbb{P}_{\mu,M}\left(v^* - v^{\pi_{alg}} \geq \epsilon\right) \geq p.$$

Proof:
The hard instance (μ, M) we used is based on Theorem B.6.1, which is described as follows.

- for the MDP $M = (\mathcal{S}, \mathcal{A}, r, P, d_1, 2H+2)$,
  - Initial distribution d_1 will always enter state s_0, and there are two actions with action a_1 always transitions to s_{yes} and action a_2 always transitions to s_{no}. The reward at the first time $r_1(s,a) = 0$ for any s, a.

– For state s_{no}, it will always transition back to itself regardless of the action and receive reward 0, *i.e.*

$$P_t(s_{\text{no}}|s_{\text{no}}, a) = 1,\ r_t(s_{\text{no}}, a) = 0,\ \forall t,\ \forall a.$$

– For state s_{yes}, it will transition to the MDP construction in Theorem B.6.1 with horizon $2H$ and s_{yes} always receives reward zero.

– For $t = 1$, choose $\mu(a_1|s_0) = \frac{1}{2}d_m SA$ and $\mu(a_2|s_0) = 1 - \frac{1}{2}d_m SA$. For $t \geq 2$, choose μ to be uniform policy, *i.e.* $\mu(a_t|s_t) = 1/A$.

Based on this construction, the optimal policy has the form $\pi^\star = (a_1, \ldots)$ and therefore the MDP branch that enters s_{no} is uninformative. Hence, data collected by that part is uninformed about the optimal policy and there is only $\frac{1}{2}d_m SA$ proportion of data from s_{yes} are useful. Moreover, by Theorem B.6.1 the rest of Markov chain succeeded from s_{yes} requires $\Omega(H^3SA/\epsilon^2)$ episodes (regardless of the exploration strategy/logging policy), so the actual data complexity needed for the whole construction (μ, M) is $\frac{\Omega(H^3SA/\epsilon^2)}{d_m SA} = \Omega(H^3/d_m\epsilon^2)$.

It remains to check this construction μ, M stays within $\mathcal{M}_{d_m}$.

- For $t = 1$, we have $d_1(s_0, a_1) = \frac{1}{2}d_m SA \geq d_m$ (since $S \geq 2$) and $d_1(s_0, a_2) = 1 - \frac{1}{2}d_m SA \geq d_m$ (this is since $d_m \leq \frac{1}{SA} \leq \frac{2}{2+SA}$);
- For $t = 2$, $d_2(s_{\text{yes}}, a) = \frac{1}{2}d_m SA \cdot \frac{1}{A} = \frac{1}{2}d_m S \geq d_m$ (since $S \geq 2$) and similar for s_{no};
- For $t \geq 3$, for g_h and b_h in the sub-chain inherited from s_{yes}, note $d_h(g_h) \leq d_{h+1}(g_{h+1})$ (since g_h and b_h are absorbing states regardless of actions), therefore $d_h(g_h) \geq d_1(g_1) = d_1(s_{\text{yes}}) \cdot \mathbb{P}(g_1|s_{\text{yes}}) = \frac{1}{2}d_m SA \cdot \frac{1}{S} = \frac{1}{2}d_m A$, since μ is uniform so $d_h(g_h, a) \geq \Omega(d_m A) \cdot \frac{1}{A} = \Omega(d_m)$ forall a. Similar result can be derived for b_h in identical way.

  For bandit state, we have for all $i \in [S-2]$,

$$\begin{aligned} d^\mu_{t+1}(s_{t+1,i}) &\geq \mathbb{P}^\mu(s_{t+1,i}, s_{t,i}, s_{t-1,i}, \ldots, s_{2,i}, s_{1,i}, s_{\text{yes}}, s_0) \\ &= \prod_{u=1}^{t} \mathbb{P}^\mu(s_{u+1,i}|s_u)\mathbb{P}^\mu(s_{1,i}|s_{\text{yes}})\mathbb{P}^\mu(s_{\text{yes}}|s_0) \\ &= (1 - \frac{1}{H})^t \left(\frac{1}{S}\right)\left(\frac{1}{2}d_m SA\right) \geq c d_m A, \end{aligned}$$

  now by μ is uniform we have $d^\mu_{t+1}(s_{t+1,i}, a) \geq \Omega(d_m A) \cdot \frac{1}{A} = \Omega(d_m)$ for all a. This concludes the proof.

Remark 14. *A directly corollary is that the sample complexity in Theorem 3.6.1 part 3. is optimal. Indeed, for the case $\epsilon_{opt} = 0$, Theorem 3.6.1 implies $\widehat{\pi}$ is the ϵ-optimal policy learned with sample complexity $O(H^3 \log(HSA/\delta)/d_m\epsilon^2)$. Theorem B.6.2 implies this sample complexity cannot be further reduced up to the logarithmic factor.*

B.6.3 Information theoretical lower sample complexity bound for uniform convergence in OPE.

By applying Theorem B.6.2, we can now prove Theorem 3.5.4.

Proof: [Proof of Theorem 3.5.4] We prove it by contradiction. Suppose there is one off-policy evaluation method $\widehat{v}^{\pi}$ such that

$$\sup_{\pi\in\Pi} |\widehat{v}^{\pi} - v^{\pi}| \leq o\left(\sqrt{\frac{H^3}{d_m n}}\right),$$

where $o(\cdot)$ represents the standard small o-notation. Then by

$$\begin{aligned} 0 \leq v^{\pi^\star} - v^{\widehat{\pi}^\star} &= v^{\pi^\star} - \widehat{v}^{\widehat{\pi}^\star} + \widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star} \\ &\leq |v^{\pi^\star} - \widehat{v}^{\pi^\star}| + |\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}| \leq 2\sup_{\pi} |v^{\pi} - \widehat{v}^{\pi}|. \end{aligned}$$

this OPE method implies a ϵ-optimal policy learning algorithm with sample complexity $o(H^3/d_m\epsilon^2)$ which is smaller than the information theoretical lower bound obtained in Theorem B.6.2. Contradiction!

B.7 Proofs of Theorem 3.6.1

Proof: [Proof of Theorem 3.6.1] Part 1. and Part 2. are just direct corollaries. We only prove Part 3. here. Indeed, by definition of empirical optimal policy we have $\widehat{Q}^{\pi^\star} \leq \widehat{Q}^{\widehat{\pi}^\star}$, so we have the following:

$$\begin{aligned} Q_1^{\pi^\star} - Q_1^{\widehat{\pi}} &= Q_1^{\pi^\star} - \widehat{Q}_1^{\widehat{\pi}^\star} + \widehat{Q}_1^{\widehat{\pi}^\star} - \widehat{Q}_1^{\widehat{\pi}} + \widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}} \\ &\leq Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star} + \widehat{Q}_1^{\widehat{\pi}^\star} - \widehat{Q}_1^{\widehat{\pi}} + \widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}} \\ &\leq Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star} + \epsilon_{\text{opt}} \cdot \mathbf{1} + \widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}} \end{aligned}$$

and $\widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}}$ can be bounded by Theorem 3.5.3 using local uniform convergence. $Q_1^{\pi^\star} - \widehat{Q}_1^{\pi^\star}$ can be bounded by $O(\sqrt{\frac{H^3\log(HSA/\delta)}{nd_m}})$ using the similar technique in Section B.5 even without introducing ϵ_{opt} since $\pi^\star$ is a fixed policy. All these implies:

$$Q_1^{\pi^\star} - Q_1^{\widehat{\pi}} \leq \left(O(\sqrt{\frac{H^3\log(HSA/\delta)}{nd_m}}) + \epsilon_{\text{opt}}\right) \cdot \mathbf{1}.$$

Especially when $\epsilon_{\text{opt}} = 0$ then this is slightly stronger than the stated result since:

$$v_1^{\pi^\star} - v_1^{\widehat{\pi}^\star} = Q_1^{\pi^\star}(\cdot, \pi^\star(\cdot)) - Q_1^{\widehat{\pi}^\star}(\cdot, \widehat{\pi}^\star(\cdot)) \leq Q_1^{\pi^\star}(\cdot, \pi^\star(\cdot)) - Q_1^{\widehat{\pi}^\star}(\cdot, \pi^\star(\cdot))$$
$$\leq \left\| Q_1^{\pi^\star} - Q_1^{\widehat{\pi}^\star} \right\|_\infty \leq O(\sqrt{\frac{H^3 \log(HSA/\delta)}{nd_m}}) \cdot \mathbf{1}$$

B.8 Simulation details

The non-stationary MDP with used for the experiments have 2 states s_0, s_1 and 2 actions a_1, a_2 where action a_1 has probability 1 always going back the current state and for action a_2, there is one state s.t. after choosing a_2 the dynamic transitions to both states with equal probability $\frac{1}{2}$ and the other one has asymmetric probability assignment ($\frac{1}{4}$ and $\frac{3}{4}$). The transition after choosing a_2 is changing over different time steps therefore the MDP is non-stationary and the change is decided by a sequence of pseudo-random numbers. More formally, P_t can be either

$$\mathbb{P}(s_0|s_0, a_1) = 1; \mathbb{P}(s_1|s_1, a_1) = 1; \mathbb{P}(\cdot|s_0, a_2) = \begin{cases} \frac{1}{2}, \text{ if } \cdot = s_1 \\ \frac{1}{2}, \text{ if } \cdot = s_0 \end{cases} ; \ \mathbb{P}(\cdot|s_1, a_2) = \begin{cases} \frac{3}{4}, \text{ if } \cdot = s_1 \\ \frac{1}{4}, \text{ if } \cdot = s_0 \end{cases}$$

or

$$\mathbb{P}(s_0|s_0, a_1) = 1; \mathbb{P}(s_1|s_1, a_1) = 1; \mathbb{P}(\cdot|s_0, a_2) = \begin{cases} \frac{1}{4}, \text{ if } \cdot = s_1 \\ \frac{3}{4}, \text{ if } \cdot = s_0 \end{cases} ; \ \mathbb{P}(\cdot|s_1, a_2) = \begin{cases} \frac{1}{2}, \text{ if } \cdot = s_1 \\ \frac{1}{2}, \text{ if } \cdot = s_0 \end{cases}$$

Moreover, to make the learning problem non-trivial we use non-stationary rewards with 4 categories, *i.e.* $r_t(s, a) \in \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\}$ and assignment of $r_t(s, a)$ for each value is changing over time. That means, one possible assignment can be

$$r_t(s_0, a_1) = 1/4, \ r_t(s_0, a_2) = 2/4, \ r_t(s_1, a_1) = 3/4, \ r_t(s_1, a_2) = 1/4.$$

Moreover, the logging policy in Figure 2.2 is uniform with $\mu_t(a_1|s) = \mu_t(a_2|s) = \frac{1}{2}$ for both states. We implement the non-stationary MDP in the `Python` environment and pseudo-random numbers p_t, r_t's are generated by keeping `numpy.random.seed(100)`.

We fix episodes $n = 2048$ and run each algorithm under $K = 100$ macro-replications with data $\mathcal{D}_{(k)} = \left\{ (s_t^{(i)}, a_t^{(i)}, r_t^{(i)}) \right\}_{(k)}^{i \in [n], t \in [H]}$, and use each $\mathcal{D}_{(k)}$ $(k = 1, ..., K)$ to construct a estimator $\widehat{v}_{[k]}^{\pi}$, then the (empirical) RMSE for fixed policy is computed as:

$$\text{RMSE\_FIX} = \sqrt{\frac{\sum_{k=1}^{K} (\widehat{v}_{[k]}^{\pi} - v_{\text{true}}^{\pi})^2}{K}},$$

and RMSE for suboptimality gap is computed as

$$\text{RMSE\_SUB} = \sqrt{\frac{\sum_{k=1}^{K}(v^{\widehat{\pi}^\star_{[k]}} - v^{\pi^\star}_{\text{true}})^2}{K}},$$

and RMSE for empirical optimal policy gap is computed as

$$\text{RMSE\_EMPIRICAL} = \sqrt{\frac{\sum_{k=1}^{K}(\widehat{v}^{\widehat{\pi}^\star}_{[k]} - v^{\widehat{\pi}^\star}_{\text{true}})^2}{K}},$$

where v^{π}_{true} is obtained by calculating $P^{\pi}_{t+1,t}(s'|s) = \sum_a P_{t+1,t}(s'|s,a)\pi_t(a|s)$, the marginal state distribution $d^{\pi}_t = P^{\pi}_{t,t-1} d^{\pi}_{t-1}$, $r^{\pi}_t(s_t) = \sum_{a_t} r_t(s_t,a_t)\pi_t(a_t|s_t)$ and $v^{\pi}_{\text{true}} = \sum_{t=1}^{H}\sum_{s_t} d^{\pi}_t(s_t) r^{\pi}_t(s_t)$. $v^{\pi^\star}_{\text{true}}$ is obtained by running Value Iteration exhaustively until the error converges to 0. The average relative error for suboptimality (average of $|v^{\widehat{\pi}^\star_{[k]}} - v^{\pi^\star}_{\text{true}}|/v^{\pi^\star}_{\text{true}}$) at $H = 1000$ is 0.0011. Lastly, we also show the scaling of $|\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}|$ in Figure B.2, which shares a similar pattern as the suboptimality plot as a whole. [10]

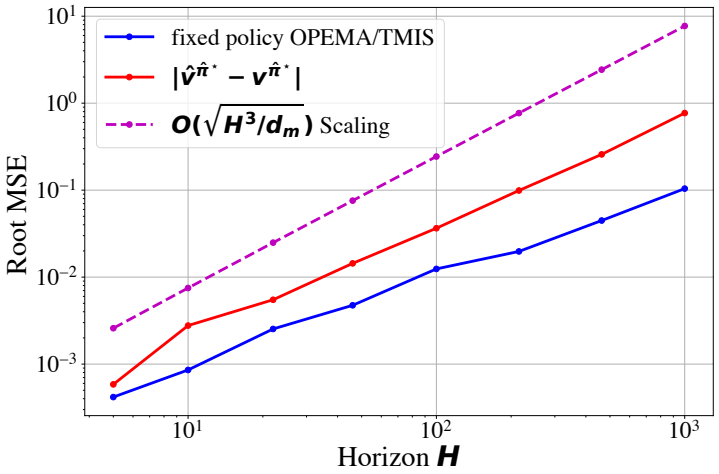

Figure B.2: Log-log plot showing the dependence on horizon of uniform OPE and pointwise OPE via learning ($|\hat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}|$) over a non-stationary MDP example.

B.9 On improvement over vanilla simulation lemma for fixed policy evaluation

Vanilla simulation lemma, Lemma 1 of [41]. Without loss of generality, assuming reward is determinsitic function over state-action. By definition of Bellman equation, we have the

[10] Here we do point out the empirical dependence on H for $|\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}|$ in the Figure B.2 is actually less than $H^{1.5}$, this comes from that the MDP example we choose is not the "hardest" example for quantity $|\widehat{v}^{\widehat{\pi}^\star} - v^{\widehat{\pi}^\star}|$.

following:

$$\widehat{V}_t^\pi = r + \widehat{P}_{t+1}^\pi \widehat{V}_{t+1}^\pi, \quad V_t^\pi = r + P_{t+1}^\pi V_{t+1}^\pi,$$

define $\epsilon_P = \sup_{t,s_t,a_t} ||\widehat{P}_t(\cdot|s_t,a_t) - P_t(\cdot|s_t,a_t)||_1$, then by Hoeffding's inequality and union bound, with probability $1-\delta$,

$$\epsilon_P \leq S \cdot \sup_{t,s_t,a_t} ||\widehat{P}_t(\cdot|s_t,a_t) - P_t(\cdot|s_t,a_t)||_\infty \leq S \cdot \sup_{t,s_t,a_t} O\left(\sqrt{\frac{\log(HSA/\delta)}{n_{s_t,a_t}}}\mathbf{1}(E_t)\right) = O\left(\sqrt{\frac{S^2\log(HSA/\delta)}{n \cdot d_m}}\right)$$

then

$$\begin{aligned}
\widehat{V}_t^\pi - V_t^\pi =& \widehat{P}_{t+1}^\pi \widehat{V}_{t+1}^\pi - P_{t+1}^\pi V_{t+1}^\pi \\
\leq& \left(\left\|\widehat{P}_{t+1}^\pi - P_{t+1}^\pi\right\|_1 \left\|\widehat{V}_{t+1}^\pi\right\|_\infty + \left\|P_{t+1}^\pi\right\|_1 \left\|\widehat{V}_{t+1}^\pi - V_{t+1}^\pi\right\|_\infty\right) \cdot \mathbf{1} \\
\leq& \left(H\epsilon_P + \left\|\widehat{V}_{t+1}^\pi - V_{t+1}^\pi\right\|_\infty\right) \cdot \mathbf{1},
\end{aligned}$$

solving recursively, we have

$$\left\|\widehat{V}_1^\pi - V_1^\pi\right\|_\infty \leq H^2\epsilon_P \leq O\left(\sqrt{\frac{H^4S^2\log(HSA/\delta)}{n \cdot d_m}}\right).$$

This verifies SL has complexity $\widetilde{O}(H^4S^2/d_m\epsilon^2)$. We do point out above standard analysis can be improved (*e.g.* [41] Section 2.2) to $\tilde{O}(H^4S/d_m\epsilon^2)$, then in this case our analysis (Lemma 3.5.2) has an improvement of H^2S with respect to the modified result.

B.10 Algorithms

Remark 15. *In short, we can see Algorithm 2 requires the splitting data size M which is undecided by [57] and that makes the hyper-parameter requiring additional concrete specifications to make the data splitting estimator sample efficient. In contrast, OPEMA in Algorithm 3 is defined without ambiguity and can be implemented without extra work.*

Their results require number of episodes in each splitted data M to satisfy $\widetilde{O}(\sqrt{nSA}) > M > O(HSA)$. To achieve data efficiency, they need $n \approx \Theta(H^2SA/\epsilon^2)$ and by that condition M has to satisfy $M \approx C \cdot HSA$. In this case, data-splitting version needs to create $N = n/M$ empirical transition dynamics and each dynamics use $H^3/N \approx C \cdot H^2SA/\epsilon^2$ episodes which is less than the lower bound ($O(H^3)$) required for learning. Most critically, due to data-splitting it has N empirical transitions hence it is not clear which transition to plan over. Therefore in this sense their result does not enables efficient offline learning. Our Analysis for unsplitted

Algorithm 3 OPEMA

Input: Logging data $\mathcal{D} = \{\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{t=1}^H\}_{i=1}^n$ from the behavior policy μ. A target policy π which we want to evaluate its cumulative reward.

1: Calculate the on-policy estimation of initial distribution $d_1(\cdot)$ by $\widehat{d}_1(s) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(s_1^{(i)} = s)$, and set $\widehat{d}_1^{\mu}(\cdot) := \widehat{d}_1(\cdot)$, $\widehat{d}_1^{\pi}(s) := \widehat{d}_1(\cdot)$.

2: **for** $t = 2, 3, \ldots, H$ **do**

3: Choose all transition data at time step t, $\{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{i=1}^n$.

4: Calculate the on-policy estimation of $d_t^{\mu}(\cdot)$ by $\widehat{d}_t^{\mu}(s) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}(s_t^{(i)} = s)$.

5: Set the off-policy estimation of $\widehat{P}_t(s_t|s_{t-1}, a_{t-1})$:

$$\widehat{P}_t(s_t|s_{t-1}, a_{t-1}) := \frac{\sum_{i=1}^n \mathbf{1}[(s_t^{(i)}, a_{t-1}^{(i)}, s_{t-1}^{(i)}) = (s_t, s_{t-1}, a_{t-1})]}{n_{s_{t-1}, a_{t-1}}}$$

when $n_{s_{t-1}, a_{t-1}} > 0$. Otherwise set it to be zero.

6: Estimate the reward function

$$\widehat{r}_t(s_t, a_t) := \frac{\sum_{i=1}^n r_t^{(i)} \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}{\sum_{i=1}^n \mathbf{1}(s_t^{(i)} = s_t, a_t^{(i)} = a_t)}.$$

when $n_{s_t, a_t} > 0$. Otherwise set it to be zero.

7: Set $\widehat{d}_t^{\pi}(\cdot, \cdot)$ according to $\widehat{d}_t^{\pi} = \widehat{P}_t^{\pi} \widehat{d}_{t-1}^{\pi}$, where $\widehat{d}_t^{\pi}(\cdot, \cdot)$ is the estimated state-action distribution.

8: **end for**

9: Substitute the all estimated values above into $\widehat{v}^{\pi} = \sum_{t=1}^H \langle \widehat{d}_t^{\pi}, \widehat{r}_t \rangle$ to obtain $\widehat{v}^{\pi}$, the estimated value of π.

version (OPEMA) addresses all these issues.

Appendix C

Supplementary Material to Chapter 4

C.1 Proof of optimal local uniform convergence

C.1.1 Model-based Offline Plug-in Estimator

Recall the model-based estimator uses empirical estimator $\widehat{P}$ for estimating P and the estimator is calculated accordingly:

$$\widehat{Q}_h^\pi = r + \widehat{P}^{\pi_{h+1}} Q_{h+1}^\pi = r + \widehat{P} V_{H+1}^\pi,$$

where $\widehat{P}(s'|s,a)$ can be expressed as:

$$\widehat{P}(s'|s,a) = \frac{\sum_{i=1}^n \sum_{h=1}^H \mathbf{1}[(s_{h+1}^{(i)}, a_h^{(i)}, s_h^{(i)}) = (s', s, a)]}{n_{s,a}}, \quad n_{s,a} = \sum_{h=1}^H \sum_{i=1}^n \mathbf{1}[(s_h^{(i)}, a_h^{(i)}) = (s,a)].$$

and $\widehat{P}(s'|s,a) = \frac{1}{S}$, if $n_{s,a} = 0$. The initial distribution is also constructed as $\widehat{d}_1^\pi(s) = n_s/n$.
First of all, we have by definition the Bellman optimality equation

$$V_t^\star(s) = \max_a \left\{ r(s,a) + \sum_{s'} P(s'|s,a) V_{t+1}^\star(s') \right\}, \quad \forall s \in \mathcal{S}. \tag{C.1}$$

and similarly the empirical version

$$\widehat{V}_t^\star(s) = \max_a \left\{ r(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \widehat{V}_{t+1}^\star(s') \right\}, \quad \forall s \in \mathcal{S}.$$

The key difficulty in obtaining the optimal dependence in stationary setting is decoupling the dependence of $P - \widehat{P}$ and $\widehat{V}^\star$. This issue is not encountered in the non-stationary setting due to the possibility to estimate different transition at each time yin2021near, but it cannot further

reduce the sample complexity on H. Moreover, the direct use of s-absorbing MDP in [61] is not sharp for finite horizon stationary setting, as it requires s-absorbing MDPs with H-dimensional cover (which has size $\approx e^H$ and it is not optimal). We design the *singleton-absorbing MDP* to get rid of the issue.

C.1.2 General absorbing MDP

The general absorbing MDP is defined as follows: for a fixed state s and a sequence $\{u_t\}_{t=1}^H$, MDP $M_{s,\{u_t\}_{t=1}^H}$ is identical to M for all states except s, and state s is absorbing in the sense $P_{M_{s,\{u_t\}_{t=1}^H}}(s|s,a) = 1$ for all a, and the instantaneous reward at time t is $r_t(s,a) = u_t$ for all $a \in \mathcal{A}$. Also, we use the shorthand notation $V^\pi_{\{s,u_t\}}$ for $V^\pi_{s,M_{s,\{u_t\}_{t=1}^H}}$ and similarly for $Q_{\{s,u_t\}}$ and transition $P_{\{s,u_t\}}$. Then the following properties hold:

Lemma C.1.1.

$$V^\star_{h,\{s,u_t\}}(s) = \sum_{t=h}^{H} u_t.$$

Proof: We prove this by backward induction. For $h = H$, under $M_{s,\{u_t\}_{t=1}^H}$ state s is absorbing (and by convention $V^\star_{H+1,\{s,u_t\}} = 0$) therefore

$$V^\star_{H,\{s,u_t\}}(s) = \max_a \left\{ r_{H,\{s,u_t\}}(s,a) + \sum_{s'} P_{\{s,u_t\}}(s'|s,a) V^\star_{H+1,\{s,u_t\}}(s') \right\} = \max_a \left\{ r_{H,\{s,u_t\}}(s,a) \right\} = u_H$$

for general h, note $\sum_{s'} P_{\{s,u_t\}}(s'|s,a) V^\star_{h+1,\{s,u_t\}}(s') = 1 \cdot V^\star_{h+1,\{s,u_t\}}(s)$, therefore using induction property $V^\star_{h+1,\{s,u_t\}}(s) = \sum_{t=h+1}^H u_t$ we can similarly obtain $V^\star_{h,\{s,u_t\}}(s) = \sum_{t=h}^H u_t$.

Lemma C.1.2. *Fix state s. For two different sequences $\{u_t\}_{t=1}^H$ and $\{u'_t\}_{t=1}^H$, we have*

$$\max_h \left\| Q^\star_{h,\{s,u_t\}} - Q^\star_{h,\{s,u'_t\}} \right\|_\infty \le H \cdot \max_{t\in[H]} \left| u_t - u'_t \right|.$$

Proof: Let $\pi^\star_{\{s,u_t\}}$ be the optimal policy in $M_{\{s,u_t\}}$. Then (by convention $\prod_{a=h+1}^{h} P^{\pi_a} = I$)

$$
\begin{aligned}
&Q^\star_{h,\{s,u_t\}} - Q^\star_{h,\{s,u'_t\}} = Q^\star_{h,\{s,u_t\}} - \max_{\pi} \sum_{i=h}^{H} \left(\prod_{a=h+1}^{i} P^{\pi_a}_{\{s,u'_t\}} \right) r_{i,\{s,u'_t\}} \\
\leq & Q^\star_{h,\{s,u_t\}} - \sum_{i=h}^{H} \left(\prod_{a=h+1}^{i} P^{\pi^\star_{a,\{s,u_t\}}}_{\{s,u'_t\}} \right) r_{i,\{s,u'_t\}} = \sum_{i=h}^{H} \left(\prod_{a=h+1}^{i} P^{\pi^\star_{a,\{s,u_t\}}}_{\{s,u'_t\}} \right) \left(r_{i,\{s,u_t\}} - r_{i,\{s,u'_t\}} \right) \\
&\leq \sum_{i=h}^{H} \max_{s,a} \left\| \left(\prod_{a=h+1}^{i} P^{\pi^\star_{a,\{s,u_t\}}}_{\{s,u'_t\}} \right)^{i-h} (\cdot|s,a) \right\|_1 \cdot \left\| r_{i,\{s,u_t\}} - r_{i,\{s,u'_t\}} \right\|_\infty \cdot \mathbf{1} = (H-h+1) \cdot \max_t \left| u_t - u'_t \right| \cdot \mathbf{1}
\end{aligned}
$$

where the first equal sign uses the definition of $Q^\star$, the second equal sign uses $P_{\{s,u_t\}}$ only depends s but not the specification of u_t's and the last equal sign comes from $r_{i,\{s,u_t\}}(s,a) = u_i$ for any $a \in \mathcal{A}$ and $r_{i,\{s,u_t\}}(\tilde{s},a) = r_{i,\{s,u'_t\}}(\tilde{s},a)$ for any $\tilde{s} \neq s$. Lastly by symmetry we finish the proof.

C.1.3 Singleton-absorbing MDP

The direct transfer of absorbing technique created in [61] will require each u_t to fill in the range of $[0,H]$ using evenly spaced elements. For finite horizon MDP there are H layers, therefore the total number of H-tuples $(u_1,\ldots,u_H)$ has order $|U_s| = Poly(H)^H$, therefore when apply the union bound, it will incur the additional H factor. We get rid of this issue by choosing one single point in H-dimensional space $[0,H]^H$. We first give the following two lemmas.

Lemma C.1.3. *$V^\star_t(s) - V^\star_{t+1}(s) \geq 0$, for all state $s \in \mathcal{S}$ and all $t \in [H]$.*

Proof: Let the optimal policy for $V^\star_{t+1}$ be $\pi^\star_{t+1:H}$, *i.e.* $V^\star_{t+1} = V^{\pi^\star_{t+1:H}}_{t+1}$, then artificially construct a policy $\pi_{t:H}$ such that $\pi_{t:H-1} = \pi^\star_{t+1:H}$ and π_H is arbitrary, then by the definition of

optimal value

$$\begin{aligned}
V_t^\star(s) \geq V_t^{\pi_{t:H}}(s) &= \mathbb{E}^{\pi_{t:H}}\left[\sum_{i=t}^{H} r(s_i,a_i)\middle| s_t = s\right] \\
&= \mathbb{E}^{\pi_{t:H-1}}\left[\sum_{i=t}^{H-1} r(s_i,a_i)\middle| s_t = s\right] + \mathbb{E}^{\pi_{t:H}}\left[r(s_H,a_H)\middle| s_t = s\right] \\
&= \mathbb{E}^{\pi^\star_{t+1:H}}\left[\sum_{i=t+1}^{H} r(s_i,a_i)\middle| s_{t+1} = s\right] + \mathbb{E}^{\pi_{t:H}}\left[r(s_H,a_H)\middle| s_t = s\right] \\
&\geq \mathbb{E}^{\pi^\star_{t+1:H}}\left[\sum_{i=t+1}^{H} r(s_i,a_i)\middle| s_{t+1} = s\right] + 0 = V_{t+1}^\star(s),
\end{aligned}$$

where the third equal sign uses exactly that *P is a STATIONARY transition* and definition $\pi_{t:H-1} = \pi^\star_{t+1:H}$. The last inequality uses the assumption that reward is always non-negative.

Remark 16. *Lemma C.1.3 leverages P is stationary and above may not be true in the non-stationary setting. This enables us to establish the following lemma, which is the key for singleton-absorbing MDP.*

Lemma C.1.4. *Fix a state s. If we choose $u_t^\star := V_t^\star(s) - V_{t+1}^\star(s)\ \forall t \in [H]$, then we have the following vector form equation*

$$V^\star_{h,\{s,u_t^\star\}} = V^\star_{h,M} \quad \forall h \in [H].$$

Similarly, if we choose $\widehat{u}_t^\star := \widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s)$, then $\widehat{V}^\star_{h,\{s,\widehat{u}_t^\star\}} = \widehat{V}^\star_{h,M}$, $\forall h \in [H]$.

Proof: We focus on the first claim. Note by Lemma C.1.3 the assignment of $u_t^\star (:= r_{t,\{s,u_t^\star\}})$ is well-defined. Next recall $V^\star_{h,M}$ is the optimal value under true MDP M and $V^\star_{h,\{s,u_t^\star\}}$ is the optimal value under the assimilating MDP $M_{s,\{u_t^\star\}_{t=1}^H}$. We prove by backward induction.

For $h = H$, note by convention $V^\star_{H+1} = 0$, therefore $u_H^\star = V_H^\star(s) - V_{H+1}^\star(s) = V_H^\star(s) - 0 = V_H^\star(s)$ and Bellman optimality equation becomes

$$V_H^\star(\tilde{s}) = \max_a \{r(\tilde{s},a)\}, \quad \forall \tilde{s} \in \mathcal{S}.$$

Under $M_{s,\{u_t^\star\}_{t=1}^H}$, for state s by Lemma C.1.1 we have $V^\star_{H,\{s,u_t^\star\}}(s) = u_H^\star = V_H^\star(s)$, for other states $\tilde{s} \neq s$, reward in $M_{s,\{u_t^\star\}_{t=1}^H} = M$ so we also have $V^\star_{H,\{s,u_t^\star\}}(\tilde{s}) = V_H^\star(\tilde{s})$ for all $\tilde{s} \neq s$.

Now for general h, for state s by Lemma C.1.1

$$V^\star_{h,\{s,u_t^\star\}}(s) = \sum_{t=h}^{H} u_t^\star = \sum_{t=h}^{H} \left(V_t^\star(s) - V_{t+1}^\star(s)\right) = V_h^\star(s),$$

for state $\tilde{s} \neq s$, by Bellman optimality equation

$$\begin{aligned} V^{\star}_{h,\{s,u_t^\star\}}(\tilde{s}) &= \max_a \left\{ r_{\{s,u_t^\star\}}(\tilde{s},a) + \sum_{s'} P_{\{s,u_t^\star\}}(s'|\tilde{s},a) V^{\star}_{h+1,\{s,u_t^\star\}}(s') \right\} \\ &= \max_a \left\{ r(\tilde{s},a) + \sum_{s'} P(s'|\tilde{s},a) V^{\star}_{h+1,\{s,u_t^\star\}}(s') \right\} \\ &= \max_a \left\{ r(\tilde{s},a) + \sum_{s'} P(s'|\tilde{s},a) V^{\star}_{h+1}(s') \right\} = V^{\star}_h(\tilde{s}), \end{aligned}$$

where the second equal sign uses when $\tilde{s} \neq s$, $M_{s,\{u_t^\star\}_{t=1}^H}$ is identical to M and the third equal sign uses induction assumption that element-wisely $V^{\star}_{h+1,\{s,u_t^\star\}} = V^{\star}_{h+1}$. Similar result can be derived for $\hat{u}^\star$ version and this completes the proof.

The singleton MDP we used is exactly $M_{s,\{u_t^\star\}_{t=1}^H}$ (or $\widehat{M}_{s,\{u_t^\star\}_{t=1}^H}$).

C.1.4 Proof for local uniform convergence

Recall the local policy class

$$\Pi_l := \left\{ \pi : \text{ s.t. } \left\| \widehat{V}_h^{\pi} - \widehat{V}_h^{\widehat{\pi}^\star} \right\|_\infty \leq \epsilon_{\text{opt}}, \forall h \in [H] \right\}.$$

For ease of exposition, we denote $N := \min_{s,a} n_{s,a}$. Note N itself is a random variable, therefore for the rest of proof we first conditional on N. Later we shall remove the conditional on N (see Section C.1.7).

For any $\widehat{\pi} \in \Pi_l$, by (empirical) Bellman equation we have element-wisely:

$$\begin{aligned} \widehat{Q}_h^{\widehat{\pi}} - Q_h^{\widehat{\pi}} &= r_h + \widehat{P}^{\widehat{\pi}_{h+1}} \widehat{Q}_{h+1}^{\widehat{\pi}} - r_h - P^{\widehat{\pi}_{h+1}} Q_{h+1}^{\widehat{\pi}} \\ &= \left(\widehat{P}^{\widehat{\pi}_{h+1}} - P^{\widehat{\pi}_{h+1}} \right) \widehat{Q}_{h+1}^{\widehat{\pi}} + P^{\widehat{\pi}_{h+1}} \left(\widehat{Q}_{h+1}^{\widehat{\pi}} - Q_{h+1}^{\widehat{\pi}} \right) \\ &= \left(\widehat{P} - P \right) \widehat{V}_{h+1}^{\widehat{\pi}} + P^{\widehat{\pi}_{h+1}} \left(\widehat{Q}_{h+1}^{\widehat{\pi}} - Q_{h+1}^{\widehat{\pi}} \right) \\ &= \ldots = \sum_{t=h}^{H} \Gamma_{h+1:t}^{\widehat{\pi}} \left(\widehat{P} - P \right) \widehat{V}_{t+1}^{\widehat{\pi}} \\ &\leq \underbrace{\sum_{t=h}^{H} \Gamma_{h+1:t}^{\widehat{\pi}} \left| \left(\widehat{P} - P \right) \widehat{V}_{t+1}^{\widehat{\pi}^\star} \right|}_{(\star)} + \underbrace{\sum_{t=h}^{H} \Gamma_{h+1:t}^{\widehat{\pi}} \left| \left(\widehat{P} - P \right) \left(\widehat{V}_{t+1}^{\widehat{\pi}} - \widehat{V}_{t+1}^{\widehat{\pi}^\star} \right) \right|}_{(\star\star)} \end{aligned}$$

where $\Gamma^{\pi}_{h+1:t} = \prod_{i=h+1}^{t} P^{\pi_i}$ is multi-step state-action transition and $\Gamma_{h+1:h} := I$.

C.1.5 Analyzing ($\star\star$)

Term ($\star\star$) can be readily bounded using the following lemma.

Lemma C.1.5. *Fix $N > 0$, we have with probability $1-\delta$, for all $t = 1, ..., H-1$*

$$\sum_{t=h}^{H} \Gamma^{\widehat{\pi}}_{h+1:t} \left|(\widehat{P} - P)(\widehat{V}^{\widehat{\pi}^\star}_{h+1} - \widehat{V}^{\widehat{\pi}}_{h+1})\right| \leq C\epsilon_{\text{opt}} \cdot \sqrt{\frac{H^2 S \log(SA/\delta)}{N}} \cdot \mathbf{1}$$

where C absorb the higher order term and absolute constants.

Proof:
First, by vector induced matrix norm[1] we have

$$\begin{aligned}
\left\| \sum_{t=h}^{H} \Gamma^{\widehat{\pi}}_{h+1:t} \cdot \left|(\widehat{P} - P)(\widehat{V}^{\widehat{\pi}^\star}_{t+1} - \widehat{V}^{\widehat{\pi}}_{t+1})\right| \right\|_\infty &\leq H \cdot \sup_t \left\| \Gamma^{\widehat{\pi}}_{h+1:t} \right\|_\infty \left\| |(\widehat{P} - P)(\widehat{V}^{\widehat{\pi}^\star}_{t+1} - \widehat{V}^{\widehat{\pi}}_{t+1})| \right\|_\infty \\
&\leq H \cdot \sup_t \left\| |(\widehat{P} - P)(\widehat{V}^{\widehat{\pi}^\star}_{t+1} - \widehat{V}^{\widehat{\pi}}_{t+1})| \right\|_\infty \\
&= H \cdot \sup_{t,s,a} \left|(\widehat{P} - P)(\cdot|s,a)(\widehat{V}^{\widehat{\pi}^\star}_{t+1} - \widehat{V}^{\widehat{\pi}}_{t+1})\right| \\
&\leq H \cdot \sup_{t,s,a} \left\|(\widehat{P} - P)(\cdot|s,a)\right\|_1 \cdot \left\|\widehat{V}^{\widehat{\pi}^\star}_{t+1} - \widehat{V}^{\widehat{\pi}}_{t+1}\right\|_\infty \cdot \mathbf{1}
\end{aligned}$$

where the second inequality uses multi-step transition $\Gamma^{\pi}_{t+1:h-1}$ is row-stochastic. Note given N, therefore by Lemma D.0.10 and a union bound we have with probability $1-\delta$,

$$\sup_{s,a} \left\|(\widehat{P} - P)(\cdot|s,a)\right\|_1 \leq C(\sqrt{\frac{S\log(SA/\delta)}{N}}),$$

(where C absorb the higher order term and absolute constants) and using definition of Π_l we have $\sup_t \left\|\widehat{V}^{\widehat{\pi}^\star}_t - \widehat{V}^{\widehat{\pi}}_t\right\|_\infty \leq \epsilon_{\text{opt}}$. This indicates

$$\sup_{t,s,a} \left\|(\widehat{P} - P)(\cdot|s,a)\right\|_1 \cdot \left\|\widehat{V}^{\widehat{\pi}^\star}_{t+1} - \widehat{V}^{\widehat{\pi}}_{t+1}\right\|_\infty \cdot \mathbf{1} \leq C(\epsilon_{\text{opt}} \sqrt{\frac{S\log(SA/\delta)}{N}} \cdot \mathbf{1}),$$

where $\mathbf{1} \in \mathbb{R}^S$ is all-one vector. Then multiple by H to get the stated result.

[1] For A a matrix and x a vector we have $\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$.

C.1.6 Analyzing ($\star$)

Concentration on $\left(\widehat{P}-P\right)\widehat{V}_h^\star$.[2] Since $\widehat{P}$ aggregates all data from different step so that $\widehat{P}$ and $\widehat{V}_h^\star$ are on longer independent, Bernstein inequality cannot be directly applied. We use the singleton-absorbing MDP $M_{s,\{u_t^\star\}_{t=1}^H}$ to handle the case (recall $u_t^\star := V_t^\star(s) - V_{t+1}^\star(s)\ \forall t \in [H]$). Again, let us fix a state s and $a \in \mathcal{A}$ be any action. Also, we use $P_{s,a}$ to denote row vector to avoid long expression. Then we have:

$$
\begin{aligned}
&\left(\widehat{P}_{s,a}-P_{s,a}\right)\widehat{V}_h^\star = \left(\widehat{P}_{s,a}-P_{s,a}\right)\left(\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}} + \widehat{V}^\star_{h,\{s,u_t^\star\}}\right)\\
&= \left(\widehat{P}_{s,a}-P_{s,a}\right)\left(\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right) + \left(\widehat{P}_{s,a}-P_{s,a}\right)\widehat{V}^\star_{h,\{s,u_t^\star\}}\\
&\le \left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1 \left\|\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty + \sqrt{\frac{2\log(4/\delta)}{N}}\sqrt{\mathrm{Var}_{s,a}(\widehat{V}^\star_{h,\{s,u_t^\star\}})} + \frac{2H\log(1/\delta)}{3N}\\
&\le \left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1 \left\|\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty + \sqrt{\frac{2\log(4/\delta)}{N}}\left(\sqrt{\mathrm{Var}_{s,a}(\widehat{V}_h^\star)} + \sqrt{\mathrm{Var}_{s,a}(\widehat{V}^\star_{h,\{s,u_t^\star\}} - \widehat{V}_h^\star)}\right) + \frac{2H\log(1/\delta)}{3N}\\
&\le \left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1 \left\|\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty + \sqrt{\frac{2\log(4/\delta)}{N}}\left(\sqrt{\mathrm{Var}_{s,a}(\widehat{V}_h^\star)} + \sqrt{\left\|\widehat{V}^\star_{h,\{s,u_t^\star\}} - \widehat{V}_h^\star\right\|_\infty^2}\right) + \frac{2H\log(1/\delta)}{3N}\\
&= \left(\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1 + \sqrt{\frac{2\log(4/\delta)}{N}}\right)\left\|\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty + \sqrt{\frac{2\log(4/\delta)}{N}}\sqrt{\mathrm{Var}_{s,a}(\widehat{V}_h^\star)} + \frac{2H\log(1/\delta)}{3N}
\end{aligned}
\tag{C.2}
$$

where the first inequality uses Bernstein inequality (Lemma D.0.3), while the second inequality uses $\sqrt{\mathrm{Var}(\cdot)}$ in norm (norm triangle inequality). Now we treat $\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1$ and $\left\|\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty$ separately.

For $\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1$. Indeed, by Lemma D.0.10 again $\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1 \le \tilde{O}(\sqrt{\frac{S\log(S/\delta)}{N}})$ and by a union bound we obtain w.p., $1-\delta$

$$
\sup_{s,a}\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1 \le C\sqrt{\frac{S\log(SA/\delta)}{N}}. \tag{C.3}
$$

where C absorbs the higher order term and constants.

For $\left\|\widehat{V}_h^\star - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty$. Note if we set $\widehat{u}_t^\star = \widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s)$, then by Lemma C.1.4

$$
\widehat{V}_h^\star = \widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}}
$$

Next since $\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}}(\tilde{s}) = \max_a \widehat{Q}^\star_{h,\{s,\hat{u}_t^\star\}}(\tilde{s},a)\ \forall \tilde{s} \in \mathcal{S}$, by generic inequality $|\max f - \max g| \le$

[2] Here we use $\widehat{V}_h^\star$ instead of $\widehat{V}_t^\star$ since we later have $\widehat{V}^\star_{h,\{s,u_t^\star\}}$. We avoid the same t twice in the expression to prevent confusion.

$\max|f-g|$, we have $|\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}}(\tilde{s}) - \widehat{V}^\star_{h,\{s,u_t^\star\}}(\tilde{s})| \le \max_a |\widehat{Q}^\star_{h,\{s,\hat{u}_t^\star\}}(\tilde{s},a) - \widehat{Q}^\star_{h,\{s,u_t^\star\}}(\tilde{s},a)|$, taking $\max_{\tilde{s}}$ on both sides, we obtain exactly

$$\left\|\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}} - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty \le \left\|\widehat{Q}^\star_{h,\{s,\hat{u}_t^\star\}} - \widehat{Q}^\star_{h,\{s,u_t^\star\}}\right\|_\infty$$

then by Lemma C.1.2,

$$\left\|\widehat{V}^\star_{h} - \widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty \le \left\|\widehat{Q}^\star_{h,\{s,\hat{u}_t^\star\}} - \widehat{Q}^\star_{h,\{s,u_t^\star\}}\right\|_\infty \le H \max_t \left|\hat{u}_t^\star - u_t^\star\right|, \tag{C.4}$$

Recall

$$\hat{u}_t^\star - u_t^\star = \widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s) - \left(V_t^\star(s) - V_{t+1}^\star(s)\right).$$

Now we denote

$$\Delta_s := \max_t |\hat{u}_t^\star - u_t^\star| = \max_t \left|\widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s) - \left(V_t^\star(s) - V_{t+1}^\star(s)\right)\right|,$$

then Δ_s itself is a scalar and a random variable.

To sum up, by (C.2), (C.3) and (C.4) and a union bound we have

Lemma C.1.6. *Fix $N > 0$. With probability $1-\delta$, element-wisely, for all $h \in [H]$,*

$$\left|\left(\widehat{P}-P\right)\widehat{V}_h^\star\right| \le C\sqrt{\frac{S\log(HSA/\delta)}{N}} \cdot H \max_s \Delta_s \cdot \mathbf{1} + \sqrt{\frac{2\log(4HSA/\delta)}{N}}\sqrt{\mathrm{Var}_P(\widehat{V}_h^\star)} + \frac{2H\log(HSA/\delta)}{3N}\cdot\mathbf{1}$$

Now plug Lemma C.1.6 back into ($\star$) and combine Lemma C.1.5, we receive:

$$\begin{aligned}
&\left|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right| \\
&\le \sum_{t=h}^H \Gamma^{\hat{\pi}}_{h+1:t}\left(C\sqrt{\frac{S\log(HSA/\delta)}{N}}\cdot H\max_s\Delta_s\cdot\mathbf{1} + \sqrt{\frac{2\log(4HSA/\delta)}{N}}\sqrt{\mathrm{Var}_P(\widehat{V}_{t+1}^\star)} + \frac{2H\log(HSA/\delta)}{3N}\cdot\mathbf{1}\right) \\
&+C\epsilon_{\mathrm{opt}}\cdot\sqrt{\frac{H^2S\log(SA/\delta)}{N}}\cdot\mathbf{1} \\
&\le \sum_{t=h}^H \Gamma^{\hat{\pi}}_{h+1:t}\sqrt{\frac{2\log(4HSA/\delta)}{N}}\sqrt{\mathrm{Var}_P(\widehat{V}_{t+1}^\star)} + CH^2\sqrt{\frac{S\log(HSA/\delta)}{N}}\cdot\max_s\Delta_s\cdot\mathbf{1} + \frac{2H^2\log(HSA/\delta)}{3N}\cdot\mathbf{1} \\
&+C\epsilon_{\mathrm{opt}}\cdot\sqrt{\frac{H^2S\log(SA/\delta)}{N}}\cdot\mathbf{1}
\end{aligned}$$

Next note

$$\begin{aligned}
&\sqrt{\mathrm{Var}_P(\widehat{V}_h^\star)} := \sqrt{\mathrm{Var}_P\left(\widehat{V}_h^{\hat{\pi}^\star}\right)} = \sqrt{\mathrm{Var}_P\left(\widehat{V}_h^{\hat{\pi}^\star} - \widehat{V}_h^{\hat{\pi}} + \widehat{V}_h^{\hat{\pi}}\right)} \\
\leq &\sqrt{\mathrm{Var}_P\left(\widehat{V}_h^{\hat{\pi}}\right)} + \sqrt{\mathrm{Var}_P\left(\widehat{V}_h^{\hat{\pi}^\star} - \widehat{V}_h^{\hat{\pi}}\right)} \leq \sqrt{\mathrm{Var}_P\left(\widehat{V}_h^{\hat{\pi}}\right)} + \left\|\widehat{V}_h^{\hat{\pi}^\star} - \widehat{V}_h^{\hat{\pi}}\right\|_\infty \\
\leq &\sqrt{\mathrm{Var}_P\left(\widehat{V}_h^{\hat{\pi}}\right)} + \epsilon_{\mathrm{opt}} \cdot \mathbf{1} \leq \sqrt{\mathrm{Var}_P\left(V_h^{\hat{\pi}}\right)} + \sqrt{\mathrm{Var}_P\left(\widehat{V}_h^{\hat{\pi}} - V_h^{\hat{\pi}}\right)} + \epsilon_{\mathrm{opt}} \cdot \mathbf{1} \\
\leq &\sqrt{\mathrm{Var}_P\left(V_h^{\hat{\pi}}\right)} + \left\|\widehat{V}_h^{\hat{\pi}} - V_h^{\hat{\pi}}\right\|_\infty + \epsilon_{\mathrm{opt}} \cdot \mathbf{1} \leq \sqrt{\mathrm{Var}_P\left(V_h^{\hat{\pi}}\right)} + \left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty + \epsilon_{\mathrm{opt}} \cdot \mathbf{1}
\end{aligned} \tag{C.5}$$

Plug (C.5) back to above we obtain $\forall h \in [H]$,

$$\begin{aligned}
&\left|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right| \leq \sqrt{\frac{2\log(4HSA/\delta)}{N}} \sum_{t=h}^{H} \Gamma_{h+1:t}^{\hat{\pi}} \left(\sqrt{\mathrm{Var}_P\left(V_{t+1}^{\hat{\pi}}\right)} + \left\|\widehat{Q}_{t+1}^{\hat{\pi}} - Q_{t+1}^{\hat{\pi}}\right\|_\infty + \epsilon_{\mathrm{opt}} \cdot \mathbf{1}\right) \\
&+ CH^2 \sqrt{\frac{S\log(HSA/\delta)}{N}} \cdot \max_s \Delta_s \cdot \mathbf{1} + \frac{2H^2\log(HSA/\delta)}{3N} \cdot \mathbf{1} + C\epsilon_{\mathrm{opt}} \cdot \sqrt{\frac{H^2 S\log(SA/\delta)}{N}} \cdot \mathbf{1} \\
&\leq \sqrt{\frac{2\log(4HSA/\delta)}{N}} \sum_{t=h}^{H} \Gamma_{h+1:t}^{\hat{\pi}} \sqrt{\mathrm{Var}_P\left(V_{t+1}^{\hat{\pi}}\right)} + \sqrt{\frac{2\log(4HSA/\delta)}{N}} \sum_{t=h}^{H} \left\|\widehat{Q}_{t+1}^{\hat{\pi}} - Q_{t+1}^{\hat{\pi}}\right\|_\infty \\
&+ CH^2 \sqrt{\frac{S\log(HSA/\delta)}{N}} \cdot \max_s \Delta_s \cdot \mathbf{1} + \frac{2H^2\log(HSA/\delta)}{3N} \cdot \mathbf{1} + C_1\epsilon_{\mathrm{opt}} \cdot \sqrt{\frac{H^2 S\log(SA/\delta)}{N}} \cdot \mathbf{1}
\end{aligned} \tag{C.6}$$

Applying Lemma D.0.8 and the coarse uniform bound (Lemma D.0.11), we obtain the following result:

Lemma C.1.7. *Given $N > 0$ and $\epsilon_{opt} \leq \sqrt{H/S}$. With probability $1-\delta$, for all $h \in [H]$,*

$$\left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty \leq \sqrt{\frac{C_0 H^3\log(4HSA/\delta)}{N}} + \sqrt{\frac{2\log(4HSA/\delta)}{N}} \sum_{t=h}^{H} \left\|\widehat{Q}_{t+1}^{\hat{\pi}} - Q_{t+1}^{\hat{\pi}}\right\|_\infty + C' H^4 \frac{S\log(HSA/\delta)}{N}$$

Proof: Since

$$\begin{aligned}
\Delta_s &:= \max_t |\hat{u}_t^\star - u_t^\star| = \max_t \left|\widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s) - \left(V_t^\star(s) - V_{t+1}^\star(s)\right)\right| \\
&\leq 2 \cdot \max_t \left|\widehat{V}_t^\star(s) - V_t^\star(s)\right| \\
&= 2 \cdot \max_t \left|\max_\pi \widehat{V}_t^\pi(s) - \max_\pi V_t^\pi(s)\right| \\
&\leq 2 \cdot \max_{\pi \in \Pi_g, t \in [H]} \left\|\widehat{V}_t^\pi - V_t^\pi\right\|_\infty \leq C \cdot H^2 \sqrt{\frac{S\log(HSA/\delta)}{N}}
\end{aligned} \tag{C.7}$$

where the last inequality uses Lemma D.0.11. Then apply union bound w.p. $1-\delta/2$, we obtain

$\max_s \Delta_s \le C \cdot H^2\sqrt{\frac{S\log(HSA/\delta)}{N}}$. Note (C.6) holds with probability $1-\delta/2$, therefore plug above into (C.6) we obtain w.p. $1-\delta$,

$$\begin{aligned}
\left|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right| &\le \sqrt{\frac{2\log(4HSA/\delta)}{N}}\sum_{t=h}^{H}\Gamma_{h+1:t}^{\hat{\pi}}\sqrt{\mathrm{Var}_P\left(V_{t+1}^{\hat{\pi}}\right)} + \sqrt{\frac{2\log(4HSA/\delta)}{N}}\sum_{t=h}^{H}\left\|\widehat{Q}_{t+1}^{\hat{\pi}} - Q_{t+1}^{\hat{\pi}}\right\|_\infty \\
&+ C'H^4\frac{S\log(HSA/\delta)}{N}\cdot\mathbf{1} + C_1\epsilon_{\mathrm{opt}}\cdot\sqrt{\frac{H^2S\log(SA/\delta)}{N}}\cdot\mathbf{1} \\
&\le \left[\sqrt{\frac{C_0H^3\log(4HSA/\delta)}{N}} + \sqrt{\frac{2\log(4HSA/\delta)}{N}}\sum_{t=h}^{H}\left\|\widehat{Q}_{t+1}^{\hat{\pi}} - Q_{t+1}^{\hat{\pi}}\right\|_\infty + C'H^4\frac{S\log(HSA/\delta)}{N}\right]\cdot\mathbf{1},
\end{aligned}$$

where the last inequality uses Lemma D.0.8 and $\epsilon_{\mathrm{opt}} \le \sqrt{H/S}$ and renames $C' = C' + C_1$. Take $\|\cdot\|_\infty$ then obtain the result.

Lemma C.1.8. *Given $N > 0$. Define $C'' := 2\cdot\max(\sqrt{C_0}, C')$ where C' is the universal constant in Lemma C.1.7. When $N \ge 8H^2\log(4HSA/\delta)$, then with probability $1-\delta$, $\forall h \in [H]$,*

$$\begin{aligned}
\left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty &\le C''\sqrt{\frac{H^3\log(4HSA/\delta)}{N}} + C''\frac{H^4S\log(HSA/\delta)}{N}. \\
\left\|\widehat{Q}_h^{\pi^\star} - Q_h^{\pi^\star}\right\|_\infty &\le C''\sqrt{\frac{H^3\log(4HSA/\delta)}{N}} + C''\frac{H^4S\log(HSA/\delta)}{N}.
\end{aligned} \tag{C.8}$$

Proof: We prove by backward induction. For $h = H$, by Lemma C.1.7

$$\begin{aligned}
\left\|\widehat{Q}_H^{\hat{\pi}} - Q_H^{\hat{\pi}}\right\|_\infty &\le \sqrt{\frac{C_0H^3\log(4HSA/\delta)}{N}} + \sqrt{\frac{2\log(4HSA/\delta)}{N}}\left\|\widehat{Q}_{H+1}^{\hat{\pi}} - Q_{H+1}^{\hat{\pi}}\right\|_\infty + C'H^4\frac{S\log(HSA/\delta)}{N} \\
&= \sqrt{\frac{C_0H^3\log(4HSA/\delta)}{N}} + 0 + C'H^4\frac{S\log(HSA/\delta)}{N} \\
&\le C''\sqrt{\frac{H^3\log(4HSA/\delta)}{N}} + C''H^4\frac{S\log(HSA/\delta)}{N},
\end{aligned}$$

for general h, by condition we have $H\sqrt{\frac{2\log(4HSA/\delta)}{N}} \le 1/2$, therefore by Lemma C.1.7

$$\begin{aligned}
&\left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty \le \sqrt{\frac{C_0 H^3 \log(4HSA/\delta)}{N}} + \sqrt{\frac{2\log(4HSA/\delta)}{N}} \sum_{t=h}^{H} \left\|\widehat{Q}_{t+1}^{\hat{\pi}} - Q_{t+1}^{\hat{\pi}}\right\|_\infty + C' H^4 \frac{S\log(HSA/\delta)}{N} \\
&\le \sqrt{\frac{C_0 H^3 \log(4HSA/\delta)}{N}} + H\sqrt{\frac{2\log(4HSA/\delta)}{N}} \max_{t+1} \left\|\widehat{Q}_{t+1}^{\hat{\pi}} - Q_{t+1}^{\hat{\pi}}\right\|_\infty + C' H^4 \frac{S\log(HSA/\delta)}{N} \\
&\le \sqrt{\frac{C_0 H^3 \log(4HSA/\delta)}{N}} + C' H^4 \frac{S\log(HSA/\delta)}{N} \\
&+ \frac{1}{2}\left(C''\sqrt{\frac{H^3\log(4HSA/\delta)}{N}} + C''\frac{H^4 S\log(HSA/\delta)}{N} \right) \\
&\le C''\sqrt{\frac{H^3\log(4HSA/\delta)}{N}} + C''\frac{H^4 S\log(HSA/\delta)}{N}
\end{aligned}$$

The proof of the second claim is even easier since $\pi^\star$ is no longer a random policy and it is really just a non-uniform point-wise OPE. There are multiple ways to prove it and we leave it as an exercise to avoid redundancy: 1. Follow the same proving pipeline as $\left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty$ used; 2. Mimic the procedure of point-wise OPE result in Lemma 3.4. in [7].

Remark 17. *Note the higher order term has dependence $H^4 S$, which is somewhat unsatisfactory. We use the* recursion-back *trick to further reduce it to $H^{3.5}S^{0.5}$.*

Lemma C.1.9. *Given $N > 0$. There exists universal constants C_1, C_2 such that when $N \ge C_1 H^2 \log(HSA/\delta)$, then with probability $1-\delta$, $\forall h \in [H]$,*

$$\left\|\widehat{Q}_h^{\hat{\pi}} - Q_h^{\hat{\pi}}\right\|_\infty \le C_2\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + C_2\frac{H^3\sqrt{HS}\log(HSA/\delta)}{N}. \tag{C.9}$$

and

$$\left\|\widehat{Q}_h^{\pi^\star} - Q_h^{\pi^\star}\right\|_\infty \le C_2\sqrt{\frac{H^3\log(HSA/\delta)}{N}} + C_2\frac{H^3\sqrt{HS}\log(HSA/\delta)}{N}.$$

Proof:
Note

$$\begin{aligned}
\widehat{V}_t^\star(s) - V_t^\star(s) &:= \widehat{V}_t^{\hat{\pi}^\star}(s) - V_t^{\pi^\star}(s) \\
&= \widehat{V}_t^{\hat{\pi}^\star}(s) - V_t^{\hat{\pi}^\star}(s) + V_t^{\hat{\pi}^\star}(s) - V_t^{\pi^\star}(s) \\
&\le \widehat{V}_t^{\hat{\pi}^\star}(s) - V_t^{\hat{\pi}^\star}(s) \le \left|\widehat{V}_t^{\hat{\pi}^\star}(s) - V_t^{\hat{\pi}^\star}(s)\right|
\end{aligned} \tag{C.10}$$

and similarly $V_t^\star(s) - \widehat{V}_t^\star(s) \le \left|\widehat{V}_t^{\pi^\star}(s) - V_t^{\pi^\star}(s)\right|$, therefore by Lemma C.1.8 (and use $||\widehat{V}_t^{\pi} -$

$V_t^\pi||_\infty \le ||\widehat{Q}_t^\pi - Q_t^\pi||_\infty)$, with probability $1-\delta$,

$$\Delta_s \le 2 \cdot \sup_t \left\| V_t^\star - \widehat{V}_t^\star \right\| \le 2 \max_{\widehat{\pi}^\star, \pi^\star} \sup_t \left\| \widehat{V}_t^\pi - V_t^\pi \right\|_\infty \le C_2 \sqrt{\frac{H^3 \log(HSA/\delta)}{N}} + C_2 \frac{H^4 S \log(HSA/\delta)}{N},$$

where the second inequality uses (C.10). This replaces the crude bound of $O(\sqrt{H^4 S \log(HSA/\delta)/N})$ for $\max_s \Delta_s$ (recall (C.7)) by $O(\sqrt{H^3 \log(HSA/\delta)/N})$.

Plug this back to (C.6) and repeat the similar analysis we end up with (C.9). The second result is similarly proved.

C.1.7 Proof of Theorem 4.6.1

Proof: [Proof of Theorem 4.6.1] Note $n_{s,a} = \sum_{i=1}^n \sum_{t=1}^H \mathbf{1}[s_t^{(i)} = s, a_t^{(i)} = a]$, which implies

$$\mathbb{E}[n_{s,a}] = \mathbb{E}\left[\sum_{i=1}^n \sum_{t=1}^H \mathbf{1}[s_t^{(i)} = s, a_t^{(i)} = a]\right] = n \cdot \sum_{t=1}^H d_t^\mu(s,a).$$

Or equivalently, $n_{s,a}$ follows Binomial$(n, \sum_{t=1}^H d_t^\mu(s,a))$. Then apply the first result of Lemma D.0.1 by taking $\theta = 1/2$, we have when $n > 1/d_m \cdot \log(HSA/\delta)$[3], then with probability $1-\delta$,

$$n_{s,a} \ge \frac{1}{2} n \cdot \sum_{t=1}^H d_t^\mu(s,a), \quad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

This further implies w.p. $1-\delta$, $n_{s,a} \ge \frac{1}{2} n \cdot \sum_{t=1}^H d_t^\mu(s,a) = \frac{1}{2} n \cdot H \cdot d^\mu(s,a) \ge \frac{1}{2} nH \cdot d_m$ and further ensures

$$N := \min_{s,a} n_{s,a} \ge \frac{1}{2} nH \cdot d_m.$$

Finally, by applying the above to Lemma C.1.9, we can overcome the condition on N and obtain the stated result.

C.2 Proof of minimax lower bound for model-based global uniform OPE

Proof: [Proof of Theorem 4.5.1] In particular, we first focus on the case where $H = 2$ and extend the result of $H = 2$ to the general $H \ge 3$ at the end.

First of all, by Definition 4.3.1 let $\widehat{P}$ be the learned transition by certain model-based method. Since we assume r_h is known and by convention $Q_{H+1}^\pi = 0$ for any π, then by Bellman

[3]The exact sufficient condition for applying Lemma D.0.1 is $n > 1/\sum_{t=1}^H d_t(s,a) \cdot \log(HSA/\delta)$ for all s, a. However, since $\sum_{t=1}^H d_t(s,a) \ge H d_m \ge d_m$, our condition $n > 1/d_m \cdot \log(HSA/\delta)$ used here is a much stronger version thus Lemma D.0.1 apply.

equation

$$\widehat{Q}_h^\pi = r_h + \widehat{P}^{\pi_{h+1}} \widehat{Q}_{h+1}^\pi, \ \forall h \in [H].$$

In particular, $\widehat{Q}_{H+1}^\pi = Q_{H+1}^\pi = 0$, and this implies

$$\widehat{Q}_H^\pi = r_H + \widehat{P}^{\pi_{H+1}} \widehat{Q}_{H+1}^\pi = r_H; \quad Q_H^\pi = r_H + P^{\pi_{H+1}} Q_{H+1}^\pi = r_H + 0 = r_H$$

Now, again by definition of Bellman equation

$$\widehat{Q}_{H-1}^\pi = r_{H-1} + \widehat{P}^{\pi_H} \widehat{Q}_H^\pi = r_{H-1} + \widehat{P}^{\pi_H} r_H$$
$$Q_{H-1}^\pi = r_{H-1} + P^{\pi_H} Q_H^\pi = r_{H-1} + P^{\pi_H} r_H$$

Therefore

$$\begin{aligned}
&\sup_{\pi \in \Pi_g} \left\| \widehat{Q}_{H-1}^\pi - Q_{H-1}^\pi \right\|_\infty = \sup_{\pi \in \Pi_g} \left\| \left(\widehat{P}^{\pi_H} - P^{\pi_H} \right) r_H \right\|_\infty \\
&= \sup_{\pi \in \Pi_g} \left\| \left(\widehat{P} - P \right) r_H^{\pi_H} \right\|_\infty = \sup_{\pi \in \Pi_g} \sup_{s,a} \left| \left(\widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right) r_H^{\pi_H} \right| \\
&= \sup_{s,a} \sup_{\pi \in \Pi_g} \left| \left(\widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right) r_H^{\pi_H} \right|,
\end{aligned}$$

where $P^{\pi_H} \in \mathbb{R}^{S \cdot A \times S \cdot A}$, $r_H \in \mathbb{R}^{S \cdot A}$, $P \in \mathbb{R}^{S \cdot A \times S}$ and $r_H^{\pi_H} \in \mathbb{R}^S$. Note $A \geq 2$, so we can choose an instance of r_H as (there are at least two actions since $A \geq 2$)

$$(r_H(s, a_1), r_H(s, a_2), ...) := (1, 0, ...) \qquad \forall s \in \mathcal{S}.$$

Above implies: if $\pi_H(s) = a_1$, then $r_H^{\pi_H}(s) = 1$; if $\pi_H(s) = a_2$, then $r_H^{\pi_H}(s) = 0$; ...

Hence, if Π_g is the global deterministic policy class, then $r_H^{\pi_H}$ can traverse all the S-dimensional vectors with either 0 or 1 in each coordinate, which is exactly

$$\left\{ r_H^{\pi_H} \in \mathbb{R}^S : \pi_H \in \Pi_g \right\} \supset \{0, 1\}^S.$$

Now let us first consider fixed s, a. Then with this choice of r, above implies

$$\begin{aligned}
&\sup_{\pi \in \Pi_g} \left| \left(\widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right) r_H^{\pi_H} \right| \geq \sup_{r \in \{0,1\}^S} \left| \left(\widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right) \cdot r \right| \\
&= \sup_{r \in \{0,1\}^S} \left| \sum_{i : r_i = 1} \left(\widehat{P}(s_i|s,a) - P(s_i|s,a) \right) \right|
\end{aligned}$$

Let $I_+ := \{i \in [S] : s.t. \ \widehat{P}(s_i|s,a) - P(s_i|s,a) > 0\}$ be the set of indices where $\widehat{P}(s_i|s,a) - P(s_i|s,a)$ are positive and $I_- := \{i \in [S] : s.t. \ \widehat{P}(s_i|s,a) - P(s_i|s,a) < 0\}$ be the set of

indices where $\widehat{P}(s_i|s,a) - P(s_i|s,a)$ are negative, then we further have

$$\begin{aligned}
&\sup_{r\in\{0,1\}^S} \left| \sum_{i:r_i=1} \left(\widehat{P}(s_i|s,a) - P(s_i|s,a) \right) \right| \\
\geq &\max \left\{ \left| \sum_{i\in I+} [\widehat{P}(s_i|s,a) - P(s_i|s,a)] \right|, \left| \sum_{i\in I-} [\widehat{P}(s_i|s,a) - P(s_i|s,a)] \right| \right\} \\
= &\max \left\{ \sum_{i\in I+} \left| \widehat{P}(s_i|s,a) - P(s_i|s,a) \right|, \sum_{i\in I-} \left| \widehat{P}(s_i|s,a) - P(s_i|s,a) \right| \right\}
\end{aligned}$$

On the other hand, we have

$$\sum_{i\in I+} \left| \widehat{P}(s_i|s,a) - P(s_i|s,a) \right| + \sum_{i\in I-} \left| \widehat{P}(s_i|s,a) - P(s_i|s,a) \right| = \left\| \widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right\|_1$$

since $\widehat{P}(s_i|s,a) - P(s_i|s,a) = 0$ contributes nothing to the l_1 norm. Combine all the steps together, we obtain

$$\sup_{\pi\in\Pi_g} \left\| \widehat{Q}^\pi_{H-1} - Q^\pi_{H-1} \right\|_\infty \geq \sup_{s,a} \frac{1}{2} \left\| \widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right\|_1 \geq 1c \cdot \sup_{s,a} \sqrt{\frac{S}{n_{s,a}}} \geq 2c' \sqrt{\frac{S}{nd_m}} \quad \text{(C.11)}$$

holds with constant probability p. Here $n_{s,a} = \sum_{h=1}^H \sum_{i=1}^n \mathbf{1}[s_h^{(i)} = s, a_h^{(i)} = a]$ is the number of data pieces visited (s,a) in n episodes. Now we explain how to obtain 1 and 2. In particular, we first explain 2.

Explain 2. Recall we consider the case $H = 2$. Then

$$\mathbb{E}\left[n_{s,a}\right] = \mathbb{E}\left[\sum_{h=1}^H \sum_{i=1}^n \mathbf{1}[s_h^{(i)} = s, a_h^{(i)} = a]\right] = n \sum_{i=1}^2 \mathbb{E}\left[\mathbf{1}[s_h^{(1)} = s, a_h^{(1)} = a]\right] = n \sum_{h=1}^2 d_h^\mu(s,a)$$

i.e. $n_{s,a}$ is a Binomial random variable with parameter n and $\sum_{h=1}^2 d_h^\mu(s,a)$. Then by Lemma D.0.1, choose $\theta = \frac{1}{2}$, apply the second result, we obtain when $n > (1/2d_m) \cdot \log(SA/\delta)$[4], with probability $1-\delta$

$$n_{s,a} \leq \frac{3}{2} n \cdot \sum_{h=1}^2 d_h^\mu(s,a), \quad \forall s,a$$

Next, similar to the lower bound proof (Theorem G.2.) of [7], we can choose μ and M (**near uniform** but not exact uniform) such that $d_h^\mu(s,a) \leq C \cdot d_m$, which further implies $n_{s,a} \leq C \cdot n \cdot d_m$, $\forall s,a$. Summarize above we end up with the following Lemma:

[4]By Lemma D.0.1,the inequality holds as long as $n \geq 1/\sum_{h=1}^2 d_h^\mu(s,a) \log(SA/\delta)$, here $n > (1/2d_m) \cdot \log(SA/\delta)$ is a stronger sufficient condition.

Lemma C.2.1. *Suppose $n \geq (1/2d_m) \cdot \log(SA/\delta)$, then*

$$\sup_{\mu,M} \mathbb{P}\left[\sqrt{\frac{1}{n_{s,a}}} \geq C \cdot \sqrt{\frac{1}{n \cdot d_m}},\ \forall s,a\right] \geq 1-\delta$$

Explain 1. To make the explanation rigorous, we first fix a pair (s,a) and conditional on $n_{s,a}$. Then by a direct translation of Lemma D.0.9, we have

$$\inf_{\widehat{P}} \sup_{P(\cdot|s,a)\in\mathcal{M}_S} \mathbb{P}\left[\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \geq \frac{1}{8}\sqrt{\frac{eS}{2n_{s,a}}} - o(\cdot)\middle| n_{s,a} \geq \frac{e}{32}S\right] \geq p,$$

where $o(\cdot)$ is some exponentially small term in S, n. Now we consider everything under the condition $n \geq \frac{e}{32} \cdot S/d_m \log(SA/\delta)$. Next again take $\theta = 1/2$, then by the first result of Lemma D.0.1, with probability $1-\delta$,

$$n_{s,a} \geq \frac{1}{2}n \cdot \sum_{h=1}^{2} d_h^{\mu}(s,a) \geq n \cdot d_m \geq \frac{e}{32}S\log(SA/\delta).$$

where the last inequality uses the condition $n \geq \frac{e}{32} \cdot S/d_m \log(SA/\delta)$. Therefore this implies

$$\begin{aligned}
&\inf_{\widehat{P}} \sup_{P(\cdot|s,a)\in\mathcal{M}_S} \mathbb{P}\left[\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \geq \frac{1}{8}\sqrt{\frac{eS}{2n_{s,a}}} - o(\cdot)\right] \\
=&\inf_{\widehat{P}} \sup_{P(\cdot|s,a)\in\mathcal{M}_S} \left(\mathbb{P}\left[\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \geq \frac{1}{8}\sqrt{\frac{eS}{2n_{s,a}}} - o(\cdot)\middle| n_{s,a} \geq \frac{e}{32}S\right] \cdot \mathbb{P}\left[n_{s,a} \geq \frac{e}{32}S\right]\right. \\
&\left.+\mathbb{P}\left[\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \geq \frac{1}{8}\sqrt{\frac{eS}{2n_{s,a}}} - o(\cdot)\middle| n_{s,a} \leq \frac{e}{32}S\right] \cdot \mathbb{P}\left[n_{s,a} \leq \frac{e}{32}S\right]\right) \\
\geq&\inf_{\widehat{P}} \sup_{P(\cdot|s,a)\in\mathcal{M}_S} \mathbb{P}\left[\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \geq \frac{1}{8}\sqrt{\frac{eS}{2n_{s,a}}} - o(\cdot)\middle| n_{s,a} \geq \frac{e}{32}S\right] \cdot \mathbb{P}\left[n_{s,a} \geq \frac{e}{32}S\right] \\
\geq& p \cdot (1-\delta),
\end{aligned}$$

To sum up, we have the following lemma:

Lemma C.2.2. *Let $n \geq \frac{e}{32}S/d_m \cdot \log(SA/\delta)$, then there exists a $0 < p < 1$,*

$$\inf_{\widehat{P}} \sup_{P(\cdot|s,a)\in\mathcal{M}_S} \mathbb{P}\left[\|\widehat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \geq \frac{1}{8}\sqrt{\frac{eS}{2n_{s,a}}} - o(\cdot)\right] \geq p \cdot (1-\delta).$$

Now we finish the proof for the case where $H = 2$. First note by (C.11),

$$\sup_{\pi\in\Pi_g} \left\| \widehat{Q}^{\pi}_{H-1} - Q^{\pi}_{H-1} \right\|_{\infty} \geq \sup_{s,a} \frac{1}{2} \left\| \widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right\|_1$$

with probability 1, therefore by (C.11), Lemma C.2.1, Lemma C.2.2 we have

$$\inf_{\hat{P}} \sup_{P\in\mathcal{M}_S} \mathbb{P}\left[\sup_{\pi\in\Pi_g} \left\| \widehat{Q}^{\pi}_{H-1} - Q^{\pi}_{H-1} \right\|_{\infty} \geq C \cdot \sqrt{\frac{S}{nd_m}} \right] \geq p(1-\delta) - \delta$$

when $n \geq c \cdot S/d_m \log(SA/\delta)$ for some $c \geq \frac{e}{32}$. Above holds for any δ.

It is easy to check $\frac{3}{2}\frac{p}{1+p} \leq 1$, therefore, in particular we set $\delta = \frac{3}{2}\frac{p}{1+p}$, direct calculation shows

$$p(1-\delta) - \delta = \frac{p}{2},$$

which completes the proof for $H = 2$.

Extend to the general $H \geq 3$.

Step 1. Similar to the decomposition in section C.1.4, we also have:

$$\hat{Q}^{\pi}_t - Q^{\pi}_t = \sum_{h=t}^{H} \hat{\Gamma}^{\pi}_{t+1:h} (\widehat{P} - P) V^{\pi}_{h+1}$$

Step 2. Now choosing rewards recursively from back (with $||r_H||_{\infty} = c$ sufficiently small) such that $1 \geq r_h \geq (||r_{h+1}||_{\infty} + \ldots + ||r_H||_{\infty})$ element-wisely $\forall h$, and $\max_{s,a} r_h(s,a) = 3\min_{s,a} r_h(s,a)$. We denote $r_{h,max} := \max_{s,a} r_h(s,a)$ and $r_{h,min} := \min_{s,a} r_h(s,a)$. This choice guarantees:

$$r_{h,min} := \min_{s,a} r_h(s,a) > ||P^{\pi_{h+1}} r_{h+1} + .. + P^{\pi_{h+1:H}} r_H||_{\infty}$$

since P^{π_h} is row-stochastic.

Step 3. Next note $V^{\pi}_h = r_h + P^{\pi_{h+1}} r_{h+1} + .. + P^{\pi_{h+1:H}} r_H$, so set $\left(r_h\left(s,a_1\right), r_h\left(s,a_2\right), \ldots\right) := (\max_{s,a} r_h(s,a), \min_{s,a} r_h(s,a), \ldots)$, then choose π_h similar to the $H = 2$ case and use **Step 1** and **Step 2** we have

$$\begin{aligned} |(\hat{P}_{s,a} - P_{s,a}) V^{\pi}_h| \geq & \frac{1}{2} ||\hat{P}_{s,a} - P_{s,a}||_1 \cdot (r_{h,max} - r_{h,min} - (P^{\pi_{h+1}} r_{h+1} + .. + P^{\pi_{h+1:H}} r_H)) \\ \geq & \frac{1}{2} ||\hat{P}_{s,a} - P_{s,a}||_1 \cdot r_{h,min} \geq \frac{1}{2} ||\hat{P}_{s,a} - P_{s,a}||_1 \cdot c \end{aligned}$$

where the reasoning of the first inequality is similar to the case of $H = 2$. Next use $\hat{\Gamma}^{\pi}_{t+1:h}$ is row-stochastic then from **Step 1** and take the sum we have

$$||\hat{Q}^{\pi}_1 - Q^{\pi}_1||_{\infty} \geq \frac{1}{2} c \cdot H \min_{s,a} ||\hat{P}_{s,a} - P_{s,a}||_1.$$

for such choice of rewards and π.

Step 4. However, in the above construction c actually depends on H due to the design $1 \geq r_h \geq (||r_{h+1}||_\infty + \ldots + ||r_H||_\infty)$. To get a universal constant c we could use the bound $||\hat{Q}_1^\pi - Q_1^\pi||_\infty \gtrsim r_{\frac{H}{2},min} \cdot \frac{H}{2} \min_{s,a} ||\hat{P}_{s,a} - P_{s,a}||_1$ instead, where $r_{\frac{H}{2},min}$ in Step 2 is universally lower bounded. Then we apply $||\hat{P}_{s,a} - P_{s,a}||_1 \gtrsim \Omega(\sqrt{S/nd_m})$ to obtain the lower bound $\Omega(\sqrt{H^2 S/nd_m})$.

Remark 18. *We point out while our lower bound of $\Omega(H^2 S/d_m\epsilon^2)$ for uniform OPE appears to be qualitatively similar to the lower bound of $\Omega(H^2 S^2 A/\epsilon^2)$ derived for the online reward-free RL setting jin2020reward, our result is not implied by theirs and cannot be proven by directly adapting their construction. Those two results are in principle, different since: the result in jin2020reward is learning-oriented where they define the problem class on $O(S)$ states and forcing $\Omega(SA/\epsilon^2)$ episodes in each state and end up with $O(S^2 A/\epsilon^2)$ complexity; our result is evaluation-oriented where we need reduce the uniform evaluation problem to estimating probability distribution in ℓ_1-error. The global uniform OPE and the reward-free setting are also different tasks (one cannot imply the other): the former deals with uniform convergence over all policies but with a fixed reward while the latter aims at learning simultaneously over all rewards.*

C.3 Proof for optimal offline learning (Corollary 4.6.1)

Proof: This is a corollary of Theorem 4.6.1. Indeed, by taking $\widehat{\pi} = \widehat{\pi}^\star$, we first have

$$\left\|\widehat{V}_1^{\widehat{\pi}^\star} - V_1^{\widehat{\pi}^\star}\right\|_\infty \leq \left\|\widehat{Q}_1^{\widehat{\pi}^\star} - Q_1^{\widehat{\pi}^\star}\right\|_\infty \leq C\left[\sqrt{\frac{H^2\iota}{nd_m}} + \frac{H^{2.5}S^{0.5}\iota}{nd_m}\right].$$

Similar to the second result in Lemma C.1.9, we also have

$$\left\|\widehat{V}_1^{\pi^\star} - V_1^{\pi^\star}\right\|_\infty \leq \left\|\widehat{Q}_1^{\pi^\star} - Q_1^{\pi^\star}\right\|_\infty \leq C\left[\sqrt{\frac{H^2\iota}{nd_m}} + \frac{H^{2.5}S^{0.5}\iota}{nd_m}\right].$$

Next, recall the definition of $\widehat{\pi} \in \Pi_l$ that

$$\left\|\widehat{V}_1^{\widehat{\pi}^\star} - \widehat{V}_1^{\widehat{\pi}}\right\|_\infty \leq \epsilon_{\text{opt}},$$

and Theorem 4.6.1 again that

$$\left\|\widehat{V}_1^{\widehat{\pi}} - V_1^{\widehat{\pi}}\right\|_\infty \leq \left\|\widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}}\right\|_\infty \leq C\left[\sqrt{\frac{H^2\iota}{nd_m}} + \frac{H^{2.5}S^{0.5}\iota}{nd_m}\right].$$

Therefore

$$\begin{aligned}
V_1^{\pi^\star} - V_1^{\widehat{\pi}} &= V_1^{\pi^\star} - V_1^{\widehat{\pi}^\star} + V_1^{\widehat{\pi}^\star} - V_1^{\widehat{\pi}} \\
&\leq \max_{\widehat{\pi}^\star, \pi^\star} \left\| \widehat{V}_1^{\pi} - V_1^{\pi} \right\|_\infty + V_1^{\widehat{\pi}^\star} - V_1^{\widehat{\pi}} \\
&= \max_{\widehat{\pi}^\star, \pi^\star} \left\| \widehat{V}_1^{\pi} - V_1^{\pi} \right\|_\infty + \left(V_1^{\widehat{\pi}^\star} - \widehat{V}_1^{\widehat{\pi}^\star} \right) + \left(\widehat{V}_1^{\widehat{\pi}^\star} - \widehat{V}_1^{\widehat{\pi}} \right) + \left(\widehat{V}_1^{\widehat{\pi}} - V_1^{\widehat{\pi}} \right) \\
&\leq 3C \left[\sqrt{\frac{H^2 \iota}{n d_m}} + \frac{H^{2.5} S^{0.5} \iota}{n d_m} \right] + \left\| \widehat{V}_1^{\widehat{\pi}^\star} - \widehat{V}_1^{\widehat{\pi}} \right\|_\infty \cdot \mathbf{1} \\
&\leq 3C \left[\sqrt{\frac{H^2 \iota}{n d_m}} + \frac{H^{2.5} S^{0.5} \iota}{n d_m} \right] + \epsilon_{\text{opt}} \cdot \mathbf{1}.
\end{aligned}$$

This completes the proof.

C.4 Proof for optimal offline Task-agnostic learning (Theorem 4.7.1)

Proof: Recall the definition of offline task-agnostic setting, where K tasks corresponds to K MDPs $M_k = (\mathcal{S}, \mathcal{A}, P, r_k, H, d_1)$ with different mean reward functions r_k's. Since the incremental number of rewards do not incur randomness, therefore by Corollary 4.6.1, choose $\widehat{\pi}_k = \widehat{\pi}_k^\star$ and apply a union bound we obtain with probability $1 - \delta$,

$$\begin{aligned}
\sup_{k \in [K]} ||V_{1,M_k}^\star - V_{1,M_k}^{\widehat{\pi}_k^\star}||_\infty &\leq O\left[\sqrt{\frac{H^2 \log(HSAK/\delta)}{n d_m}} + \frac{H^{2.5} S^{0.5} \log(HSAK/\delta)}{n d_m} \right] \\
&= O\left[\sqrt{\frac{H^2 (\iota + \log(K))}{n d_m}} + \frac{H^{2.5} S^{0.5} (\iota + \log(K))}{n d_m} \right],
\end{aligned}$$

which completes the proof.

Remark 19. *We stress that Section 3 of [84] claims the definition of task-agnostic RL setting embraces one challenge that $r_k^{(i)}$'s are the observed random realizations and the need to accurately estimate mean rewards r_k's causes the additional $\log(K)$ dependence. However, for offline case, this is not essential since, by straightforward calculation, estimating $r_k^{(i)}$'s accurately only requires $\tilde{O}(\log(K)/d_m \epsilon^2)$ samples, which is of lower order comparing to $\tilde{O}(H^2 \log(K)/d_m \epsilon^2)$ learning bound. Therefore, in Definition 4.7.1 we do not incorporate the random version statement for reward r_k.*

C.4.1 Offline Learning in the Constrained MDPs (CMDP)

Recently, there is a line of studies in the Constrained Markov Decision Processes (CMDP) (*e.g.* [**?**]), where the MDP $M = (S, \mathcal{A}, P, H, d_1)$. When the reward is set to be r, it defines the objective function V_r^π and there is another utility function g that defines the constraint. To be concrete, the objective formualted as:

$$\underset{\pi\in\Delta(\mathcal{A}|S,H)}{\text{maximize}} V_{r,1}^\pi\left(x_1\right) \text{ subject to } V_{g,1}^\pi\left(x_1\right) \geq b \tag{C.12}$$

where $b \in (0, H]$ is some constraint threshold. In addition, the formulation needs a Slater condition that: there exists $\gamma > 0$ and $\bar{\pi} \in \Delta(\mathcal{A}|S, H)$ such that $V_{g,1}^{\bar{\pi}}(x_1) \geq b + \gamma$.

Let $\pi^\star$ be the optimal solution that is compatible with the programming (C.12) (note this is **different** from the optimal policy that maximizes $V_{r,1}^\pi$ only), then by feasibility it satisfies $V_{g,1}^{\pi^\star} \geq b$.

Now let $\hat{\pi}^\star$ be the solution of the empirical program:

$$\underset{\pi\in\Delta(\mathcal{A}|S,H)}{\text{maximize}} \widehat{V}_{r,1}^\pi\left(x_1\right) \text{ subject to } \widehat{V}_{g,1}^\pi\left(x_1\right) \geq b \tag{C.13}$$

then we can show $\hat{\pi}^\star$ is a near-optimal solution for (C.12) via the local uniform convergence guarantee (Theorem 4.6.1).

Indeed, define a surrogate program:

$$\underset{\pi\in\Delta(\mathcal{A}|S,H)}{\text{maximize}} \widehat{V}_{r,1}^\pi\left(x_1\right) \text{ subject to } V_{g,1}^\pi\left(x_1\right) \geq b \tag{C.14}$$

and let $\bar{\pi}^\star$ be the solution for (C.14). Then apparently $\bar{\pi}^\star$ satisfies $V_{g,1}^{\bar{\pi}^\star}\left(x_1\right) \geq b$. Moreover, we have

$$\begin{aligned}
V_{r,1}^{\pi^\star} - V_{r,1}^{\bar{\pi}^\star} =& V_{r,1}^{\pi^\star} - \widehat{V}_{r,1}^{\pi^\star} + \widehat{V}_{r,1}^{\pi^\star} - \widehat{V}_{r,1}^{\bar{\pi}^\star} + \widehat{V}_{r,1}^{\bar{\pi}^\star} - V_{r,1}^{\bar{\pi}^\star} \\
\leq& V_{r,1}^{\pi^\star} - \widehat{V}_{r,1}^{\pi^\star} + 0 + \widehat{V}_{r,1}^{\bar{\pi}^\star} - V_{r,1}^{\bar{\pi}^\star} \\
\leq& 2\sup_\pi |V_{r,1}^\pi - \widehat{V}_{r,1}^\pi|
\end{aligned}$$

On the other hand, by local uniform convergence guarantee, $|V_{g,1}^\pi - \widehat{V}_{g,1}^\pi| \leq \tilde{O}(\sqrt{H^2/nd_m})$ for all π in the $\sqrt{H/S}$-neighborhood of $\hat{\pi}^\star$ (w.r.t g). This implies

$$V_{r,1}^{\pi^\star} - V_{r,1}^{\hat{\pi}^\star} \leq 2\sup_\pi |V_{r,1}^\pi - \widehat{V}_{r,1}^\pi| + \tilde{O}(\sqrt{H^2/nd_m})$$

and the violation of the constraint is bounded by $\tilde{O}(\sqrt{H^2/nd_m})$. This means any approach that solves (C.13) is near-optimal for the original constrained MDP task given the uniform convergence guarantee.

C.5 Proof for optimal offline Reward-free learning (Theorem 4.7.2)

Similar to before, recall $n_{s,a} = \sum_{h=1}^{H}\sum_{i=1}^{n} \mathbf{1}[s_h^{(i)} = s, a_h^{(i)} = a]$. We first prove two lemmas which essentially provide a version of *"Maximal Bernstein inequality"*. We first fix a pair (s,a) and then conditional on $n_{s,a}$.

Lemma C.5.1. *We define* $\epsilon_1 = \sqrt{\frac{1}{HS^2}}$. *Let* $\mathcal{G} = \{[i_1\epsilon_1, i_2\epsilon_1, \ldots, i_S\epsilon_1]^\top | i_1, i_2, \ldots, i_S \in \mathbb{Z}\} \cap [0,H]^S$ *be the* S*-dimensional grid. Next define* $\iota_1 = \log[(\sqrt{H^3S^2})^S/\delta]$. *Then with probability* $1-\delta$,

$$\left|(P_{s,a} - \widehat{P}_{s,a})w\right| \le \sqrt{\frac{2\mathrm{Var}_{s,a}(w)\iota_1}{n_{s,a}}} + \frac{2H\iota_1}{3n_{s,a}}, \quad \forall w \in \mathcal{G}.$$

This is by the direct application of Bernstein inequality with a union bound, where the cardinality of $\mathcal{G}$ is

$$\left(\frac{H}{\epsilon_1}\right)^S = \left(\sqrt{H^3S^2}\right)^S.$$

Lemma C.5.2. *Let the* S*-dimensional grid be* $\mathcal{G} = \{[i_1\epsilon_1, i_2\epsilon_1, \ldots, i_S\epsilon_1]^\top | i_1, i_2, \ldots, i_S \in \mathbb{Z}\} \cap [0,H]^S$ *and define* $\iota_1 = \log[(\sqrt{H^3S^2})^S/\delta]$. *It holds with probability* $1-\delta$,

$$\left|(P_{s,a} - \widehat{P}_{s,a})v\right| \le \sqrt{\frac{2\mathrm{Var}_{s,a}(v)\iota_1}{n_{s,a}}} + C\sqrt{\frac{\iota_1}{n_{s,a}HS}} + \frac{2H\iota_1}{3n_{s,a}}, \quad \forall\, v \in [0,H]^S.$$

Proof: Let $z := \mathrm{Proj}_{\mathcal{G}}(v)$. Then by design of $\mathcal{G}$ we have

$$\|z - v\|_\infty \le \epsilon_1 = \sqrt{\frac{1}{HS^2}}.$$

Therefore we obtain $\forall v \in [0, H]^S$,

$$\begin{aligned}
\left|(P_{s,a} - \widehat{P}_{s,a})v\right| &\le \left|(P_{s,a} - \widehat{P}_{s,a})(v - z)\right| + \left|(P_{s,a} - \widehat{P}_{s,a})z\right| \\
&\le \left\|P_{s,a} - \widehat{P}_{s,a}\right\|_1 \|z - v\|_\infty + \left|(P_{s,a} - \widehat{P}_{s,a})z\right| \\
&\le c\sqrt{\frac{S}{n_{s,a}}}\|z - v\|_\infty + \sqrt{\frac{2\mathrm{Var}_{s,a}(z)\iota_1}{n_{s,a}}} + \frac{2H\iota_1}{3n_{s,a}} \\
&\le c\sqrt{\frac{S}{n_{s,a}}}\|z - v\|_\infty + \sqrt{\frac{2\|z - v\|_\infty^2\,\iota_1}{n_{s,a}}} + \sqrt{\frac{2\mathrm{Var}_{s,a}(v)\iota_1}{n_{s,a}}} + \frac{2H\iota_1}{3n_{s,a}} \\
&\le C\sqrt{\frac{S\iota_1}{n_{s,a}}}\|z - v\|_\infty + \sqrt{\frac{2\mathrm{Var}_{s,a}(v)\iota_1}{n_{s,a}}} + \frac{2H\iota_1}{3n_{s,a}} \\
&\le C\sqrt{\frac{\iota_1}{n_{s,a}HS}} + \sqrt{\frac{2\mathrm{Var}_{s,a}(v)\iota_1}{n_{s,a}}} + \frac{2H\iota_1}{3n_{s,a}}.
\end{aligned}$$

where the third inequality uses Lemma C.5.1 and Lemma D.0.10.

Then recall $N := \min_{s,a} n_{s,a}$, by Lemma C.5.2 and a union bound we obtain with probability $1 - \delta$, element-wisely,

$$\left|(P - \widehat{P})v\right| \le C \cdot \left(\sqrt{\frac{2\mathrm{Var}_{s,a}(v)\iota_2}{N}} + 2\sqrt{\frac{\iota_2}{N \cdot HS}} + \frac{2H\iota_2}{3N}\right) \cdot \mathbf{1}, \quad \forall\, v \in [0, H]^S, \tag{C.15}$$

where $\iota_2 = S\log(HSA/\delta)$.

Remark 20. *Equation C.15 is a form of maximal Bernstein inequality as it keeps validity for all $v \in [0, H]^S$. The price for this stronger result is the extra S factor (coming from ι_2) in the dominate term.*

Now, for *any* reward r, by (empirical) Bellman equation we have element-wisely:

$$\begin{aligned}
\widehat{Q}_h^{\widehat{\pi}^\star} - Q_h^{\widehat{\pi}^\star} &= r_h + \widehat{P}^{\widehat{\pi}^\star_{h+1}}\widehat{Q}_{h+1}^{\widehat{\pi}^\star} - r_h - P^{\widehat{\pi}^\star_{h+1}}Q_{h+1}^{\widehat{\pi}^\star} \\
&= \left(\widehat{P}^{\widehat{\pi}^\star_{h+1}} - P^{\widehat{\pi}^\star_{h+1}}\right)\widehat{Q}_{h+1}^{\widehat{\pi}^\star} + P^{\widehat{\pi}^\star_{h+1}}\left(\widehat{Q}_{h+1}^{\widehat{\pi}^\star} - Q_{h+1}^{\widehat{\pi}^\star}\right) \\
&= \left(\widehat{P} - P\right)\widehat{V}_{h+1}^{\widehat{\pi}^\star} + P^{\widehat{\pi}^\star_{h+1}}\left(\widehat{Q}_{h+1}^{\widehat{\pi}^\star} - Q_{h+1}^{\widehat{\pi}^\star}\right) \\
&= \ldots = \sum_{t=h}^{H}\Gamma_{h+1:t}^{\widehat{\pi}^\star}\left(\widehat{P} - P\right)\widehat{V}_{t+1}^{\widehat{\pi}^\star}
\end{aligned}$$

where $\Gamma^{\pi}_{h+1:t} = \prod_{i=h+1}^{t} P^{\pi_i}$ is multi-step state-action transition and $\Gamma_{h+1:h} := I$.

Concentration on $\left(\widehat{P} - P\right)\widehat{V}^{\star}_{h}$. Now by (C.15), we have the following:

$$
\begin{aligned}
&\left(\widehat{P}_{s,a} - P_{s,a}\right)\widehat{V}^{\star}_{h} \\
&\leq C \cdot \left(\sqrt{\frac{2\mathrm{Var}_{s,a}(\widehat{V}^{\star}_{h})\iota_2}{N}} + 2\sqrt{\frac{\iota_2}{N \cdot HS}} + \frac{2H\iota_2}{3N} \right) \\
&\leq C \cdot \left(\sqrt{\frac{2\mathrm{Var}_{s,a}(V^{\widehat{\pi}^{\star}}_{h})\iota_2}{N}} + 2\sqrt{\frac{\iota_2}{N \cdot HS}} + \sqrt{\frac{2\iota_2}{N}} \cdot \left\| \widehat{V}^{\widehat{\pi}^{\star}}_{h} - V^{\widehat{\pi}^{\star}}_{h} \right\|_{\infty} + \frac{2H\iota_2}{3N} \right) \\
&\leq C \cdot \left(\sqrt{\frac{2\mathrm{Var}_{s,a}(V^{\widehat{\pi}^{\star}}_{h})\iota_2}{N}} + 2\sqrt{\frac{\iota_2}{N \cdot HS}} + \sqrt{\frac{2\iota_2}{N}} \cdot H^2\sqrt{\frac{S}{N}} + \frac{2H\iota_2}{3N} \right) \\
&\leq C' \cdot \left(\sqrt{\frac{2\mathrm{Var}_{s,a}(V^{\widehat{\pi}^{\star}}_{h})\iota_2}{N}} + 2\sqrt{\frac{\iota_2}{N \cdot HS}} + \frac{2H^2 S\log(HSA/\delta)}{N} \right),
\end{aligned}
\tag{C.16}
$$

where the third inequality uses Lemma D.0.11[5]. Then above implies

[5]Note the use of Lemma D.0.11 also works for any rewards since the only high probability result they used is for $||P - \hat{P}||_1$. Therefore conditional on the concentration for $||P - \hat{P}||_1$, the argument follows for any arbitrary reward as well.

$$
\begin{aligned}
&\widehat{Q}_h^{\widehat{\pi}^\star} - Q_h^{\widehat{\pi}^\star} \\
\leq & C' \sum_{t=h}^{H} \Gamma_{h+1:t}^{\widehat{\pi}^\star} \cdot \left(\sqrt{\frac{2\mathrm{Var}_{s,a}(V_h^{\widehat{\pi}^\star})\iota_2}{N}} + 2\sqrt{\frac{\iota_2}{N \cdot HS}} + \frac{2H^2 S \log(HSA/\delta)}{N} \right) \\
\leq & C' \left[\sum_{t=h}^{H} \Gamma_{h+1:t}^{\widehat{\pi}^\star} \cdot \sqrt{\frac{2\mathrm{Var}_{s,a}(V_h^{\widehat{\pi}^\star})\iota_2}{N}} + 2\sqrt{\frac{H \log(HSA/\delta)}{N}} + \frac{2H^3 S \log(HSA/\delta)}{N} \right] \\
\leq & C' \left[\sqrt{\frac{2H^3 S \log(HSA/\delta)}{N}} + 2\sqrt{\frac{H \log(HSA/\delta)}{N}} + \frac{2H^3 S \log(HSA/\delta)}{N} \right] \\
\leq & C'' \left[\sqrt{\frac{H^3 S \log(HSA/\delta)}{N}} + \frac{H^3 S \log(HSA/\delta)}{N} \right] \\
\leq & O\left[\sqrt{\frac{H^2 S \log(HSA/\delta)}{nd_m}} + \frac{H^2 S \log(HSA/\delta)}{nd_m} \right],
\end{aligned}
$$

where the third inequality uses Lemma D.0.8 and the last one uses $N \geq \frac{1}{2} nd_m$ with high probability. Similar result holds for $\widehat{Q}_h^{\pi^\star} - Q_h^{\pi^\star}$. Combing those results we have reward-free bound (for any reward simultaneously)

$$
O\left[\sqrt{\frac{H^2 S \log(HSA/\delta)}{nd_m}} + \frac{H^2 S \log(HSA/\delta)}{nd_m} \right],
$$

which finishes the proof of Theorem 4.7.2.

Remark 21. *Note above result is tight in both the dominant term AND the higher order term. Therefore this result cannot be further improved even in the higher order term.*

C.6 Discussion of Section 4.7

In this section we explain why Theorem 4.7.1 and Theorem 4.7.2 are optimal in the offline RL.

We begin with the offline task-agnostic setting. For the exquisite readers who check the proof of Theorem 5 of [84], the proving procedure of their lower bound follows the standard reduction to best-arm identification in multi-armed bandit problems. More specifically, to incorporate the dependence of $\log(K)$, they rely on the Theorem 10 of [84] (which is originated

from [116]) to show in order to be (ϵ, δ)-correct for a problem with A arms and with K tasks, it need at least $\Omega(\frac{A}{\epsilon^2}\log(\frac{K}{\delta}))$ samples. Such a result updates the Lemma G.1. in [8] by the extra factor $\log(K)$ for the bandit problem with K tasks. With no modification, the rest of the proof in Section E of [8] follows though and one can end up with the lower bound $\Omega(H^2\log(K)/d_m\epsilon^2)$ over the problem class $\mathcal{M}_{d_m} := \left\{(\mu, M) \mid \min_{t,s_t,a_t} d_t^{\mu}\left(s_t, a_t\right) \geq d_m\right\}$. The case for the offline reward-free setting is also similar. Indeed, the $\Omega(SA/\epsilon^2)$ trajectories in Lemma 4.2 in [80] could be replaced by $\Omega(1/d_m\epsilon^2)$ by choosing some hard *near-uniform* behavior policy instances (see Section E.2 in [8]) and the rest follows since by forcing S such instances (Section 4.2 of [80]) to obtain $\Omega(S/d_m\epsilon^2)$ and create a chain of $\Omega(H)$ rewards for $\Omega(H^2S/d_m\epsilon^2)$.

C.7 Proof of the linear MDP with anchor representations

Recall that we assume a generative oracle here. Sometimes we abuse the notation $\mathcal{K}$ for either anchor point set or the anchor point indices set. The meaning should be clear in each context.

C.7.1 Model-based Plug-in Estimator for Anchor Representations

Step 1: For each (s_k, a_k) where index $k \in \mathcal{K}$, collect N samples from $P(\cdot|s_k, a_k)$; compute

$$\widehat{P}_{\mathcal{K}}(s'|s_k, a_k) = \frac{count(s, a, s')}{N};$$

Step 2: Compute the linear combination coefficients $\lambda_k^{s,a}$ satisfies $\phi(s, a) = \sum_{k\in\mathcal{K}} \lambda_k^{s,a}\phi(s_k, a_k)$;

Step 3: Estimate transition distribution

$$\widehat{P}(s'|s, a) = \sum_{k\in\mathcal{K}} \lambda_k^{s,a} \cdot \widehat{P}_{\mathcal{K}}(s'|s_k, a_k).$$

We need to check such $\widehat{P}(s'|s, a)$ is a valid distribution. This is due to:

$$\begin{aligned}\sum_{k\in\mathcal{K}} \lambda_k^{s,a} &= \sum_{k\in\mathcal{K}}\sum_{s'} \lambda_k^{s,a} P(s'|s_k, a_k) = \sum_{s'}\sum_{k\in\mathcal{K}} \lambda_k^{s,a} P(s'|s_k, a_k)\\ &= \sum_{s'}\sum_{k\in\mathcal{K}} \lambda_k^{s,a}\langle\phi(s_k, a_k), \psi(s')\rangle = \sum_{s'}\langle\phi(s, a), \psi(s')\rangle = \sum_{s'} P(s'|s, a) = 1\end{aligned}$$

and

$$\sum_{s'} \widehat{P}(s'|s,a) = \sum_{s'} \sum_{k\in\mathcal{K}} \lambda_k^{s,a} \widehat{P}_{\mathcal{K}}\left(s' \mid s_k, a_k\right) = \sum_{k\in\mathcal{K}} \sum_{s'} \lambda_k^{s,a} \widehat{P}_{\mathcal{K}}\left(s' \mid s_k, a_k\right)$$
$$= \sum_{k\in\mathcal{K}} \lambda_k^{s,a} \frac{N}{N} = 1.$$

Step 4: construct empirical model $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, r, H)$ and output $\widehat{\pi}^\star = \operatorname{argmax}_\pi \widehat{V}_1^\pi$. Similarly, Bellman (optimality) equations hold[6]

$$V_t^\star(s) = \max_a \left\{ r(s,a) + \int_{s'} V_{t+1}^\star(s') d\, P(s'|s,a) \right\}, \quad \forall s \in \mathcal{S}.$$
$$\widehat{V}_t^\star(s) = \max_a \left\{ r(s,a) + \int_{s'} \widehat{V}_{t+1}^\star(s') d\, \widehat{P}(s'|s,a) \right\}, \quad \forall s \in \mathcal{S}.$$

C.7.2 General absorbing MDP

The definition of the general absorbing MDP remains the same: *i.e.* for a fixed state s and a sequence $\{u_t\}_{t=1}^H$, MDP $M_{s,\{u_t\}_{t=1}^H}$ is identical to M for all states except s, and state s is absorbing in the sense $P_{M_{s,\{u_t\}_{t=1}^H}}(s|s,a) = 1$ for all a, and the instantaneous reward at time t is $r_t(s,a) = u_t$ for all $a \in \mathcal{A}$. Also, we use the shorthand notation $V^\pi_{\{s,u_t\}}$ for $V^\pi_{s,M_{s,\{u_t\}_{t=1}^H}}$ and similarly for $Q_{\{s,u_t\}}$ and transition $P_{\{s,u_t\}}$. Then the following properties mirroring the Lemma C.1.1 and Lemma C.1.2 with nearly identical proof but for the integral version (which we skip):

Lemma C.7.1.

$$V^\star_{h,\{s,u_t\}}(s) = \sum_{t=h}^{H} u_t.$$

Lemma C.7.2. *Fix state s. For two different sequences $\{u_t\}_{t=1}^H$ and $\{u'_t\}_{t=1}^H$, we have*

$$\max_h \left\| Q^\star_{h,\{s,u_t\}} - Q^\star_{h,\{s,u'_t\}} \right\|_\infty \le H \cdot \max_{t\in[H]} \left| u_t - u'_t \right|.$$

C.7.3 Singleton-absorbing MDP

The well-definedness of singleton-absorbing MDP for linear MDP with anchor points depends on the following two lemmas whose proofs are still nearly identical to Lemma C.1.3 and Lemma C.1.4 which we skip.

[6] We use the integral only to denote $\mathcal{S}$ could be exponentially large.

Lemma C.7.3. *$V_t^\star(s) - V_{t+1}^\star(s) \geq 0$, for all state $s \in \mathcal{S}$ and all $t \in [H]$.*

Lemma C.7.4. *Fix a state s. If we choose $u_t^\star := V_t^\star(s) - V_{t+1}^\star(s)\ \forall t \in [H]$, then we have the following vector form equation*

$$V^\star_{h,\{s,u_t^\star\}} = V^\star_{h,M} \quad \forall h \in [H].$$

Similarly, if we choose $\widehat{u}_t^\star := \widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s)$, then $\widehat{V}^\star_{h,\{s,\widehat{u}_t^\star\}} = \widehat{V}^\star_{h,M}$, $\forall h \in [H]$.

The singleton MDP we used is exactly $M_{s,\{u_t^\star\}_{t=1}^H}$ (or $\widehat{M}_{s,\{u_t^\star\}_{t=1}^H}$).

C.7.4 Proof for the optimal sample complexity

For $\widehat{\pi}^\star$, by (empirical) Bellman equation we have element-wisely:

$$\begin{aligned}
\widehat{Q}_h^{\widehat{\pi}^\star} - Q_h^{\widehat{\pi}^\star} &= r_h + \widehat{P}^{\widehat{\pi}^\star_{h+1}} \widehat{Q}_{h+1}^{\widehat{\pi}^\star} - r_h - P^{\widehat{\pi}^\star_{h+1}} Q_{h+1}^{\widehat{\pi}^\star} \\
&= \left(\widehat{P}^{\widehat{\pi}^\star_{h+1}} - P^{\widehat{\pi}^\star_{h+1}}\right) \widehat{Q}_{h+1}^{\widehat{\pi}^\star} + P^{\widehat{\pi}^\star_{h+1}} \left(\widehat{Q}_{h+1}^{\widehat{\pi}^\star} - Q_{h+1}^{\widehat{\pi}^\star}\right) \\
&= \left(\widehat{P} - P\right) \widehat{V}_{h+1}^{\widehat{\pi}^\star} + P^{\widehat{\pi}^\star_{h+1}} \left(\widehat{Q}_{h+1}^{\widehat{\pi}^\star} - Q_{h+1}^{\widehat{\pi}^\star}\right) \\
&= \ldots = \sum_{t=h}^{H} \Gamma^{\widehat{\pi}^\star}_{h+1:t} \left(\widehat{P} - P\right) \widehat{V}_{t+1}^{\widehat{\pi}^\star} \leq \underbrace{\sum_{t=h}^{H} \Gamma^{\widehat{\pi}^\star}_{h+1:t} \left|\left(\widehat{P} - P\right) \widehat{V}_{t+1}^{\widehat{\pi}^\star}\right|}_{(\star)}
\end{aligned}$$

where $\Gamma^{\pi^\star}_{h+1:t} = \prod_{i=h+1}^{t} P^{\pi_i^\star}$ is multi-step state-action transition and $\Gamma_{h+1:h} := I$.

C.7.5 Analyzing ($\star$)

Concentration on $\left(\widehat{P} - P\right) \widehat{V}_h^\star$. Since $\widehat{P}$ aggregates all data from different step so that $\widehat{P}$ and $\widehat{V}_h^\star$ are on longer independent. We use the singleton-absorbing MDP $M_{s,\{u_t^\star\}_{t=1}^H}$ to handle the case (recall $u_t^\star := V_t^\star(s) - V_{t+1}^\star(s)\ \forall t \in [H]$). **Here, we fix the state action $(s,a) \in \mathcal{K}$.**

Then we have:

$$
\begin{aligned}
&\left(\widehat{P}_{s,a}-P_{s,a}\right)\widehat{V}_h^\star=\left(\widehat{P}_{s,a}-P_{s,a}\right)\left(\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}+\widehat{V}^\star_{h,\{s,u_t^\star\}}\right)\\
&=\left(\widehat{P}_{s,a}-P_{s,a}\right)\left(\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right)+\left(\widehat{P}_{s,a}-P_{s,a}\right)\widehat{V}^\star_{h,\{s,u_t^\star\}}\\
&\le\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1\left\|\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty+\sqrt{\frac{2\log(4/\delta)}{N}}\sqrt{\mathrm{Var}_{s,a}(\widehat{V}^\star_{h,\{s,u_t^\star\}})}+\frac{2H\log(1/\delta)}{3N}\\
&\le\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1\left\|\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty+\sqrt{\frac{2\log(4/\delta)}{N}}\left(\sqrt{\mathrm{Var}_{s,a}(\widehat{V}_h^\star)}+\sqrt{\mathrm{Var}_{s,a}(\widehat{V}^\star_{h,\{s,u_t^\star\}}-\widehat{V}_h^\star)}\right)+\frac{2H\log(1/\delta)}{3N}\\
&\le\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1\left\|\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty+\sqrt{\frac{2\log(4/\delta)}{N}}\left(\sqrt{\mathrm{Var}_{s,a}(\widehat{V}_h^\star)}+\sqrt{\left\|\widehat{V}^\star_{h,\{s,u_t^\star\}}-\widehat{V}_h^\star\right\|_\infty^2}\right)+\frac{2H\log(1/\delta)}{3N}\\
&=\left(\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1+\sqrt{\frac{2\log(4/\delta)}{N}}\right)\left\|\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty+\sqrt{\frac{2\log(4/\delta)}{N}}\sqrt{\mathrm{Var}_{s,a}(\widehat{V}_h^\star)}+\frac{2H\log(1/\delta)}{3N}
\end{aligned}
\tag{C.17}
$$

where the first inequality uses Bernstein inequality (Lemma D.0.3) (**note here** $P_{s,a}V=\int_{s'}V(s')dP(s'|s,a)$ **since** $\mathcal{S}$ **could be continuous space, but this does not affect the availability of Bernstein inequality!**), the second inequality uses $\sqrt{\mathrm{Var}(\cdot)}$ is norm (norm triangle inequality). Now we treat $\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1$ and $\left\|\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty$ separately.

For $\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1$**.** Recall here $(s,a)\in\mathcal{K}$. By Lemma D.0.10 we obtain w.p. $1-\delta$

$$
\left\|\widehat{P}_{s,a}-P_{s,a}\right\|_1\le C\sqrt{\frac{|\mathcal{S}|\log(1/\delta)}{N}}. \tag{C.18}
$$

where C absorbs the higher order term and constants.

For $\left\|\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty$**.** Note if we set $\hat{u}_t^\star=\widehat{V}_t^\star(s)-\widehat{V}_{t+1}^\star(s)$, then by Lemma C.7.4

$$
\widehat{V}_h^\star=\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}}
$$

Next since $\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}}(\tilde{s})=\max_a\widehat{Q}^\star_{h,\{s,\hat{u}_t^\star\}}(\tilde{s},a)\ \forall\tilde{s}\in\mathcal{S}$, by generic inequality $|\max f-\max g|\le\max|f-g|$, we have $|\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}}(\tilde{s})-\widehat{V}^\star_{h,\{s,u_t^\star\}}(\tilde{s})|\le\max_a|\widehat{Q}^\star_{h,\{s,\hat{u}_t^\star\}}(\tilde{s},a)-\widehat{Q}^\star_{h,\{s,u_t^\star\}}(\tilde{s},a)|$, taking $\max_{\tilde{s}}$ on both sides, we obtain exactly

$$
\left\|\widehat{V}^\star_{h,\{s,\hat{u}_t^\star\}}-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty\le\left\|\widehat{Q}^\star_{h,\{s,\hat{u}_t^\star\}}-\widehat{Q}^\star_{h,\{s,u_t^\star\}}\right\|_\infty
$$

then by Lemma C.7.2,

$$
\left\|\widehat{V}_h^\star-\widehat{V}^\star_{h,\{s,u_t^\star\}}\right\|_\infty\le\left\|\widehat{Q}^\star_{h,\{s,\hat{u}_t^\star\}}-\widehat{Q}^\star_{h,\{s,u_t^\star\}}\right\|_\infty\le H\max_t\left|\hat{u}_t^\star-u_t^\star\right|, \tag{C.19}
$$

Recall

$$\hat{u}_t^\star - u_t^\star = \widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s) - \left(V_t^\star(s) - V_{t+1}^\star(s)\right).$$

Now we denote

$$\Delta_s := \max_t |\hat{u}_t^\star - u_t^\star| = \max_t \left|\widehat{V}_t^\star(s) - \widehat{V}_{t+1}^\star(s) - \left(V_t^\star(s) - V_{t+1}^\star(s)\right)\right|,$$

then Δ_s itself is a scalar and a random variable.

To sum up, by (C.17), (C.3) and (C.19) and a union bound over all $(s,a) \in \mathcal{K}$ we have

Lemma C.7.5. *Fix $N > 0$. With probability $1-\delta$, element-wisely, for all $h \in [H]$ and all $(s_k, a_k) \in \mathcal{K}$,*

$$\begin{aligned}\left|\left(\widehat{P}_{s_k,a_k} - P_{s_k,a_k}\right)\widehat{V}_h^\star\right| \leq & C\sqrt{\frac{|S|\log(HK/\delta)}{N}} \cdot H \max_{s_k} \Delta_{s_k} \\ & + \sqrt{\frac{2\log(4HK/\delta)}{N}}\sqrt{\mathrm{Var}_{P_{s_k,a_k}}(\widehat{V}_h^\star)} + \frac{2H\log(HK/\delta)}{3N}\end{aligned}$$

Now we extend Lemma C.7.5 to any arbitrary (s,a) by proving the following lemma:

Lemma C.7.6 (recover lemma)**.** *For any function V and any state action (s,a), we have*

$$\sum_{k\in\mathcal{K}} \lambda_k^{s,a}\sqrt{\mathrm{Var}_{P_{s_k,a_k}}(V)} \leq \sqrt{\mathrm{Var}_{P_{s,a}}(V)}$$

Proof: [Proof of Lemma C.7.6] Since $\lambda_k^{s,a}$ are probability distributions, by Jensen's inequality twice

$$\begin{aligned}&\sum_{k\in\mathcal{K}} \lambda_k^{s,a}\sqrt{\mathrm{Var}_{P_{s_k,a_k}}(V)} \leq \sqrt{\sum_{k\in\mathcal{K}} \lambda_k^{s,a}\mathrm{Var}_{P_{s_k,a_k}}(V)} \\ =&\sqrt{\sum_{k\in\mathcal{K}} \lambda_k^{s,a}\mathrm{Var}_{P_{s_k,a_k}}(V)} = \sqrt{\sum_{k\in\mathcal{K}} \lambda_k^{s,a}(P_{s_k,a_k}V^2 - (P_{s_k,a_k}V)^2)} \\ \leq&\sqrt{\sum_{k\in\mathcal{K}} \lambda_k^{s,a}\cdot P_{s_k,a_k}V^2 - (\sum_{k\in\mathcal{K}} \lambda_k^{s,a}P_{s_k,a_k}V)^2} \\ =&\sqrt{P_{s,a}V^2 - (P_{s,a}V)^2} = \sqrt{\mathrm{Var}_{P_{s,a}}(V)},\end{aligned}$$

where we use $P_{s,a} = \sum_{k\in\mathcal{K}} \lambda_k^{s,a}P_{s_k,a_k}$.

Therefore for all (s,a), using Lemma C.7.5 and Lemma C.7.6 we obtain w.p. $1-\delta$,

$$\begin{aligned}
&\left|\left(\widehat{P}_{s,a}-P_{s,a}\right)\widehat{V}_h^\star\right| \le \sum_{k\in\mathcal{K}}\lambda_k^{s,a}\left|\left(\widehat{P}_{s_k,a_k}-P_{s_k,a_k}\right)\widehat{V}_h^\star\right| \\
&\le C\sum_{k\in\mathcal{K}}\lambda_k^{s,a}\sqrt{\frac{S\log(HK/\delta)}{N}}\cdot H\max_{s_k}\Delta_{s_k}+\sum_{k\in\mathcal{K}}\lambda_k^{s,a}\sqrt{\frac{2\log(4HK/\delta)}{N}}\sqrt{\mathrm{Var}_{P_{s_k,a_k}}(\widehat{V}_h^\star)} \\
&+\sum_{k\in\mathcal{K}}\lambda_k^{s,a}\frac{2H\log(HK/\delta)}{3N} \\
&= C\sqrt{\frac{S\log(HK/\delta)}{N}}\cdot H\max_{s_k}\Delta_{s_k}+\sum_{k\in\mathcal{K}}\lambda_k^{s,a}\sqrt{\frac{2\log(4HK/\delta)}{N}}\sqrt{\mathrm{Var}_{P_{s_k,a_k}}(\widehat{V}_h^\star)} \\
&+\frac{2H\log(HK/\delta)}{3N} \\
&\le C\sqrt{\frac{S\log(HK/\delta)}{N}}\cdot H\max_{s_k}\Delta_{s_k}+\sqrt{\frac{2\log(4HK/\delta)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}(\widehat{V}_h^\star)}+\frac{2H\log(HK/\delta)}{3N}
\end{aligned}$$

Now plug above back into $(\star)$, we receive:

$$\begin{aligned}
&\left|\widehat{Q}_h^{\widehat{\pi}^\star}-Q_h^{\widehat{\pi}^\star}\right| \\
&\le\sum_{t=h}^{H}\Gamma_{h+1:t}^{\widehat{\pi}^\star}\left(C\sqrt{\frac{S\log(HK/\delta)}{N}}\cdot H\max_{s_k}\Delta_{s_k}\cdot\mathbf{1}+\sqrt{\frac{2\log(4HK/\delta)}{N}}\sqrt{\mathrm{Var}_P(\widehat{V}_{t+1}^\star)}+\frac{2H\log(HK/\delta)}{3N}\cdot\mathbf{1}\right) \\
&\le\sum_{t=h}^{H}\Gamma_{h+1:t}^{\widehat{\pi}}\sqrt{\frac{2\log(4HK/\delta)}{N}}\sqrt{\mathrm{Var}_P(\widehat{V}_{t+1}^\star)}+CH^2\sqrt{\frac{S\log(HK/\delta)}{N}}\cdot\max_s\Delta_s\cdot\mathbf{1}+\frac{2H^2\log(HK/\delta)}{3N}\cdot\mathbf{1}
\end{aligned}$$

Similar to before, we get

$$\sqrt{\mathrm{Var}_P(\widehat{V}_h^\star)}:=\sqrt{\mathrm{Var}_P\left(\widehat{V}_h^{\widehat{\pi}^\star}\right)}\le\sqrt{\mathrm{Var}_P\left(V_h^{\widehat{\pi}^\star}\right)}+\left\|\widehat{Q}_h^{\widehat{\pi}^\star}-Q_h^{\widehat{\pi}^\star}\right\|_\infty \tag{C.20}$$

Plug (C.20) back to above we obtain $\forall h \in [H]$,

$$\begin{aligned}
&\left|\widehat{Q}_h^{\widehat{\pi}^\star} - Q_h^{\widehat{\pi}^\star}\right| \leq \sqrt{\frac{2\log(4HK/\delta)}{N}} \sum_{t=h}^{H} \Gamma_{h+1:t}^{\widehat{\pi}^\star} \left(\sqrt{\mathrm{Var}_P\left(V_{t+1}^{\widehat{\pi}^\star}\right)} + \left\|\widehat{Q}_{t+1}^{\widehat{\pi}^\star} - Q_{t+1}^{\widehat{\pi}^\star}\right\|_\infty \right) \\
&+ CH^2 \sqrt{\frac{S\log(HK/\delta)}{N}} \cdot \max_{s_k} \Delta_{s_k} \cdot \mathbf{1} + \frac{2H^2\log(HK/\delta)}{3N} \cdot \mathbf{1} \\
&\leq \sqrt{\frac{2\log(4HK/\delta)}{N}} \sum_{t=h}^{H} \Gamma_{h+1:t}^{\widehat{\pi}^\star} \sqrt{\mathrm{Var}_P\left(V_{t+1}^{\widehat{\pi}^\star}\right)} + \sqrt{\frac{2\log(4HK/\delta)}{N}} \sum_{t=h}^{H} \left\|\widehat{Q}_{t+1}^{\widehat{\pi}^\star} - Q_{t+1}^{\widehat{\pi}^\star}\right\|_\infty \\
&+ CH^2 \sqrt{\frac{S\log(HK/\delta)}{N}} \cdot \max_{s_k} \Delta_{s_k} \cdot \mathbf{1} + \frac{2H^2\log(HK/\delta)}{3N} \cdot \mathbf{1}
\end{aligned} \tag{C.21}$$

Apply Lemma D.0.8 and the (anchor version using recover lemma C.7.6) coarse uniform bound (Lemma D.0.11) we obtain the following lemma:

Lemma C.7.7. *With probability $1-\delta$, for all $h \in [H]$,*

$$\left\|\widehat{Q}_h^{\widehat{\pi}^\star} - Q_h^{\widehat{\pi}^\star}\right\|_\infty \leq \sqrt{\frac{C_0 H^3 \log(4HK/\delta)}{N}} + \sqrt{\frac{2\log(4HK/\delta)}{N}} \sum_{t=h}^{H} \left\|\widehat{Q}_{t+1}^{\widehat{\pi}^\star} - Q_{t+1}^{\widehat{\pi}^\star}\right\|_\infty + C' H^4 \frac{S\log(HK/\delta)}{N}$$

Proof: Since

$$\begin{aligned}
\Delta_{s_k} &:= \max_t |\hat{u}_t^\star - u_t^\star| = \max_t \left| \widehat{V}_t^\star(s_k) - \widehat{V}_{t+1}^\star(s_k) - \left(V_t^\star(s_k) - V_{t+1}^\star(s_k)\right) \right| \\
&\leq 2 \cdot \max_t \left| \widehat{V}_t^\star(s_k) - V_t^\star(s_k) \right| \\
&= 2 \cdot \max_t \left| \max_\pi \widehat{V}_t^\pi(s_k) - \max_\pi V_t^\pi(s_k) \right| \\
&\leq 2 \cdot \max_{\pi \in \Pi_g, t \in [H]} \left\| \widehat{V}_t^\pi - V_t^\pi \right\|_\infty \leq C \cdot H^2 \sqrt{\frac{|\mathcal{S}|\log(HK/\delta)}{N}}
\end{aligned} \tag{C.22}$$

where the last inequality uses (the anchor version) of Lemma D.0.11.[7] Then apply union bound w.p. $1-\delta/2$, we obtain $\max_{s_k} \Delta_{s_k} \leq C \cdot H^2 \sqrt{\frac{|\mathcal{S}|\log(HK^2/\delta)}{N}}$. Note (C.21) holds with probability $1-\delta/2$, therefore plug above into (C.21) and uses Lemma D.0.8 and take $||\cdot||_\infty$ we obtain w.p. $1-\delta$, the result holds.

Lemma C.7.8. *Given $N > 0$. Define $C'' := 2 \cdot \max(\sqrt{C_0}, C')$ where C' is the universal constant in Lemma C.7.7. When $N \geq 8H^2|S|\log(4HK/\delta)$, then with probability $1-\delta$, $\forall h \in$*

[7] Here the anchor version means for any (s,a) we can apply $||\widehat{P}_{s,a} - P_{s,a}||_1 = ||\sum_k \lambda_k^{s,a}(\widehat{P}_{s,a} - P_{s,a})||_1 \leq \sum_k \lambda_k^{s,a}||\widehat{P}_{s,a} - P_{s,a}||_1$.

$[H]$,

$$\begin{aligned}\left\|\widehat{Q}_h^{\hat{\pi}^\star} - Q_h^{\hat{\pi}^\star}\right\|_\infty &\le C''\sqrt{\frac{H^3\log(4HK/\delta)}{N}} + C''\frac{H^4 S\log(HK/\delta)}{N}.\\ \left\|\widehat{Q}_h^{\pi^\star} - Q_h^{\pi^\star}\right\|_\infty &\le C''\sqrt{\frac{H^3\log(4HK/\delta)}{N}} + C''\frac{H^4 S\log(HK/\delta)}{N}.\end{aligned} \tag{C.23}$$

Proof: The proof is very similar to that of Lemma C.1.8.

Remark 22. *Note the higher order term has dependence of the order of H^4S. However, using the same* self-bounding *trick, we can reduce it to $H^{3.5}S^{0.5}$.*

Lemma C.7.9. *Given $N > 0$. There exists universal constants C_1, C_2 such that when $N \ge C_1 H^2|S|\log(HK/\delta)$, then with probability $1-\delta$, $\forall h \in [H]$,*

$$\left\|\widehat{Q}_h^{\hat{\pi}^\star} - Q_h^{\hat{\pi}^\star}\right\|_\infty \le C_2\sqrt{\frac{H^3\log(HK/\delta)}{N}} + C_2\frac{H^3\sqrt{HS}\log(HK/\delta)}{N}. \tag{C.24}$$

and

$$\left\|\widehat{Q}_h^{\pi^\star} - Q_h^{\pi^\star}\right\|_\infty \le C_2\sqrt{\frac{H^3\log(HK/\delta)}{N}} + C_2\frac{H^3\sqrt{HS}\log(HK/\delta)}{N}.$$

Proof: The proof is similar to Lemma C.1.9.

C.7.6 Proof of Theorem 4.8.1

Proof: By the direct computing of the suboptimality,

$$Q_1^\star - Q_1^{\hat{\pi}^\star} = Q_1^\star - \widehat{Q}_1^{\pi^\star} + \widehat{Q}_1^{\pi^\star} - \widehat{Q}_1^{\hat{\pi}^\star} + \widehat{Q}_1^{\hat{\pi}^\star} - Q_1^{\hat{\pi}^\star} \le |Q_1^\star - \widehat{Q}_1^{\pi^\star}| + |\widehat{Q}_1^{\hat{\pi}^\star} - Q_1^{\hat{\pi}^\star}|,$$

then by Lemma C.7.9 we can finish the proof.

C.7.7 Take-away in the linear MDP with anchor setting.

Under the setting S could be exponential large, $\mathcal{A}$ could be infinite (or even continuous space), with anchor representations ($K \ll |S|$), our Theorem 4.8.1 has order $\widetilde{O}(\sqrt{H^3/N})$ when N is sufficiently large. This translate to $N = \widetilde{O}(H^3/\epsilon^2)$ and the total sample used is $KN = \widetilde{O}(KH^3/\epsilon^2)$. This improves the total complexity $\widetilde{O}(KH^4/\epsilon^2)$ in [68] and is optimal.

C.8 The computational efficiency for the model-based offline plug-in estimators

For completeness, we discuss the computational and storage aspect of our model-based method. Its computational cost is $\widetilde{O}(H^4/d_m\epsilon^2)$ for computing $\widehat{P}$, the same as its sample com-

plexity in steps (H steps is an episode), and running value iteration causes $O(HS^2A)$ time (here we assume the bit complexity $L(P, r, H) = 1$, see [117] Section 1.3). The total computational complexity is $\widetilde{O}(H^4/d_m\epsilon^2) + O(HS^2A)$, with a memory cost of $O(HS^2A)$.

Appendix D

Some Technical Lemmas

Lemma D.0.1 (Multiplicative Chernoff bound [118]). *Let X be a Binomial random variable with parameter p, n. For any $\delta > 0$, we have that*

$$\mathbb{P}[X < (1-\delta)pn] < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{np}.$$

A slightly looser bound that suffices for our propose is the following:

$$\mathbb{P}[X < (1-\delta)pn] < e^{-\frac{\delta^2 pn}{2}}.$$

Lemma D.0.2 (Hoeffding's Inequality [119]). *Let $x_1, ..., x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \leq \xi_i$ with probability 1. Then for any $\epsilon > 0$ we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \geq \epsilon\right) \leq e^{-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}\xi_i^2}}.$$

Lemma D.0.3 (Bernstein's Inequality). *Let $x_1, ..., x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \leq \xi$ with probability 1. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[x_i]$, then with probability $1-\delta$ we have*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \sqrt{\frac{2\sigma^2 \cdot \log(1/\delta)}{n}} + \frac{2\xi}{3n}\log(1/\delta)$$

Lemma D.0.4 (Mcdiarmid's Inequality: [119]). *Let $x_1, ..., x_n$ be independent random variables and $S : X^n \to \mathbb{R}$ be a measurable function which is invariant under permutation and let the random variable Z be given by $Z = S(x_1, x_2, ..., x_n)$. Assume S has bounded difference: i.e.*

$$\sup_{x_1,...,x_n,x_i'} |S(x_1, ..., x_i, ..., x_n) - S(x_1, ..., x_i', ..., x_n)| \leq \xi_i,$$

then for any $\epsilon > 0$ we have

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \epsilon) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} \xi_i^2}}.$$

Lemma D.0.5 (Azuma-Hoeffding inequality)**.** *Suppose X_k, $k = 1, 2, 3, ...$ is a martingale and $|X_k - X_{k-1}| \leq c_k$ almost surely. Then for all positive integers N and any $\epsilon > 0$,*

$$\mathbb{P}(|X_N - X_0| \geq \epsilon) \leq 2e^{-\frac{\epsilon^2}{2\sum_{i=1}^{N} c_i^2}}.$$

Lemma D.0.6 (Freedman's inequality [120])**.** *Let X be the martingale associated with a filter $\mathcal{F}$ (i.e. $X_i = \mathbb{E}[X|\mathcal{F}_i]$) satisfying $|X_i - X_{i-1}| \leq M$ for $i = 1, ..., n$. Denote $W := \sum_{i=1}^{n} \mathrm{Var}(X_i|\mathcal{F}_{i-1})$ then we have*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon, W \leq \sigma^2) \leq 2e^{-\frac{\epsilon^2}{2(\sigma^2 + M\epsilon/3)}}.$$

Or in other words, with probability $1 - \delta$,

$$|X - \mathbb{E}[X]| \leq \sqrt{8\sigma^2 \cdot \log(1/\delta)} + \frac{2M}{3} \cdot \log(1/\delta), \quad Or \quad W \geq \sigma^2.$$

Lemma D.0.7 (Best arm identification lower bound [115])**.** *For any $A \geq 2$ and $\tau \leq \sqrt{1/8}$ and any best arm identification algorithm that produces an estimate $\hat{a}$, there exists a multi-arm bandit problem for which the best arm $a^\star$ is τ better than all others, but $\mathbb{P}[\hat{a} \neq a^\star] \geq 1/3$ unless the number of samples T is at least $\frac{A}{72\tau^2}$.*

Lemma D.0.8 (Sum of expectation of conditional variance of value; Lemma F.3 of [7])**.**

$$\begin{aligned}
&\mathrm{Var}_\pi \left[\sum_{t=h}^{H} r_t^{(1)} \mid s_h^{(1)} = s_h, a_h^{(1)} = a_h \right] \\
&= \sum_{t=h}^{H} \Bigg(\mathbb{E}_\pi \left[\mathrm{Var} \left[r_t^{(1)} + V_{t+1}^{\pi} \left(s_{t+1}^{(1)} \right) \mid s_t^{(1)}, a_t^{(1)} \right] \mid s_h^{(1)} = s_h, a_h^{(1)} = a_h \right] \\
&+ \mathbb{E}_\pi \left[\mathrm{Var} \left[\mathbb{E} \left[r_t^{(1)} + V_{t+1}^{\pi} \left(s_{t+1}^{(1)} \right) \mid s_t^{(1)}, a_t^{(1)} \right] \mid s_t^{(1)} \right] \mid s_h^{(1)} = s_h, a_h^{(1)} = a_h \right] \Bigg)
\end{aligned}$$

By apply above, one can show

$$\sum_{t=h}^{H} \Gamma_{h+1:t}^{\pi} \sqrt{\mathrm{Var}_P \left(V_{t+1}^{\pi} \right)} \leq \sqrt{(H-h)^3} \cdot \mathbf{1}.$$

Remark 23. *The infinite horizon discounted setting counterpart result is $(I - \gamma P^\pi)^{-1} \sigma_{V^\pi} \leq (1-\gamma)^{-3/2}$.*

D.0.1 Minimax rate of discrete distributions under l_1 loss.

This Section provides the minimax rate for $\left\|\widehat{P} - P\right\|_1$ for any model-based algorithms and is based on [89]. Let P be S dimensional distribution.

Lemma D.0.9 (Minimax lower bound for $\left\|\widehat{P} - P\right\|_1$)**.** *Let n be the number of data-points sampled from P. If $n > \frac{e}{32}S$, then there exists a constant $p > 0$, such that*

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{P}\left[\left\|\widehat{P} - P\right\|_1 \geq \frac{1}{8}\sqrt{\frac{eS}{2n}} - o(e^{-n}) - o(e^{-S})\right] \geq p,$$

*where $\mathcal{M}_S$ denotes the set of distributions with support size S and the infimum is taken over **ALL** estimators.*

Remark 24. *Note the $\widehat{P}$ in above carries over all estimators but not just empirical estimator. This provides the minimax result.*

Proof: The proof comes from Theorem 2 of [89], where we pick $\zeta = 1$. Note they establish the minimax result for $\mathbb{E}_P\|\hat{P} - P\|_1$. However, by a simple contradiction we can get the above. Indeed, suppose

$$\inf_{\widehat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{P}\left[\left\|\widehat{P} - P\right\|_1 < \frac{1}{8}\sqrt{\frac{eS}{2n}} - o(e^{-n}) - o(e^{-S})\right] = 1,$$

then this implies $\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P\|\hat{P} - P\|_1 < \frac{1}{8}\sqrt{\frac{eS}{2n}} - o(e^{-n}) - o(e^{-S})$ which contradicts Theorem 2 of [89].

Lemma D.0.10 (Upper bound for $\left\|\widehat{P} - P\right\|_1$)**.** *Let n be the number of data-points sampled from P. Then with probability $1 - \delta$*

$$\left\|\widehat{P} - P\right\|_1 \leq C\left(\sqrt{\frac{S\log(S/\delta)}{n}} + \frac{S\log(S/\delta)}{n}\right)$$

for any $P \in \mathcal{M}_S$. Here $\widehat{P}$ is the empirical (MLE) estimator.

Proof: First fix a state s. Let $X_i = \mathbf{1}[s_i = s]$, then $X_i \sim Bern(p_s(1 - p_s))$ and $X_s = \sum_{i=1}^n X_i \sim Binomial(n, p_i)$. By Bernstein inequality,

$$\left|\frac{X_s}{n} - P_s\right| \leq \sqrt{\frac{2p_s(1 - p_s)\log(1/\delta)}{n}} + \frac{3}{n}\log(1/\delta)$$

Apply a union bound we obtain w.p. $1-\delta$

$$\left|\frac{X_s}{n}-P_s\right| \le \sqrt{\frac{2p_s(1-p_s)\log(S/\delta)}{n}}+\frac{3}{n}\log(S/\delta) \quad \forall s\in\mathcal{S}$$

which implies

$$\begin{aligned}
\left\|\widehat{P}-P\right\|_1 &= \sum_{s\in\mathcal{S}}\left|\frac{X_s}{n}-P_s\right| \\
&\le \sum_{s\in\mathcal{S}}\sqrt{\frac{2p_s(1-p_s)\log(S/\delta)}{n}}+\frac{3S}{n}\log(S/\delta) \\
&= \sqrt{\frac{1}{n}}\sum_{s\in\mathcal{S}}\frac{1}{S}\cdot\sqrt{2S^2p_s(1-p_s)\log(S/\delta)}+\frac{3S}{n}\log(S/\delta) \\
&\le \sqrt{\frac{1}{n}}\sqrt{2S^2\cdot\frac{\sum_{s\in\mathcal{S}}p_s}{S}\left(1-\frac{\sum_{s\in\mathcal{S}}p_s}{S}\right)\log(S/\delta)}+\frac{3S}{n}\log(S/\delta) \\
&= \sqrt{\frac{2(S-1)\log(S/\delta)}{n}}+\frac{3S}{n}\log(S/\delta).
\end{aligned}$$

where the last inequality uses the concavity of $\sqrt{x(1-x)}$.

Finally, we can absorb the higher order term using the mild condition $n > c\cdot S\log(S/\delta)$.

D.0.2 A crude uniform convergence bound

Here we provide a crude bound for $\sup_{\pi\in\Pi_g}\left\|\widehat{V}_1^\pi-V_1^\pi\right\|_\infty$, which is the finite horizon counterpart of Section 2.2 of [41] and is a form of simulation lemma.

Lemma D.0.11 (Crude bound by Simulation Lemma)**.** *Fix $N>0$ to be number of samples for each coordinates. Recall Π_g is the global policy class. Then w.p. $1-\delta$,*

$$\sup_{\pi\in\Pi_g,h\in[H]}\left\|\widehat{Q}_h^\pi-Q_h^\pi\right\|_\infty \le C\cdot H^2\sqrt{\frac{S\log(SA/\delta)}{N}},$$

which further implies

$$\sup_{\pi\in\Pi_g,h\in[H]}\left\|\widehat{V}_h^\pi-V_h^\pi\right\|_\infty \le C\cdot H^2\sqrt{\frac{S\log(SA/\delta)}{N}},$$

Proof:

$$\begin{aligned}
\widehat{Q}_h^\pi - Q_h^\pi &= r_h + \widehat{P}^{\pi_{h+1}}\widehat{Q}_{h+1}^\pi - r_h - P^{\pi_{h+1}}Q_{h+1}^\pi \\
&= \left(\widehat{P}^{\pi_{h+1}} - P^{\pi_{h+1}}\right)\widehat{Q}_{h+1}^\pi + P^{\pi_{h+1}}\left(Q_{h+1}^\pi - Q_{h+1}^\pi\right) \\
&= \left(\widehat{P} - P\right)\widehat{V}_{h+1}^\pi + P^{\pi_{h+1}}\left(\widehat{Q}_{h+1}^\pi - Q_{h+1}^\pi\right) \\
&= \ldots = \sum_{t=h}^{H} \Gamma_{h+1:t}^\pi \left(\widehat{P} - P\right)\widehat{V}_{t+1}^\pi \\
&\leq \sum_{t=h}^{H} \Gamma_{h+1:t}^\pi \left|\left(\widehat{P} - P\right)\widehat{V}_{t+1}^\pi\right| \\
&\leq \sum_{t=h}^{H} 1 \cdot \max_{s,a} \left\|(\widehat{P} - P)(\cdot|s,a)\right\|_1 \cdot \left\|\widehat{V}_{t+1}^\pi\right\|_\infty \cdot \mathbf{1} \\
&\leq H^2 \cdot \max_{s,a} \left\|(\widehat{P} - P)(\cdot|s,a)\right\|_1 \cdot \mathbf{1} \leq C \cdot H^2 \sqrt{\frac{S\log(SA/\delta)}{N}}\mathbf{1}
\end{aligned}$$

with probability $1-\delta$, where the last inequality is by Lemma D.0.10. By symmetry and taking the $\|\cdot\|_\infty$, we obtain w.p. $1-\delta$

$$\sup_{\pi\in\Pi_g, h\in[H]} \left\|\widehat{Q}_h^\pi - Q_h^\pi\right\|_\infty \leq C \cdot H^2 \sqrt{\frac{S\log(SA/\delta)}{N}}.$$

The above holds for $\forall\pi \in \Pi_g$ since Lemma D.0.10 acts on $\left\|\widehat{P} - P\right\|_1$ and is irrelevant to π.

Bibliography

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

[2] S. Lange, T. Gabel, and M. Riedmiller, *Batch reinforcement learning*, in *Reinforcement learning*, pp. 45–73. Springer, 2012.

[3] H. Le, C. Voloshin, and Y. Yue, *Batch policy learning under constraints*, in *International Conference on Machine Learning*, pp. 3703–3712, 2019.

[4] J. Chen and N. Jiang, *Information-theoretic considerations in batch reinforcement learning*, in *International Conference on Machine Learning*, pp. 1042–1051, 2019.

[5] T. Xie and N. Jiang, *Q* approximation schemes for batch reinforcement learning: A theoretical comparison*, in *Uncertainty in Artificial Intelligence*, pp. 550–559, 2020.

[6] T. Xie and N. Jiang, *Batch value-function approximation with only realizability*, *International Conference on Machine Learning* (2021).

[7] M. Yin, Y. Bai, and Y.-X. Wang, *Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning*, in *International Conference on Artificial Intelligence and Statistics*, pp. 1567–1575, PMLR, 2021.

[8] M. Yin, Y. Bai, and Y.-X. Wang, *Near-optimal offline reinforcement learning via double variance reduction*, *Advances in neural information processing systems* (2021).

[9] T. Ren, J. Li, B. Dai, S. S. Du, and S. Sanghavi, *Nearly horizon-free offline reinforcement learning*, *Advances in neural information processing systems* (2021).

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

[11] T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai, *Policy finetuning: Bridging sample-efficient offline and online reinforcement learning*, *Advances in neural information processing systems* (2021).

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] L. Li, W. Chu, J. Langford, and X. Wang, *Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms*, in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, ACM, 2011.

[14] M. Dudík, J. Langford, and L. Li, *Doubly robust policy evaluation and learning*, *arXiv preprint arXiv:1103.4601* (2011).

[15] A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni, *Off-policy evaluation for slate recommendation*, in *Advances in Neural Information Processing Systems*, pp. 3632–3642, 2017.

[16] Y.-X. Wang, A. Agarwal, and M. Dudik, *Optimal and adaptive off-policy evaluation in contextual bandits*, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3589–3597, JMLR. org, 2017.

[17] L. Li, R. Munos, and C. Szepesvári, *Toward minimax off-policy value estimation*, .

[18] N. Jiang and L. Li, *Doubly robust off-policy value evaluation for reinforcement learning*, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pp. 652–661, JMLR. org, 2016.

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

[20] M. Farajtabar, Y. Chow, and M. Ghavamzadeh, *More robust doubly robust off-policy evaluation*, *arXiv preprint arXiv:1802.03493* (2018).

[21] T. Xie, Y. Ma, and Y.-X. Wang, *Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling*, in *Advances in Neural Information Processing Systems*, pp. 9665–9675, 2019.

[22] G. Theocharous, P. S. Thomas, and M. Ghavamzadeh, *Personalized ad recommendation systems for life-time value optimization with guarantees*, in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[23] P. S. Thomas, G. Theocharous, M. Ghavamzadeh, I. Durugkar, and E. Brunskill, *Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing*, in *Twenty-Ninth IAAI Conference*, 2017.

[24] T. Mandel, Y.-E. Liu, S. Levine, E. Brunskill, and Z. Popovic, *Offline policy evaluation across representations with applications to educational games*, in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1077–1084, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[25] S. A. Murphy, M. J. van der Laan, J. M. Robins, and C. P. P. R. Group, *Marginal mean models for dynamic regimes*, *Journal of the American Statistical Association* **96** (2001), no. 456 1410–1423.

[26] K. Hirano, G. W. Imbens, and G. Ridder, *Efficient estimation of average treatment effects using the estimated propensity score*, *Econometrica* **71** (2003), no. 4 1161–1189.

[27] Y. Liu, O. Gottesman, A. Raghu, M. Komorowski, A. A. Faisal, F. Doshi-Velez, and E. Brunskill, *Representation balancing mdps for off-policy policy evaluation*, in *Advances in Neural Information Processing Systems*, pp. 2644–2653, 2018.

[28] O. Gottesman, Y. Liu, S. Sussex, E. Brunskill, and F. Doshi-Velez, *Combining parametric and nonparametric models for off-policy evaluation*, *arXiv preprint arXiv:1905.05787* (2019).

[29] Q. Liu, L. Li, Z. Tang, and D. Zhou, *Breaking the curse of horizon: Infinite-horizon off-policy estimation*, in *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018.

[30] A. Hallak and S. Mannor, *Consistent on-line off-policy evaluation*, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1372–1383, JMLR. org, 2017.

[31] C. Gelada and M. G. Bellemare, *Off-policy deep reinforcement learning by bootstrapping the covariate shift*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3647–3655, 2019.

[32] N. Kallus and M. Uehara, *Double reinforcement learning for efficient off-policy evaluation in markov decision processes*, *arXiv preprint arXiv:1908.08526* (2019).

[33] N. Kallus and M. Uehara, *Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes*, *arXiv preprint arXiv:1909.05850* (2019).

[34] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, .

[35] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998.

[36] N. Jiang and A. Agarwal, *Open problem: The dependence of sample complexity lower bounds on planning horizon*, in *Conference On Learning Theory*, pp. 3395–3398, 2018.

[37] A. R. Mahmood, H. P. van Hasselt, and R. S. Sutton, *Weighted importance sampling for off-policy learning with linear function approximation*, in *Advances in Neural Information Processing Systems*, pp. 3014–3022, 2014.

[38] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.

[39] M. G. Azar, I. Osband, and R. Munos, *Minimax regret bounds for reinforcement learning*, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272, JMLR. org, 2017.

[40] M. Kearns and S. Singh, *Near-optimal reinforcement learning in polynomial time*, *Machine learning* **49** (2002), no. 2-3 209–232.

[41] N. Jiang, *Notes on tabular methods*, .

[42] A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye, *Near-optimal time and sample complexities for solving markov decision processes with a generative model*, in *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018.

[43] M. G. Azar, R. Munos, and H. J. Kappen, *Minimax pac bounds on the sample complexity of reinforcement learning with a generative model*, *Machine learning* **91** (2013), no. 3 325–349.

[44] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, *Counterfactual reasoning and learning systems: The example of computational advertising*, *The Journal of Machine Learning Research* **14** (2013), no. 1 3207–3260.

[45] L. Tang, R. Rosales, A. Singh, and D. Agarwal, *Automatic ad format selection via contextual bandits*, in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 1587–1594, ACM, 2013.

[46] F. Bertoluzzo and M. Corazza, *Testing different reinforcement learning configurations for financial trading: Introduction and applications*, *Procedia Economics and Finance* **3** (2012) 68–77.

[47] D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine, *Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods*, in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6284–6291, IEEE, 2018.

[48] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, *Robonet: Large-scale multi-robot learning*, in *Conference on Robot Learning*, pp. 885–897, 2020.

[49] N. Jaques, A. Ghandeharioun, J. H. Shen, C. Ferguson, A. Lapedriza, N. Jones, S. Gu, and R. Picard, *Way off-policy batch deep reinforcement learning of implicit human preferences in dialog*, *arXiv preprint arXiv:1907.00456* (2019).

[50] D. Ernst, G.-B. Stan, J. Goncalves, and L. Wehenkel, *Clinical data based optimal sti strategies for hiv: a reinforcement learning approach*, in *Decision and Control, 2006 45th IEEE Conference on*, pp. 667–672, IEEE, 2006.

[51] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi, *Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach*, in *Machine Learning for Healthcare Conference*, pp. 147–163, 2017.

[52] A. Raghu, O. Gottesman, Y. Liu, M. Komorowski, A. Faisal, F. Doshi-Velez, and E. Brunskill, *Behaviour policy estimation in off-policy policy evaluation: Calibration matters*, *arXiv preprint arXiv:1807.01066* (2018).

[53] O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi, *Guidelines for reinforcement learning in healthcare*, *Nat Med* **25** (2019), no. 1 16–18.

[54] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[55] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Learnability, stability and uniform convergence*, *The Journal of Machine Learning Research* **11** (2010) 2635–2670.

[56] M. Farajtabar, Y. Chow, and M. Ghavamzadeh, *More robust doubly robust off-policy evaluation*, in *Proceedings of the 35th International Conference on Machine Learning*, (Stockholmsmässan, Stockholm Sweden), pp. 1447–1456, PMLR, 10–15 Jul, 2018.

[57] M. Yin and Y.-X. Wang, *Asymptotically efficient off-policy evaluation for tabular reinforcement learning*, in *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958, PMLR, 2020.

[58] R. Munos, *Error bounds for approximate policy iteration*, in *International Conference on Machine Learning*, pp. 560–567, 2003.

[59] Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill, *Provably good batch reinforcement learning without great exploration*, *Advances in neural information processing systems* (2020).

[60] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, *Provably efficient reinforcement learning with linear function approximation*, in *Conference on Learning Theory*, pp. 2137–2143, PMLR, 2020.

[61] A. Agarwal, S. Kakade, and L. F. Yang, *Model-based reinforcement learning with a generative model is minimax optimal*, in *Conference on Learning Theory*, pp. 67–83, 2020.

[62] A. Tewari, *Reinforcement learning in large or unknown MDPs*. University of California, Berkeley, 2007.

[63] Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill, *Off-policy policy gradient with state distribution correction*, in *Uncertainty in Artificial Intelligence*.

[64] Y. Duan, Z. Jia, and M. Wang, *Minimax-optimal off-policy evaluation with linear function approximation*, in *International Conference on Machine Learning*, pp. 8334–8342, 2020.

[65] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire, *Contextual decision processes with low bellman rank are pac-learnable*, in *International Conference on Machine Learning-Volume 70*, pp. 1704–1713, 2017.

[66] Y. Ye, *The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate*, *Mathematics of Operations Research* **36** (2011), no. 4 593–603.

[67] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[68] Q. Cui and L. F. Yang, *Is plug-in solver sample-efficient for feature-based reinforcement learning?*, in *Advances in neural information processing systems*, 2020.

[69] C. Szepesvári and R. Munos, *Finite time bounds for sampling based fitted value iteration*, in *Proceedings of the 22nd international conference on Machine learning*, pp. 880–887, 2005.

[70] A. Antos, R. Munos, and C. Szepesvari, *Fitted q-iteration in continuous action-space mdps*, in *Advances in Neural Information Processing Systems*, pp. 9–16, 2008.

[71] A. Antos, C. Szepesvári, and R. Munos, *Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path*, *Machine Learning* **71** (2008), no. 1 89–129.

[72] Y. Jin, Z. Yang, and Z. Wang, *Is pessimism provably efficient for offline rl?*, *International Conference on Machine Learning* (2020).

[73] R. Wang, D. P. Foster, and S. M. Kakade, *What are the statistical limits of offline rl with linear function approximation?*, *International Conference on Machine Learning* (2021).

[74] A. Zanette, *Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl*, *International Conference on Machine Learning* (2021).

[75] Y. Efroni, N. Merlis, M. Ghavamzadeh, and S. Mannor, *Tight regret bounds for model-based reinforcement learning with greedy policies*, in *Advances in Neural Information Processing Systems*, 2019.

[76] K. Zhang, S. M. Kakade, T. Başar, and L. F. Yang, *Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity*, *arXiv preprint arXiv:2007.07461* (2020).

[77] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, *Breaking the sample size barrier in model-based reinforcement learning with a generative model*, *Advances in Neural Information Processing Systems* **33** (2020).

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

[79] R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims, *Morel: Model-based offline reinforcement learning*, *Advances in neural information processing systems* (2020).

[80] C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu, *Reward-free exploration for reinforcement learning*, in *International Conference on Machine Learning*, pp. 4870–4879, PMLR, 2020.

[81] E. Kaufmann, P. Ménard, O. D. Domingues, A. Jonsson, E. Leurent, and M. Valko, *Adaptive reward-free exploration*, *arXiv preprint arXiv:2006.06294* (2020).

[82] P. Menard, O. D. Domingues, A. Jonsson, E. Kaufmann, E. Leurent, and M. Valko, *Fast active learning for pure exploration in reinforcement learning*, *arXiv preprint arXiv:2007.13442* (2020).

[83] Z. Zhang, S. S. Du, and X. Ji, *Nearly minimax optimal reward-free reinforcement learning*, *arXiv preprint arXiv:2010.05901* (2020).

[84] X. Zhang, A. Singla, *et. al.*, *Task-agnostic exploration in reinforcement learning*, *Advances in Neural Information Processing Systems* (2020).

[85] R. Wang, S. S. Du, L. F. Yang, and R. Salakhutdinov, *On reward-free reinforcement learning with linear function approximation*, *arXiv preprint arXiv:2006.11274* (2020).

[86] Q. Liu, T. Yu, Y. Bai, and C. Jin, *A sharp analysis of model-based reinforcement learning with self-play*, *arXiv preprint arXiv:2010.01604* (2020).

[87] T. Jaksch, R. Ortner, and P. Auer, *Near-optimal regret bounds for reinforcement learning.*, *Journal of Machine Learning Research* **11** (2010), no. 4.

[88] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang, *Model-based reinforcement learning with value-targeted regression*, in *International Conference on Machine Learning*, pp. 463–474, PMLR, 2020.

[89] Y. Han, J. Jiao, and T. Weissman, *Minimax estimation of discrete distributions under l_1 loss*, *IEEE Transactions on Information Theory* **61** (2015), no. 11 6343–6354.

[90] L. Yang and M. Wang, *Sample-optimal parametric q-learning using linearly additive features*, in *International Conference on Machine Learning*, pp. 6995–7004, PMLR, 2019.

[91] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[92] D. Zhou, Q. Gu, and C. Szepesvari, *Nearly minimax optimal reinforcement learning for linear mixture markov decision processes*, *arXiv preprint arXiv:2012.08507* (2020).

[93] M. Yin and Y.-X. Wang, *Optimal uniform ope and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings*, *Advances in Neural Information Processing Systems* (2021).

[94] C. Dann, L. Li, W. Wei, and E. Brunskill, *Policy certificates: Towards accountable reinforcement learning*, in *International Conference on Machine Learning*, pp. 1507–1516, PMLR, 2019.

[95] Z. Zhang, X. Ji, and S. S. Du, *Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon*, *Conference of Learning Theory* (2021).

[96] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, *Is q-learning provably efficient?*, in *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

[97] M. Yin and Y.-X. Wang, *Towards instance-optimal offline reinforcement learning with pessimism*, *Advances in Neural Information Processing Systems,* (2021).

[98] M. Yin, Y. Duan, M. Wang, and Y.-X. Wang, *Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism*, *International Conference on Learning Representations,* (2022).

[99] T. Nguyen-Tan, M. Yin, S. Gupta, S. Venkates, and R. Arora, *On instance-dependent bounds for offline reinforcement learning with linear function approximation*, *Association for the Advancement of Artificial Intelligence,* (2023).

[100] M. Yin, M. Wang, and Y.-X. Wang, *Offline reinforcement learning with differentiable function approximation is provably efficient*, *International Conference on Learning Representations,* (2023).

[101] M. Yin, W. Chen, M. Wang, and Y.-X. Wang, *Offline stochastic shortest path: Learning, evaluation and towards optimality*, *Uncertainty in Artificial Intelligence,* (2022).

[102] D. Qiao, M. Yin, M. Min, and Y.-X. Wang, *Sample-efficient reinforcement learning with loglog(t) switching cost*, *International Conference on Machine Learning,* (2022).

[103] J. Li, E. Zhang, M. Yin, Q. Bai, Y.-X. Wang, and W. Y. Wang, *Offline reinforcement learning with closed-form policy improvement operators*, *NeurIPS workshop in Offline RL,* (2022).

[104] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine, *D4rl: Datasets for deep data-driven reinforcement learning*, *arXiv preprint arXiv:2004.07219* (2020).

[105] C. Jin, Q. Liu, and S. Miryoosefi, *Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms*, *Advances in neural information processing systems* **34** (2021) 13406–13418.

[106] S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang, *Bilinear classes: A structural framework for provable generalization in rl*, in *International Conference on Machine Learning*, pp. 2826–2836, PMLR, 2021.

[107] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, *The statistical complexity of interactive decision making*, *arXiv preprint arXiv:2112.13487* (2021).

[108] C. Dann, T. Lattimore, and E. Brunskill, *Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning*, in *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

[109] M. J. Kearns and S. P. Singh, *Finite-sample convergence rates for q-learning and indirect algorithms*, in *Advances in neural information processing systems*, pp. 996–1002, 1999.

[110] A. Agarwal, S. Kakade, and L. F. Yang, *On the optimality of sparse model-based planning for markov decision processes*, *arXiv preprint arXiv:1906.03804* (2019).

[111] P. S. Thomas, *Safe reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 2015.

[112] R. I. Brafman and M. Tennenholtz, *R-max-a general polynomial time algorithm for near-optimal reinforcement learning*, *Journal of Machine Learning Research* **3** (2002), no. Oct 213–231.

[113] F. Chung and L. Lu, *Concentration inequalities and martingale inequalities: a survey*, *Internet Mathematics* **3** (2006), no. 1 79–127.

[114] C. Dann and E. Brunskill, *Sample complexity of episodic fixed-horizon reinforcement learning*, in *Advances in Neural Information Processing Systems*, pp. 2818–2826, 2015.

[115] A. Krishnamurthy, A. Agarwal, and J. Langford, *PAC reinforcement learning with rich observations*, in *Advances in Neural Information Processing Systems*, pp. 1840–1848, 2016.

[116] S. Mannor and J. N. Tsitsiklis, *The sample complexity of exploration in the multi-armed bandit problem*, *Journal of Machine Learning Research* **5** (2004), no. Jun 623–648.

[117] A. Agarwal, N. Jiang, and S. M. Kakade, *Reinforcement learning: Theory and algorithms*, *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep* (2019).

[118] H. Chernoff *et. al.*, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, *The Annals of Mathematical Statistics* **23** (1952), no. 4 493–507.

[119] K. Sridharan, *A gentle introduction to concentration inequalities*, *Dept. Comput. Sci., Cornell Univ., Tech. Rep* (2002).

[120] J. Tropp *et. al.*, *Freedman's inequality for matrix martingales*, *Electronic Communications in Probability* **16** (2011) 262–270.