DISTRIBUTIONAL OFF-POLICY EVALUATION WITH BELLMAN RESIDUAL MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the problem of distributional off-policy evaluation which serves as the foundation of many distributional reinforcement learning (DRL) algorithms. In contrast to most existing works (that rely on supremum-extended statistical distances), we study the expectation-extended statistical distance for quantifying the distributional Bellman residuals and provide the corresponding theoretical supports. Extending the framework of Bellman residual minimization to DRL, we propose a method called Energy Bellman Residual Minimizer (EBRM) to estimate the return distribution. We establish a finite-sample error bound for the EBRM estimator under the realizability assumption. Additionally, we introduce a variant of our method based on a multi-step bootstrapping procedure to enable multi-step extension. By selecting an appropriate step level, we obtain a better error bound for this variant of EBRM compared to a single-step EBRM, under non-realizability settings. Finally, we demonstrate the superior performance of our method through simulation studies, comparing with other existing methods.

1 INTRODUCTION

In reinforcement learning (RL), the cumulative (discounted) reward, also known as the return, is a crucial quantity for evaluating the performance of a policy. Most existing RL methods focus on only the expectation of the return distribution. In Bellemare et al. (2017a), the focus has been extended to the whole return distribution, and they introduce a distributional RL (DRL) algorithm (hereafter called Categorical algorithm) that achieves a considerably better performance in Atari games than expectation-oriented Deep-Q Networks (Mnih et al., 2015). This has sparked significant interests among the RL community, and was later followed by a series of quantile-based methods including QRDQN, QRTD (Dabney et al., 2018b), IQN (Dabney et al., 2018a), FQF (Yang et al., 2019), EDRL (Rowland et al., 2019) and particle-based methods including MMDRL (Nguyen-Tang et al., 2021), SinkhornDRL (Sun et al., 2022), MD3QN (Zhang et al., 2021). In this paper, we consider the problem of off-policy evaluation in DRL, i.e., estimating the (conditional) return distribution of a target policy based on offline data.

Despite their competitive performances, distributional RL methods are significantly underdeveloped compared with the traditional expectation-based RL, especially in the theoretical development under the offline setting. All aforementioned methods are motivated by supremum-extended distances due to the contraction property (see (4) below), but their algorithms essentially minimize an expectation-extended distance (see (6)), as summarized in the column "Distance Mismatch" of Table 1. This leads to a theory-practice gap. Also, most of these work does not provide any statistical guarantee such as the convergence rate. We note that Rowland et al. (2018) establishes the consistency of their estimator, but no error bound analysis (and convergence rate) is provided. In terms of statistical analysis, a very recent work FLE (Wu et al., 2023) only offers error bound analysis of their estimator for the marginal distribution of return, which is hard to use for policy learning. In addition, their analysis is based on a strong condition called completeness, which in general significantly restricts model choices of return distributions and excludes the non-realizable scenario.

This paper proposes novel estimators, which we call Energy Bellman Residual Minimizer (EBRM), based on the idea of Bellman residual minimization for the conditional distribution of the return. In contrast to existing work, we provide solid theoretical ground for the application of expectation-extended distance in measuring (distributional) Bellman residual. A multi-step extension of our

estimator is proposed for non-realizability settings. Our method comes with statistical error bound analyses in both realizable and non-realizable settings. Table 1 provides some key comparisons between our method and some existing works. More details is given in Table 3 in the Appendix D.1. Finally, we summarize our contributions as follows. (1) We provide theoretical foundation of the application of expectation-extended distance for Bellman residual minimization in DRL. See Section 2.3. (2) We develop a novel distributional off-policy evaluation method (EBRM), together with its finite-sample error bound. See Section 3. (3) We develop a multi-step extension of EBRM for non-realizabile settings in Section 4. We also provide corresponding finite-sample error bound under non-realizable settings. (4) Our numerical experiments in Section 5 demonstrate the strong performance of EBRM compared with some baseline methods.

Table 1: Comparison among DRL methods in off-policy evaluation.

Method	Distance match	Statistical error bound	Non-realizable	Multi-dimension
Categorical (Bellemare et al., 2017a)	✗	✗	NA	✓
QRTD (Dabney et al., 2018b)	✗	✗	NA	✗
IQN (Dabney et al., 2018a)	✗	✗	NA	✗
FQF (Yang et al., 2019)	✗	✗	NA	✗
EDRL (Rowland et al., 2019)	✗	✗	NA	✗
MMDRL (Nguyen-Tang et al., 2021)	✗	✗	NA	✓
SinkhornDRL (Sun et al., 2022)	✗	✗	NA	✓
MD3QN (Zhang et al., 2021)	✗	✗	NA	✓
FLE (Wu et al., 2023)	✓	✓	NA	✓
EBRM (our method)	✓	✓	✓	✓

2 Off-policy Evaluation Based on Bellman Equation

2.1 Background

We consider an off-policy evaluation (OPE) problem under the framework of infinite-horizon Markov Decision Process (MDP), which is characterized by a state space $\mathcal{S}$, a discrete action space $\mathcal{A}$, and a transition probability $p : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}^d \times \mathcal{S})$ with $\mathcal{P}(\mathcal{X})$ denoting the class of probability measures over a generic space $\mathcal{X}$. In other words, p defines a joint distribution of a d-dimensional immediate reward and the next state conditioned on a state-action pair. At each time point, an action is chosen by the agent based on a current state according to a (stochastic) policy, a mapping from $\mathcal{S}$ to $\mathcal{P}(\mathcal{A})$. With the initial state-action pair $(S^{(0)}, A^{(0)})$, a trajectory generated by such an MDP can be written as $\{S^{(t)}, A^{(t)}, R^{(t+1)}\}_{t\geq 0}$. The return variable is defined as $Z := \sum_{t=1}^{\infty} \gamma^{t-1} R^{(t)}$ with $\gamma \in [0, 1)$ being a discount factor, based on which we can evaluate the performance of some target policy π.

Traditional OPE methods are mainly focused on estimating the expectation of return Z under the target policy π, whereas DRL aims to estimate the whole distribution of Z. Letting $\mathcal{L}(X)$ be the probability measure of some random variable (or vector) X, our target is to estimate the collection of return distributions conditioned on different initial state-action pairs $(S^{(0)}, A^{(0)}) = (s, a)$:

$$\Upsilon_\pi(s,a) := \mathcal{L}\bigg(\sum_{t=1}^{\infty} \gamma^{t-1} R^{(t)}\bigg),\ (R^{(t+1)}, S^{(t+1)}) \sim p(\cdot|S^{(t)}, A^{(t)}),\ A^{(t+1)} \sim \pi(\cdot|S^{(t+1)}), \quad (1)$$

collectively written as $\Upsilon_\pi \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$. It is analogous to the Q-function in traditional RL, whose evaluation at a state-action pair (s, a) is the expectation of the distribution $\Upsilon_\pi(s, a)$. Our goal in this paper is to use the offline data generated by the behavior policy b to estimate Υ_π.

Similar to most existing DRL methods, our proposal is based on the distributional Bellman equation (Bellemare et al., 2017a). Define the distributional Bellman operator by $\mathcal{T}^\pi : \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}} \to \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ such that, for any $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$,

$$\big(\mathcal{T}^\pi \Upsilon\big)(s,a) := \int_{\mathbb{R}^d\times\mathcal{S}\times\mathcal{A}} (g_{r,\gamma})_{\#}\Upsilon(s',a')\mathrm{d}\pi(a'|s')\mathrm{d}p(r,s'|s,a), \quad (s,a) \in \mathcal{S}\times\mathcal{A}, \quad (2)$$

where $(g_{r,\gamma})_{\#} : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$ maps the distribution of any random vector X to the distribution of $r + \gamma X$. One can show that Υ_π is the unique solution to the distributional Bellman equation:

$$\mathcal{T}^\pi \Upsilon = \Upsilon. \tag{3}$$

Letting $Z_\pi(s,a)$ be the random vector that follows the distribution $\Upsilon_\pi(s,a)$, one can also express the distributional Bellman equation (3) in a more intuitive way: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$Z_\pi(s,a) \overset{D}{=} R + \gamma Z_\pi(S', A') \quad \text{where} \quad (R, S') \sim p(\cdot|s,a), A' \sim \pi(\cdot|S'),$$

where $\overset{D}{=}$ refers to the equivalence in terms of the underlying distributions. Due to the distributional Bellman equation (3), a sensible approach to find Υ_π is based on minimizing the discrepancy between $\mathcal{T}^\pi \Upsilon$ and Υ with respect to $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$, which will be called *Bellman residual* hereafter. To proceed with this approach, two important issues need to be addressed. First, both $\mathcal{T}^\pi \Upsilon$ and Υ are collections of distributions over $\mathbb{R}^d$, based on which Bellman residual shall be quantified. Second, $\mathcal{T}^\pi$ may not be available and therefore needs to be estimated through data. We will focus on the quantification of Bellman residual first, and defer the proposed estimator of $\mathcal{T}^\pi$ and the formal description of our estimator for Υ_π to Section 3.

2.2 Existing Measures of Bellman Residuals

To quantify the discrepancy between the two sides of the distributional Bellman equation (3), one can use a distance over $\mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$. Fixing a state-action pair, one can solely compare two distributions from $\mathcal{P}(\mathbb{R}^d)$. Therefore, a common strategy is to start by selecting a statistical distance $\eta(\cdot,\cdot) : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0,\infty]$, and then define an extended-distance over $\mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ through combining the statistical distances over different state-action pairs. As shown in Table 3 in Appendix D.1, most existing methods (e.g., Bellemare et al., 2017b;a; Nguyen et al., 2020) are based on some *supremum-extended* distance η_∞:

$$\eta_\infty(\Upsilon_1, \Upsilon_2) := \sup_{s,a} \eta\Big\{\Upsilon_1(s,a), \Upsilon_2(s,a)\Big\}. \tag{4}$$

Under various choices of η including Wasserstein-p metric with $1 \le p \le \infty$ (Bellemare et al., 2017a; Dabney et al., 2018b) and maximum mean discrepancy (Nguyen-Tang et al., 2021), it is shown that $\mathcal{T}^\pi$ is a contraction with respect to η_∞. More specifically, $\eta_\infty(\mathcal{T}^\pi\Upsilon_1, \mathcal{T}^\pi\Upsilon_2) \le \gamma^{\beta_0} \cdot \eta_\infty(\Upsilon_1, \Upsilon_2)$ holds for any $\Upsilon_1, \Upsilon_2 \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$, where the value of $\beta_0 > 0$ depends on the choice of η. If η_∞ is a metric, then the contractive property implies, for any $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$,

$$\eta_\infty(\Upsilon, \Upsilon_\pi) \le \sum_{k=1}^{\infty} \eta_\infty\big\{(\mathcal{T}^\pi)^{k-1}\Upsilon, (\mathcal{T}^\pi)^k\Upsilon\big\} \le \frac{1}{1-\gamma^{\beta_0}} \cdot \eta_\infty(\Upsilon, \mathcal{T}^\pi\Upsilon). \tag{5}$$

As such, minimizing Bellman residual measured by η_∞ would be a sensible approach for finding Υ_π. However, as surveyed in Appendix D.1, most existing methods in practice essentially minimize an empirical (and approximated) version of the *expectation-extended* distance defined by

$$\bar{\eta}(\Upsilon_1, \Upsilon_2) := \mathbb{E}_{(S,A)\sim b_\mu} \eta\Big\{\Upsilon_1(S,A), \Upsilon_2(S,A)\Big\}, \tag{6}$$

with $(S,A) \sim b_\mu$. Here $b_\mu = \mu \times b$ refer to data distribution over $\mathcal{S} \times \mathcal{A}$ induced by the behavior policy b. With a slight abuse of notation, we will overload the notation b_μ with its density (with respect to some appropriate base measure of $\mathcal{S} \times \mathcal{A}$, e.g., counting measure and Lebesgue measure). We remark that (5) does not hold under $\bar{\eta}$ because η_∞ and $\bar{\eta}$ are not necessarily equivalent for the general state-action space, leading to a theory-practice gap in most methods (Column 1 of Table 1).

2.3 Expectation-extended Distance

Despite the implicit use of expectation-extended distances in some prior works, the corresponding theoretical foundations are not well established. Regarding Bellman residual minimization, a very natural and crucial question is:

> *In terms of an expectation-extended distance, does small Bellman residual of Υ lead to closeness between Υ and Υ_π?*

To proceed, we focus on settings such that the state-action pairs of interest can be well covered by b_μ, as formally stated in the following assumption. Let $q^\pi(s,a|\tilde{s},\tilde{a})$ be the conditional probability density of the next state-action pair at (s,a) conditional on the current state-action pair at $(\tilde{s},\tilde{a})$, defined by the transition probability p and the target policy π.

Assumption 1. *There exists $p_{\min} > 0$ and $p_{\max} < \infty$ such that $b_\mu(s,a) \geq p_{\min}$ for all $s,a \in \mathcal{S} \times \mathcal{A}$ and $q^\pi(s,a|\tilde{s},\tilde{a}) \leq p_{\max}$ for all $(\tilde{s},\tilde{a}),(\tilde{s},\tilde{a}) \in \mathcal{S} \times \mathcal{A}$.*

Let $q_{b_\mu}^{\pi:t}(s,a)$ be the probability density (or mass) of $(S^{(t)},A^{(t)})$ at (s,a), given $(S^{(0)},A^{(0)}) \sim b_\mu$ and the target policy π. Assumption 1 implies uniformly bounded density ratio, that is $q_{b_\mu}^{\pi:t}(s,a)/b_\mu(s,a) \leq C_{\sup}(< \infty)$ for all $t \in \mathbb{N}$, as proved in Appendix A.1.

In the following Theorem 1 (proved in Appendix A.2), we provide a solid ground for Bellman residual minimization based on expectation-extended distances.

Theorem 1. *Under Assumption 1, if the statistical distance η satisfies translation-invariance, scale-sensitivity of order $\beta_0 > 0$, convexity, and relaxed triangular inequality defined in Appendix A.2.1, then we can bound the inaccuracy:*

$$\bar{\eta}(\Upsilon,\Upsilon_\pi) \leq 2C_{\sup}B_1(\gamma;\beta_0)\cdot\bar{\eta}(\Upsilon,\mathcal{T}^\pi\Upsilon), \tag{7}$$

where $B_1(\gamma;\beta_0) := \frac{1}{2(1-\gamma^{\beta_0})}\sum_{k=1}^{\infty} 4^k\gamma^{(2^{k-1}-1)\beta_0} < \infty$ is an increasing function of $\gamma \in (0,1)$, and $C_{\sup}$ is defined in (25).

Inequality (7) provides an analogy to Bound (5) for expectation-based distances, answering our prior question positively for some expectation-extended distances. Note that Theorem 1 can be applied to the settings with general state-action space, including continuous one.

In order to take advantage of Theorem 1, we should select a statistical distance that satisfies all the properties stated in Theorem 1. One example is energy distance (Székely & Rizzo, 2013) as proved in Appendix A.3, which is in fact a squared maximum mean discrepancy (Gretton et al., 2012) with kernel $k(\mathbf{x},\mathbf{y}) = \|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x}-\mathbf{y}\|$. The energy distance is defined as

$$\mathcal{E}\{\mathcal{L}(\mathbf{X}),\mathcal{L}(\mathbf{Y})\} := 2\mathbb{E}\|\mathbf{X}-\mathbf{Y}\| - \mathbb{E}\|\mathbf{X}-\mathbf{X}'\| - \mathbb{E}\|\mathbf{Y}-\mathbf{Y}'\|, \tag{8}$$

where $\mathbf{X}'$ and $\mathbf{Y}'$ are independent copies of $\mathbf{X}$ and $\mathbf{Y}$ respectively, and $\mathbf{X},\mathbf{X}',\mathbf{Y},\mathbf{Y}'$ are independent. In below, we will use energy distance to construct our estimator.

3 Energy Bellman Residual Minimizer

3.1 Estimated Bellman Residual

Despite applicability of Theorem 1 to general state-action space, we will focus on tabular case with finite cardinality $|\mathcal{S}\times\mathcal{A}| < \infty$ for simpler construction of estimation, which enables an in-depth theoretical study under both realizable and non-realizable settings in Sections 3.2 and 4.3. But the reward can be continuous. Our target objective of Bellman residual minimization is

$$\bar{\mathcal{E}}(\Upsilon,\mathcal{T}^\pi\Upsilon) = \sum_{s,a} b_\mu(s,a)\cdot\mathcal{E}\big\{\Upsilon(s,a),\mathcal{T}^\pi\Upsilon(s,a)\big\}, \quad \text{where} \tag{9}$$

$$\begin{aligned}\mathcal{E}\big\{\Upsilon(s,a),\mathcal{T}^\pi\Upsilon(s,a)\big\} = 2\mathbb{E}\|Z_\alpha(s,a) - Z_\beta^{(1)}(s,a)\| - \mathbb{E}\|Z_\alpha(s,a) - Z_\beta(s,a)\| \\ - \mathbb{E}\|Z_\alpha^{(1)}(s,a) - Z_\beta^{(1)}(s,a)\|,\end{aligned}$$

where $Z_\alpha(s,a), Z_\beta(s,a) \sim \Upsilon(s,a)$ and $Z_\alpha^{(1)}(s,a), Z_\beta^{(1)}(s,a) \sim \mathcal{T}^\pi\Upsilon(s,a)$ are all independent. For the tabular case with offline data, we can estimate b_μ and the transition p simply by empirical distributions. That is, given observations $\mathcal{D} = \{(s_i,a_i,r_i,s_i')\}_{i=1}^N$, we consider

$$\hat{b}_\mu(s,a) := \frac{N(s,a)}{N} \quad \text{where} \quad N(s,a) := \sum_{i=1}^N \mathbf{1}\big\{(s_i,a_i) = (s,a)\big\}, \quad \text{and} \tag{10}$$

$$\hat{p}(E|s,a) := \begin{cases} \frac{1}{N(s,a)}\sum_{i:(s_i,a_i)=(s,a)}\delta_{r_i,s_i'}(E) & \text{if } N(s,a) \geq 1, \\ \delta_{\mathbf{0},s}(E) & \text{if } N(s,a) = 0 \end{cases} \quad \text{for any measurable set } E,$$

where $\delta_{r,s'}$ is the Dirac measure at (r, s'). Based on this, we can estimate $\mathcal{T}^\pi$ for any $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ by the estimated transition $\hat{p}$ and the target policy π, by replacing p of (2) with $\hat{p}$.

Denoting the conditional expectation by $\tilde{\mathbb{E}}(\cdots) := \mathbb{E}(\cdots|\mathcal{D})$, we can compute

$$\begin{aligned}\mathcal{E}\big\{\Upsilon(s,a), \hat{\mathcal{T}}^\pi\Upsilon(s,a)\big\} = 2\tilde{\mathbb{E}}\|Z_\alpha(s,a) - \hat{Z}_\beta^{(1)}(s,a)\| - \tilde{\mathbb{E}}\|Z_\alpha(s,a) - Z_\beta(s,a)\| \\ - \tilde{\mathbb{E}}\|\hat{Z}_\alpha^{(1)}(s,a) - \hat{Z}_\beta^{(1)}(s,a)\|, \end{aligned} \tag{11}$$

where $Z_\alpha(s,a), Z_\beta(s,a) \sim \Upsilon_\theta(s,a)$ and $\hat{Z}_\alpha^{(1)}(s,a), \hat{Z}_\beta^{(1)}(s,a) \sim \hat{\mathcal{T}}^\pi\Upsilon(s,a)$ are all independent conditioned on the observed data $\mathcal{D}$ that determines $\hat{\mathcal{T}}^\pi$. With the above construction, we can estimate the objective function by

$$\hat{\bar{\mathcal{E}}}(\Upsilon, \hat{\mathcal{T}}^\pi\Upsilon) = \sum_{s,a} \hat{b}_\mu(s,a) \cdot \mathcal{E}\big\{\Upsilon(s,a), \hat{\mathcal{T}}^\pi\Upsilon(s,a)\big\}. \tag{12}$$

Now letting $\{\Upsilon_\theta : \theta \in \Theta\} \subseteq \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ be the hypothesis class of Υ_π, where each distribution Υ_θ is indexed by an element of candidate space Θ, a special case of which is the parametric case $\Theta \subseteq \mathbb{R}^p$. Then the proposed estimator of Υ_π is $\Upsilon_{\hat{\theta}}$ where

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \hat{\bar{\mathcal{E}}}(\Upsilon_\theta, \hat{\mathcal{T}}^\pi\Upsilon_\theta). \tag{13}$$

We call our method the *Energy Bellman Residual Minimizer* (EBRM) and summarize it in Algorithm 1. We will refer to the approach here as EBRM-single-step, as opposed to the multi-step extension EBRM-multi-step in Section 4.2.

Algorithm 1 EBRM-single-step

Input: $\Theta, \mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$
Output: $\hat{\theta}$
Estimate $\hat{b}_\mu$ and $\hat{p}$. ▷ Refer to Equation (10).
Compute $\hat{\theta} = \arg\min_{\theta\in\Theta} \hat{\bar{\mathcal{E}}}(\Upsilon_\theta, \hat{\mathcal{T}}^\pi\Upsilon_\theta)$. ▷ Refer to Equations (11) and (12).

3.2 Statistical Error Bound

In this subsection, we will provide a statistical error bound for EBRM-single-step. As shown in Table 1, most existing distributional OPE methods do not have a finite sample error bound for their estimators. To the best of our knowledge, the only exception is the very recent work named FLE (Wu et al., 2023), which is only able to analyze the marginal distribution of the return instead of conditional distributions of the return on each state-action pair studied in this paper. In passing, we also note that Rowland et al. (2018) also shows the consistency of their estimator, but no error bound analysis (and so convergence rate) is provided. We will first focus on the realizability setting and defer the analysis for the non-realizable case in Section 4.

Assumption 2. *There exists a unique $\theta \in \Theta$ such that $\Upsilon_\pi(s,a) = \Upsilon_\theta(s,a)$ for all $s, a \in \mathcal{S}\times\mathcal{A}$.*

Note that realizability is a generally weaker assumption than the widely-assumed completeness assumption (e.g., used in FLE (Wu et al., 2023)) which states that for all $\theta \in \Theta$, there exist $\theta' \in \Theta$ such that $\mathcal{T}^\pi\Upsilon_\theta = \Upsilon_{\theta'}$, in that it implies realizability due to $\Upsilon_\pi = \lim_{T\to\infty}(\mathcal{T}^\pi)^T\Upsilon_\theta$ under mild conditions. In contrast with non-realizability settings (Section 4), the realizability assumption aligns the minimizer of inaccuracy $\mathcal{E}(\Upsilon, \Upsilon_\pi)$ (best approximation) and the minimizer of Bellman residual, leading to stronger arguments and results.

Additionally, we make several mild assumptions regarding the transition probability p and the candidate space Θ, including the subgaussian rewards. A random variable (vector) $\mathbf{X}$ being subgaussian implies its tail probability decaying as fast as gaussian distribution (e.g., gaussian mixture, bounded random variable), quantified with finite subgaussian norm $\|\mathbf{X}\|_{\psi_2} < \infty$, as explained in Appendix A.4.

Assumption 3. *For any $\theta \in \Theta$, the random element $Z(s,a;\theta)$, which follows $\Upsilon_\theta(s,a)$, has finite expectation with respect to their norms, and the reward distribution are subgaussian, i.e.,*

$$\sup_{\theta\in\Theta}\sup_{s,a} \mathbb{E}\|Z(s,a;\theta)\| < \infty \quad \textit{and} \quad \sup_{s,a} \|R(s,a)\|_{\psi_2} < \infty.$$

Assumption 4. *The offline data* $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ *are iid draws from* $b_\mu \times p$.

Assumption 5. *There exists a metric* $\tilde{\eta}$ *over* $\mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ *such that* $\operatorname{diam}(\Theta; \tilde{\eta}) := \sup_{\theta_1,\theta_2\in\Theta} \tilde{\eta}(\theta_1, \theta_2) < \infty$, *where* $\tilde{\eta}(\theta_1, \theta_2) := \tilde{\eta}(\Upsilon_{\theta_1}, \Upsilon_{\theta_2})$. *For arbitrary* $c \in \mathbb{R}^d$, $\gamma_1, \gamma_2 \in [0, 1]$, $(s, a), (\tilde{s}, \tilde{a}) \in \mathcal{S} \times \mathcal{A}$, *letting* $Z_i(s, a) \sim \Upsilon_i(s, a)$ *be such that* $(Z_1(s, a), Z_3(s, a)) \in \mathbb{R}^d \times \mathbb{R}^d$ *and* $(Z_2(\tilde{s}, \tilde{a}), Z_4(\tilde{s}, \tilde{a})) \in \mathbb{R}^d \times \mathbb{R}^d$ *are mutually independent,* $\tilde{\eta}$ *should satisfy*

$$\left|\mathbb{E}\|c + \gamma_1 Z_1(s, a) - \gamma_2 Z_2(\tilde{s}, \tilde{a})\| - \mathbb{E}\|c + \gamma_1 Z_3(s, a) - \gamma_2 Z_4(\tilde{s}, \tilde{a})\|\right| \leq \gamma_1 \cdot \tilde{\eta}(\Upsilon_1, \Upsilon_3) + \gamma_2 \cdot \tilde{\eta}(\Upsilon_2, \Upsilon_4). \tag{14}$$

Supremum-extended Wasserstein-1 metric $\mathbb{W}_{1,\infty}$, which is shown to be a metric by Lemma 2 of Bellemare et al. (2017b), is an example that satisfies (14), as proved in Appendix A.5. Then we can obtain the convergence rate $O(\sqrt{\log(N/\delta)/N})$ as follows, with the exact finite-sample error bound demonstrated in Appendix A.6.7. Its proof can be found in Appendix A.6, and its special case for $\Theta \subseteq \mathbb{R}^p$ is covered in Corollary 3 of Appendix A.7.

Theorem 2. (Inaccuracy for realizable scenario) *Under Assumptions 1–5, for any* $\delta \in (0, 1)$, *given large enough sample size* $N \geq N(\delta)$, *our estimator* $\hat{\theta} \in \Theta$ *given by (13) satisfies the following bound with probability at least* $1 - \delta$,

$$\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_\pi) \lesssim \sqrt{\frac{1}{N} \log(\frac{(|\mathcal{S} \times \mathcal{A}| + N)}{\delta})}, \tag{15}$$

where $N(\delta)$ *depends on the complexity of* Θ *and* $\lesssim$ *means bounded by the given bound (RHS) multiplied by a positive number that does not depend on* N, *as defined in Appendix A.6.8.*

4 Non-realizable Settings

4.1 Combating Non-realizability With Multi-step Extensions

In the tabular case, most traditional OPE/RL methods do not suffer from model mis-specification and thus realizability always holds. In contrast, in DRL, as our target is to estimate the conditional distribution of return given any state-action pair, which is an infinite-dimensional object, non-realizability could still happen. Hence understanding and analyzing DRL methods for the tabular case under the non-realizable scenario is both important and challenging.

In the previous section under realizability, Theorem 1 played a fundamental role in our analysis. Indeed, Theorem 1 is valid regardless of realizability (Assumption 2), and essentially implies

$$0 \leq \min_{\theta\in\Theta} \bar{\mathcal{E}}(\Upsilon_\theta, \Upsilon_\pi) \leq \bar{\mathcal{E}}(\Upsilon_{\theta_*}, \Upsilon_\pi) \leq 2C_{\sup} B_1(\gamma; \beta_0) \cdot \bar{\mathcal{E}}(\Upsilon_{\theta_*}, \mathcal{T}^\pi \Upsilon_{\theta_*}), \tag{16}$$

where $\theta_* := \arg\min_{\theta\in\Theta} \bar{\mathcal{E}}(\Upsilon_\theta, \mathcal{T}^\pi \Upsilon_\theta)$. Violation of Assumption 2 (that is, non-realizability) implies nonzero value of $\bar{\mathcal{E}}(\Upsilon_{\theta_*}, \mathcal{T}^\pi \Upsilon_{\theta_*}) > 0$, and so Theorem 1 no longer ensures that θ_* has the smallest inaccuracy among $\theta \in \Theta$. Thus non-realizability may lead to the following mismatch:

$$\tilde{\theta} := \arg\min_{\theta\in\Theta} \bar{\mathcal{E}}(\Upsilon_\theta, \Upsilon_\pi) \neq \arg\min_{\theta\in\Theta} \bar{\mathcal{E}}(\Upsilon_\theta, \mathcal{T}^\pi \Upsilon_\theta) =: \theta_*. \tag{17}$$

Clearly, this mismatch is not due to sample variability, so it is unrealistic to hope that $\hat{\theta}$ defined by (13) would necessarily converge in probability to $\tilde{\theta}$ as $N \to \infty$.

To solve this issue, we propose a new approach. Temporarily ignoring mathematical rigor, the most important insight is that we can approximate $(\mathcal{T}^\pi)^m \Upsilon \approx \Upsilon_\pi$ with sufficiently large step level $m \in \mathbb{N}$. Thanks to the properties of energy distance, we have the following

$$\sup_{\theta\in\Theta} |\bar{\mathcal{E}}(\Upsilon_\theta, (\mathcal{T}^\pi)^m \Upsilon_\theta) - \bar{\mathcal{E}}(\Upsilon_\theta, \Upsilon_\pi)| \leq C\gamma^m, \quad \text{for some constant } C > 0. \tag{18}$$

(See Appendix C.2.9 under assumptions of Theorem 3.) As $m \to \infty$, the RHS of (18) shrinks to zero, making m-step Bellman residual $\bar{\mathcal{E}}(\Upsilon_\theta, (\mathcal{T}^\pi)^m \Upsilon_\theta)$ approximate the inaccuracy $\bar{\mathcal{E}}(\Upsilon_\theta, \Upsilon_\pi)$. This leads the two minimizers to be close, as illustrated schematically in Figure 1:

$$\theta_*^{(m)} := \arg\min_{\theta\in\Theta} \bar{\mathcal{E}}(\Upsilon_\theta, (\mathcal{T}^\pi)^m \Upsilon_\theta) \approx \arg\min_{\theta\in\Theta} \bar{\mathcal{E}}(\Upsilon_\theta, \Upsilon_\pi) =: \tilde{\theta} \text{ for large enough } m. \tag{19}$$

One can intuitively guess that larger step level m is required when the extent of non-realizability is large. Although multi-step idea has been widely employed for the purpose of improving sample efficiency particularly in traditional RL (e.g., Chen et al., 2021), ours is the first approach to use it in DRL for the purpose of overcoming non-realizability, to the best of our knowledge.

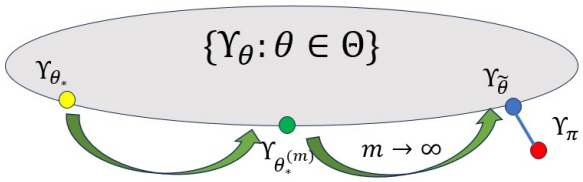

Figure 1: Larger m makes $(\mathcal{T}^\pi)^m \Upsilon_\theta \approx \Upsilon_\pi$ in Energy Distance, and thereby leads to $\theta_*^{(m)} \approx \tilde{\theta}$.

4.2 Bootstrap Operator

Generalizing from definition of $\hat{\mathcal{T}}^\pi$ based on (10), we consider $\hat{Z}^{(m)}(s,a;\theta) \sim (\hat{\mathcal{T}}^\pi)^m \Upsilon_\theta(s,a)$ as the distribution of an m-lengthed trajectories of tuples (s,a,r,s') that is generated under the estimated transition $\hat{p}$ and the target policy π:

$$\hat{Z}^{(m)}(s,a;\theta) \stackrel{D}{=} \sum_{t=1}^{m} \gamma^{t-1} \hat{R}^{(t)} + \gamma^m Z(\hat{S}^{(m)}, \hat{A}^{(m)}; \theta), \quad \text{where} \tag{20}$$

$$(\hat{R}^{(t)}, \hat{S}^{(t)}) \sim \hat{p}(\cdots | \hat{S}^{(t-1)}, \hat{A}^{(t-1)}) \quad \text{and} \quad \hat{A}^{(t)} \sim \pi(\cdot | \hat{S}^{(t)}) \quad \forall t \geq 1, \ (\hat{S}^{(0)}, \hat{A}^{(0)}) = (s,a).$$

Now we can define the estimated and the population Bellman residual, as well as the inaccuracy function, along with their minimizers as:

$$\hat{F}_m(\theta) := \hat{\bar{\mathcal{E}}}\big(\Upsilon_\theta, (\hat{\mathcal{T}}^\pi)^m \Upsilon_\theta\big), \quad F_m(\theta) := \bar{\mathcal{E}}\big(\Upsilon_\theta, (\mathcal{T}^\pi)^m \Upsilon_\theta\big), \quad F(\theta) := \bar{\mathcal{E}}\big(\Upsilon_\theta, \Upsilon_\pi\big), \tag{21}$$

$$\hat{\theta}^{(m)} := \arg\min_{\theta \in \Theta} \hat{F}_m(\theta), \quad \theta_*^{(m)} := \arg\min_{\theta \in \Theta} F_m(\theta), \quad \tilde{\theta} := \arg\min_{\theta \in \Theta} F(\theta).$$

However, the estimation of m-step Bellman operator (20) generally requires computation of N^m trajectories (as discussed in Appendix B.1), which amounts to a heavy computational burden.

To alleviate such burden, we will instead bootstrap $M \ll N^m$ many trajectories by first sampling the initial state-action pairs $(s_i^{(0)}, a_i^{(0)})$ $(1 \leq i \leq M)$ from $\hat{b}_\mu$ and then resampling the subsequent $r_i^{(t+1)}, s_i^{(t+1)} \sim \hat{p}(\cdots | s_i^{(t)}, a_i^{(t)})$ and $a_i^{(t+1)} \sim \pi(\cdot | s_i^{(t+1)})$ for m steps. Let $\hat{p}_m^{(B)}(\cdots | s,a)$ be the empirical probability measure of $(\sum_{t=1}^m \gamma^{t-1} r_i^{(t)}, s_i^{(m)})$ conditioning on $(s_i^{(0)}, a_i^{(0)}) = (s,a)$. We define the *bootstrap operator* as follows, with an abuse of notation $\mathcal{B}_m Z(s,a;\theta) \sim \mathcal{B}_m \Upsilon_\theta(s,a)$,

$$\mathcal{B}_m Z(s,a;\theta) :\stackrel{D}{=} \sum_{t=1}^{m} \gamma^{t-1} \hat{R}_b^{(t)} + \gamma^m Z(\hat{S}_b^{(m)}, \hat{A}_b^{(m)}; \theta), \tag{22}$$

$$\text{where} \quad (\sum_{t=1}^{m} \gamma^{t-1} \hat{R}_b^{(t)}, \hat{S}_b^{(m)}) \sim \hat{p}_m^{(B)}(\cdots | s,a) \quad \text{and} \quad \hat{A}_b^{(m)} \sim \pi(\cdot | \hat{S}_b^{(m)}).$$

Then we can compute our objective function and derive the bootstrap-based multi-step estimator.

$$\hat{F}_m^{(B)}(\theta) := \hat{\bar{\mathcal{E}}}\big(\Upsilon_\theta, \mathcal{B}_m \Upsilon_\theta\big) \quad \text{and} \quad \hat{\theta}_m^{(B)} := \arg\min_{\theta \in \Theta} \hat{F}_m^{(B)}(\theta). \tag{23}$$

We will refer to this method as EBRM-multi-step, whose procedure is summarized in Algorithm 2.

4.3 Statistical Error Bound

In this section, we develop a theoretical guarantee for $\bar{\mathcal{E}}\big(\Upsilon_{\hat{\theta}_m^{(B)}}, \Upsilon_{\tilde{\theta}}\big)$, where $\Upsilon_{\tilde{\theta}}$ is the best one we can achieve under the non-realizability. To proceed, we need to first deal with the parameter convergence from $\hat{\theta}_m^{(B)}$ to $\tilde{\theta}$, which relies on the following assumptions regarding the inaccuracy function $F(\cdot)$ (21), distance $\tilde{\eta}$ (Assumpion 5), and candidate space Θ.

Algorithm 2 EBRM-multi-step

Input: $\Theta, \mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N, m, M$
Output: $\hat{\theta}_m^{(B)}$
Estimate $\hat{b}_\mu$ and $\hat{p}$. ▷ Refer to Equation (10).
Randomly generate M tuples of $(\sum_{t=1}^m \gamma^{t-1} r_i^{(t)}, s_i^{(m)})$ $(1 \le i \le M)$.
$\hat{\theta}_m^{(B)} = \arg\min_{\theta\in\Theta} \hat{\bar{\mathcal{E}}}(\Upsilon_\theta, \mathcal{B}_m \Upsilon_\theta)$. ▷ Refer to Equations (22) and (23).

Assumption 6. *The inaccuracy function* (21) $F(\cdot) : \Theta \subset \mathbb{R}^p \to \mathbb{R}$ *has a unique minimizer* $\tilde{\theta}$*, and lower bounded by a polynomial of degree* $q \ge 1$*. That is, for all* $\theta \in \Theta$*, we have* $F(\theta) \ge F(\tilde{\theta}) + c_q \cdot \|\theta - \tilde{\theta}\|^q$ *for some constant* $c_q > 0$*.*

Assumption 7. *The candidate space* Θ *is compact (i.e.,* $\mathrm{diam}(\Theta; \|\cdot\|) < \infty$*). Furthermore, there exists* $L > 0$ *such that*

$$\tilde{\eta}(\theta_1, \theta_2) \le L\|\theta_1 - \theta_2\| \quad \textit{for} \quad \forall \theta_1, \theta_2 \in \Theta.$$

Assumption 8. $\tilde{\eta}$ *satisfies contractive property, i.e.,* $\tilde{\eta}(\mathcal{T}^\pi \Upsilon_1, \mathcal{T}^\pi \Upsilon_2) \le \gamma \cdot \tilde{\eta}(\Upsilon_1, \Upsilon_2)$*, where* $\mathcal{T}^\pi$ (2) *may correspond to an arbitrary transition* $p(\cdots|s, a)$*.*

Assumption 6 is used in quantifying the convergence rate. Compactness in Assumption 7 is for ensuring the existence of a minimizer of the estimated objective function (23), which is proved to be continuous in Appendix C.2.11. Compactness can be relaxed to "bounded" under mild conditions. Assumption 8 makes (18) feasible, and thereby shrinks the disparity between Bellman minimizer and the best approximation (19). This is satisfied by $\mathbb{W}_{1,\infty}$ that also satisfies property (14), as proved in Lemma 3 of Bellemare et al. (2017b).

Due to space constraints, we only present a simplified result below (proof in Appendix B.4), and a more detailed version of the finite-sample error bound for a fixed m is given in Appendix B.4.3.

Theorem 3. *Under Assumptions 1, 3–8, letting* $M = \lfloor C_1 \cdot N \rfloor$ *and* $m = \lfloor \frac{1}{4} \log_{(1/\gamma)}(C_2 N / \log N) \rfloor$ *for arbitrary constants* $C_1, C_2 > 0$*, we have the optimal convergence rate of the upper bound*

$$\bar{\mathcal{E}}\big(\Upsilon_{\hat{\theta}_m^{(B)}}, \Upsilon_{\tilde{\theta}}\big) \le \tilde{O}_p\left[\frac{1}{N^{1/(4q)}} \cdot \left\{\log_{\frac{1}{\gamma}}\left(\frac{N}{\log N}\right)\right\}^{2/q}\right],$$

where $\tilde{O}_p$ *indicates the rate of convergence up to logarithmic order.*

The convergence rate of Theorem 3 is the result of the (asymptotically) optimal choice of M and m. In our analysis, we notice a form of bias-variance trade-off in the selection of m, as explained in Appendix B.4.4. Practically, we set $M = N$ which works fine in the simulations of Section 5. A practical rule of m will be discussed in Section 4.4.

Note that the finite-sample error bound in Appendix B.4.3 is applicable to the setting with $m = 1$ and realizability assumption. For instance, assuming that the inaccuracy function $F(\theta)$ is lower-bounded by a quadratic polynomial $q = 2$, it gives us the bound $O[\{\log(N/\delta_1)/N\}^{1/8}]$ under the ideal case where we can ignore the last two sources of inaccuracy specified in Appendix B.4.4, each associated with bootstrap and non-realizability. We can see that it is much slower than the convergence rate $O(\sqrt{\log(N/\delta)/N})$ of Theorem 2, implying that it does not degenerate into Theorem 2. This is fundamentally due to a different proof structure that can be introduced via the application of Theorem 1 in the proof of Theorem 2, as intuitively explained in Appendix C.3.1. As explained earlier in Section 4.1, Theorem 1 can be used effectively to construct convergence of $\hat{\theta}$ under realizability.

4.4 Data-adaptive way of selecting step level

We need to choose m in practice. We will apply Lepski's rule (Lepskii, 1991). Since multi-step construction includes bootstrapping from the observed samples $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ (Section 4.2), this enables us to form a confidence interval. Starting from large enough m, we can decrease it until the intersection of the confidence intervals becomes a null set. To elaborate, given the data $\mathcal{D}$, we first generate multiple estimates of $\hat{\theta}_m^{(B)}$ (say $\hat{\theta}_{m,j}^{(B)}$ for $1 \le j \le J$), and calculate the disparity

from the single-step estimation which has no randomness once $\mathcal{D}$ is given, that is $\hat{\bar{\mathcal{E}}}(\Upsilon_{\hat{\theta}}, \Upsilon_{\hat{\theta}^{(B)}_{m,j}})$ $(1 \le j \le J)$. Then we calculate their means and standard deviations, forming a (Mean $\pm$ SD) interval for each m. Starting from a large enough value, we decrease m by one or more at a time, and select the m that makes the intersection become a null set $\cap_{k \ge m} I^{(k)} = \varnothing$. If we did not obtain the null set until $\cap_{k \ge 2} I^{(k)} \neq \varnothing$, then we use EBRM-single-step (13) without boostrap. Details are explained in Algorithm 3 of Appendix D.3.1.

5 Experiments

We assume a state space $\mathcal{S} = \{1, 2, \cdots, 30\}$ and an action space $\mathcal{A} = \{-1, 1\}$, each action representing left or right. With the details of the environment in Appendix D.2.1, the initial state distribution and behavior / target policies are

$$S \sim \text{Unif}\{1, 2, \cdots, 30\} \quad \text{and} \quad A \sim b(\cdot|S), \quad \text{where} \quad b(a|s) = 1/2 \quad \text{for} \quad \forall s, a \in \mathcal{S} \times \mathcal{A},$$
$$\pi(-1|s) = 0 \quad \text{and} \quad \pi(1|s) = 1 \quad \text{for} \quad \forall s \in \mathcal{S}. \tag{24}$$

We compare three methods: EBRM, FLE (Wu et al., 2023), and QRTD (Dabney et al., 2018b). Here, we assume realizability where the correct model is known (details in Appendix D.2.2) under two settings (with small and large variances), and the step level m for EBRM is chosen in a data-adaptive way in Section 4.4. With other tuning parameter selections explained in Appendix D.3, we repeated 100 simulations with the given sample size for each case, whose mean and standard deviation (within parenthesis) are recorded in Table 2. EBRM showed the lowest inaccuracy values measured by both $\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_{\pi})$ and $\overline{\mathbb{W}}_1(\Upsilon_{\hat{\theta}}, \Upsilon_{\pi})$, where $\overline{\mathbb{W}}_1$ indicates expectation-extended (6) Wasserstein-1 metric.

We also performed simulations in non-realizable scenarios (Appendix D.2.3) with more variety of sample sizes, and included Wasserstein-1 metric between marginal return distributions (Tables 8–13 of Appendix D.4). In most cases, EBRM showed outstanding performance.

Table 2: Mean $\bar{\mathcal{E}}$-inaccuracy (top) and $\overline{\mathbb{W}}_1$-inaccuracy (bottom) (standard deviation in parenthesis) over 100 simulations under realizability ($\gamma = 0.99$). Smallest inaccuracy values are in boldface.

	Small variance			**Large variance**		
Sample size	2000	5000	10000	2000	5000	10000
EBRM (Ours)	**0.046** (0.060)	**0.019** (0.022)	**0.008** (0.010)	**0.728** (0.920)	**0.301** (0.354)	**0.128** (0.167)
FLE	5.533 (6.448)	2.385 (2.883)	1.220 (1.618)	24.603 (25.768)	14.482 (16.101)	6.528 (7.814)
QRTD	48.679 (34.323)	46.032 (30.909)	49.402 (34.617)	105.274 (11.728)	75.173 (21.515)	70.483 (33.965)

	Small variance			**Large variance**		
Sample size	2000	5000	10000	2000	5000	10000
EBRM (Ours)	**1.339** (0.651)	**0.985** (0.388)	**0.782** (0.227)	**21.221** (10.337)	**15.532** (6.117)	**12.371** (3.595)
FLE	12.374 (7.843)	8.036 (5.091)	5.694 (3.773)	101.232 (58.586)	79.628 (46.772)	53.745 (33.948)
QRTD	56.739 (23.716)	54.397 (22.259)	57.145 (24.314)	274.405 (11.003)	236.383 (22.376)	223.537 (38.935)

6 Conclusion

In this paper, we justify the use of expectation-extended distances for Bellman residual minimization in DRL under general state-action space, based on which we propose a distributional OPE method called EBRM. We establish finite sample error bounds of the proposed estimator for the tabular case with or without realizability assumption. One interesting future direction is to extend EBRM to non-tabular case via linear MDP (e.g., Lazic et al., 2020; Bradtke & Barto, 1996), as we will briefly discuss in Appendix C.3.2.

REFERENCES

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017a.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017b.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22:33–57, 1996.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of off-policy td-learning via generalized bellman operators. *Advances in Neural Information Processing Systems*, 34:21440–21452, 2021.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pp. 1895–1904. PMLR, 2017.

Nevena Lazic, Dong Yin, Mehrdad Farajtabar, Nir Levine, Dilan Gorur, Chris Harris, and Dale Schuurmans. A maximum-entropy approach to off-policy evaluation in average-reward mdps. *Advances in Neural Information Processing Systems*, 33:12461–12471, 2020.

OV Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Thanh Tang Nguyen, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning with maximum mean discrepancy. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

Thanh Nguyen-Tang, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning via moment matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9144–9152, 2021.

Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 5528–5536. PMLR, 2019.

Bodhisattva Sen. A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 11:28–29, 2018.

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9196–9205. PMLR, 2020.

Ke Sun, Yingnan Zhao, Yi Liu, Wulong Liu, Bei Jiang, and Linglong Kong. Distributional reinforcement learning via sinkhorn iterations. *arXiv preprint arXiv:2202.00769*, 2022.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Jiayi Wang, Raymond KW Wong, and Xiaoke Zhang. Low-rank covariance function estimation for multidimensional functional data. *Journal of the American Statistical Association*, 117(538): 809–822, 2022.

Runzhe Wu, Masatoshi Uehara, and Wen Sun. Distributional offline policy evaluation with predictive error guarantees. *arXiv preprint arXiv:2302.09456*, 2023.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional reinforcement learning for multi-dimensional reward functions. *Advances in Neural Information Processing Systems*, 34:1519–1529, 2021.

A PROOFS FOR SECTIONS 2 AND 3

A.1 BOUNDING RADON-NIKODYM DERIVATIVE (25)

Let $t \in \mathbb{N}$ be arbitrary. Then we have the following for all $s, a \in \mathcal{S} \times \mathcal{A}$,

$$\frac{q_{b_\mu}^{\pi:t}(s,a)}{b_\mu(s,a)} \leq \int_{\mathcal{S}\times\mathcal{A}} q^\pi(s,a|\tilde{s},\tilde{a}) \cdot q_{b_\mu}^{\pi:t-1}(\tilde{s},\tilde{a}) \mathrm{d}\nu(\tilde{s},\tilde{a}) \cdot \frac{1}{b_\mu(s,a)} \leq \frac{p_{\max}}{p_{\min}} < \infty,$$

where $q_{b_\mu}^{\pi:0} = b_\mu$. Since $t \in \mathbb{N}$ was arbitrary, this implies existence of $C(t), C_{\sup} > 0$ such that

$$\sup_{s,a} \left\{ \frac{q_{b_\mu}^{\pi:t}(s,a)}{b_\mu(s,a)} \right\} \leq C(t) \leq C_{\sup} \leq \frac{p_{\max}}{p_{\min}} \text{ for } \forall\, t \in \mathbb{N}. \tag{25}$$

A.2 PROOF OF THEOREM 1

A.2.1 PROPERTIES OF DISTANCE

Property 1. *η satisfies translation-invariance and scale-sensitivity of order $\beta_0 > 0$. That is, with z being an arbitrary (nonrandom) constant of computable size and $c \in \mathbb{R}$,*

$$\eta\{\mathcal{L}(z+X), \mathcal{L}(z+Y)\} \leq \eta\{\mathcal{L}(X), \mathcal{L}(Y)\} \;\; \& \;\; \eta\{\mathcal{L}(cX), \mathcal{L}(cY)\} \leq |c|^{\beta_0} \eta\{\mathcal{L}(X), \mathcal{L}(Y)\}.$$

Property 2. *Letting $\mu(\cdot), \nu(\cdot) : \mathcal{Z} \to \mathcal{P}(\mathbb{R}^d)$ have different probability measures depending on the index random variable $Z \in \mathcal{Z}$ that follows a distribution $P(\cdot)$, the distance between probability-mixtures $\int_{\mathcal{Z}} \mu(z)\mathrm{d}P(z)$ and $\int_{\mathcal{Z}} \nu(z)\mathrm{d}P(z)$ satisfies convexity, that is*

$$\eta\left\{ \int_{\mathcal{Z}} \mu(z)\mathrm{d}P(z), \int_{\mathcal{Z}} \nu(z)\mathrm{d}P(z) \right\} \leq \int_{\mathcal{Z}} \eta\{\mu(z), \nu(z)\}\mathrm{d}P(z) = \mathbb{E}_{Z\sim P}[\eta\{\mu(Z), \nu(Z)\}].$$

Property 3. *It satisfies the following Relaxed Triangular Inequality for all integers $K \geq 2$,*

$$\eta(\mathcal{L}(X_0), \mathcal{L}(X_K)) \leq K \sum_{i=0}^{K-1} \rho^2(\mathcal{L}(X_i), \mathcal{L}(X_{i+1})) = K \sum_{i=0}^{K-1} \eta(\mathcal{L}(X_i), \mathcal{L}(X_{i+1})). \tag{26}$$

This is satisfied by all squared metric, that is $\eta(P,Q) = \rho^2(P,Q)$ for some probability metric ρ.

A.2.2 PROOF

Let $\Upsilon \in \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$ be arbitarily chosen. Starting with Relaxed Triangular Inequality (26) with $p = 2$, we obtain the following for an arbitrary $t \in \mathbb{N}$,

$$\eta\left\{ \Upsilon(s,a), \Upsilon_\pi(s,a) \right\} \leq 2\Bigg[\underbrace{\eta\left\{ \Upsilon(s,a), (\mathcal{T}^\pi)^t \Upsilon(s,a) \right\}}_{(a)} + \underbrace{\eta\left\{ (\mathcal{T}^\pi)^t \Upsilon(s,a), \Upsilon_\pi(s,a) \right\}}_{(b)} \Bigg]. \tag{27}$$

Let us first deal with (b). Define $P_t^\pi(\cdots|s,a)$ to be the probability measure of $(\sum_{i=1}^t \gamma^{i-1} R^{(i)}, S^{(t)}, A^{(t)})$ after t steps starting from the intial state-action pair $(S,A) = (s,a)$ under the given transition probability $(R^{(t)}, S^{(t)}) \sim p(\cdots|S^{(t-1)}, A^{(t-1)})$ and the target policy $A^{(t)} \sim \pi(\cdot|S^{(t)})$. Further denoting the probability measure of $y + \gamma^t Z(s^{(t)}, a^{(t)})$ with $Z(s,a) \sim \Upsilon(s,a)$ as $(g_{y,\gamma^t})_\# \Upsilon(s^{(t)}, a^{(t)})$ for the fixed value of $s^{(t)}, a^{(t)}$, $y = \sum_{i=1}^t \gamma^{i-1} r^{(i)}$

(which aligns with the notation $(g_{r,\gamma})_{\#}$ in (2)), we can obtain the following,

$$
\begin{aligned}
(b) &= \eta\Big\{(\mathcal{T}^\pi)^t\Upsilon(s,a), (\mathcal{T}^\pi)^t\Upsilon_\pi(s,a)\Big\} \quad \because (\mathcal{T}^\pi)^t\Upsilon_\pi(s,a) = \Upsilon_\pi(s,a) \\
&= \eta\Bigg\{ \int (g_{y,\gamma^t})_{\#}\Upsilon(s^{(t)},a^{(t)})\mathrm{d}P_t^\pi(y,s^{(t)},a^{(t)}|s,a), \\
&\qquad\qquad \int (g_{y,\gamma^t})_{\#}\Upsilon_\pi(s^{(t)},a^{(t)})\mathrm{d}P_t^\pi(y,s^{(t)},a^{(t)}|s,a)\Bigg\} \\
&\leq \int \eta\big\{(g_{y,\gamma^t})_{\#}\Upsilon(s^{(t)},a^{(t)}), (g_{y,\gamma^t})_{\#}\Upsilon_\pi(s^{(t)},a^{(t)})\big\}\mathrm{d}P_t^\pi(y,s^{(t)},a^{(t)}|s,a) \text{ by Property 2} \\
&\leq \int \gamma^{t\beta_0}\eta\Big\{\Upsilon(s^{(t)},a^{(t)}), \Upsilon_\pi(s^{(t)},a^{(t)})\Big\}\mathrm{d}P_t^\pi(s^{(t)},a^{(t)}|s,a) \text{ by Property 1} \\
&= \gamma^{t\beta_0}\mathbb{E}_\pi\Big[\eta\Big\{\Upsilon(S^{(t)},A^{(t)}), \Upsilon_\pi(S^{(t)},A^{(t)})\Big\}\Big|S=s,A=a\Big]. \qquad (28)
\end{aligned}
$$

From now on, we will use $\mathbb{E}_{b_\mu}$ and $\mathbb{E}_{q_{b_\mu}^{\pi:t}}$ to denote the expectation with respect to the probability $(S,A)\sim b_\mu$ and $(S,A)\sim q_{b_\mu}^{\pi:t}$. Treating (S,A) in Inequality (27) as random, we can obtain the following based on Inequality (28),

$$
\begin{aligned}
\mathbb{E}_{b_\mu}\Big[(b)\Big] &\leq \gamma^{t\beta_0}\mathbb{E}\Big[\eta\Big\{\Upsilon(S^{(t)},A^{(t)}), \Upsilon_\pi(S^{(t)},A^{(t)}))\Big\}\Big|(S,A)\sim b_\mu,\ A^{(i)}\sim\pi(\cdot|S^{(i)})\ \forall i\geq 1\Big] \\
&= \gamma^{t\beta_0}\mathbb{E}_{b_\mu}\Big[\mathbb{E}_\pi\Big\{\eta\Big(\Upsilon(S^{(t)},A^{(t)}), \Upsilon_\pi(S^{(t)},A^{(t)})\Big)\Big|S,A\Big\}\Big] \\
&= \gamma^{t\beta_0}\mathbb{E}_{q_{b_\mu}^{\pi:t}}\Big[\eta\Big\{\Upsilon(S,A), \Upsilon_\pi(S,A)\Big\}\Big] \text{ by definition of } q_{b_\mu}^{\pi:t} \text{ under Assumption 1} \\
&= \gamma^{t\beta_0}\mathbb{E}_{b_\mu}\Big[\eta\Big\{\Upsilon(S,A), \Upsilon_\pi(S,A)\Big\}\cdot\frac{q_{b_\mu}^{\pi:t}(S,A)}{b_\mu(S,A)}\Big] \\
&\leq C(t)\cdot\gamma^{t\beta_0}\cdot\bar{\eta}(\Upsilon,\Upsilon_\pi) \text{ by Inequality (25).} \qquad (29)
\end{aligned}
$$

Now let us deal with (a) of Inequality (27) using Property 3. Let $\Upsilon_k(s,a) = (\mathcal{T}^\pi)^k\Upsilon(s,a)$. For sufficiently large $t\in\mathbb{N}$, we obtain the following by relaxed triangle inequality (26) with $K=2$,

$$
\begin{aligned}
&\eta(\Upsilon(s,a), (\mathcal{T}^\pi)^t\Upsilon(s,a)) \\
&\quad\leq 2\cdot\Big\{\eta(\Upsilon_0(s,a),\Upsilon_1(s,a)) + \eta(\Upsilon_1(s,a),\Upsilon_t(s,a))\Big\} \\
&\quad\leq 2\cdot\eta(\Upsilon_0(s,a),\Upsilon_1(s,a)) + 2\cdot 2\cdot\Big\{\eta(\Upsilon_1(s,a),\Upsilon_3(s,a)) + \eta(\Upsilon_3(s,a),\Upsilon_t(s,a))\Big\} \\
&\quad\leq 2\cdot\eta(\Upsilon_0(s,a),\Upsilon_1(s,a)) + 2^2\cdot\eta(\Upsilon_1(s,a),\Upsilon_3(s,a)) \\
&\qquad\qquad + 2^2\cdot 2\cdot\Big\{\eta(\Upsilon_3(s,a),\Upsilon_7(s,a)) + \eta(\Upsilon_7(s,a),\Upsilon_t(s,a))\Big\} \\
&\quad\leq 2\cdot\eta(\Upsilon_0(s,a),\Upsilon_1(s,a)) + 2^2\cdot\eta(\Upsilon_1(s,a),\Upsilon_3(s,a)) + 2^3\cdot\eta(\Upsilon_3(s,a),\Upsilon_7(s,a)) + \cdots.
\end{aligned}
$$

This can be further bounded as follows by relaxed triangle inequality (26), this time with general $K\geq 2$,

$$
\begin{aligned}
\eta(\Upsilon(s,a), (\mathcal{T}^\pi)^t\Upsilon(s,a)) &\leq 2\cdot\eta(\Upsilon_0(s,a),\Upsilon_1(s,a)) \\
&\quad + 2^2\cdot 2\cdot\big\{\eta(\Upsilon_1(s,a),\Upsilon_2(s,a)) + \eta(\Upsilon_2(s,a),\Upsilon_3(s,a))\big\} \\
&\quad + 2^3\cdot 2^2\cdot\big\{\eta(\Upsilon_3(s,a),\Upsilon_4(s,a)) + \cdots + \eta(\Upsilon_6(s,a),\Upsilon_7(s,a))\big\} \\
&\quad + \cdots,
\end{aligned}
$$

which can finally be formalized into

$$
(a) \leq \sum_{k=1}^{\infty} 2^{2k-1}\cdot\sum_{j=0}^{s_1(k)}\eta\Big(\Upsilon_{s_1(k)+j}(s,a), \Upsilon_{s_1(k)+j+1}(s,a)\Big) \text{ where } s_1(k) = 2^{k-1}-1.
$$

Therefore we can further obtain following using similar logic as Inequality (29),

$$\mathbb{E}_{b_\mu}\Big[(a)\Big] \leq \sum_{k=1}^{\infty} 2^{2k-1} \cdot \sum_{j=0}^{s_1(k)} \mathbb{E}_{b_\mu}\Big[\eta\Big\{\Upsilon_{s_1(k)+j}(S,A), \Upsilon_{s_1(k)+j+1}(S,A)\Big\}\Big]$$
$$\leq \sum_{k=1}^{\infty} 2^{2k-1} \cdot \sum_{j=0}^{s_1(k)} \mathbb{E}_{b_\mu}\Big[\eta\Big\{(\mathcal{T}^\pi)^{s_1(k)+j}\Upsilon(S,A), (\mathcal{T}^\pi)^{s_1(k)+j+1}\Upsilon(S,A)\Big\}\Big]$$
$$\leq \sum_{k=1}^{\infty} 2^{2k-1} \cdot \sum_{j=0}^{s_1(k)} \gamma^{(s_1(k)+j)\beta_0} \cdot \mathbb{E}_{q^{b:\pi}_{T:s_1(k)+j}}\Big[\eta\Big\{\Upsilon(S,A), \mathcal{T}^\pi\Upsilon(S,A)\Big\}\Big]$$
$$\leq \sum_{k=1}^{\infty} 2^{2k-1} \cdot \sum_{j=0}^{s_1(k)} \gamma^{(s_1(k)+j)\beta_0} \cdot C(s_1(k)+j) \cdot \mathbb{E}_{b_\mu}\Big[\eta\Big\{\Upsilon(S,A), \mathcal{T}^\pi\Upsilon(S,A)\Big\}\Big]$$
$$= \sum_{k=1}^{\infty} 2^{2k-1} \cdot \sum_{j=0}^{2^{k-1}-1} \gamma^{(2^{k-1}-1+j)\beta_0} \cdot C(2^{k-1}-1+j) \cdot \mathbb{E}_{b_\mu}\Big[\eta\Big\{\Upsilon(S,A), \mathcal{T}^\pi\Upsilon(S,A)\Big\}\Big]. \quad (30)$$

Note that we have

$$B(\gamma;\beta_0) = \sum_{k=1}^{\infty} 2^{2k-1} \cdot \sum_{j=0}^{2^{k-1}-1} \gamma^{(2^{k-1}-1+j)\beta_0} = \sum_{k=1}^{\infty} 2^{2k-1} \cdot \gamma^{(2^{k-1}-1)\beta_0} \cdot \sum_{j=0}^{2^{k-1}-1} (\gamma^{\beta_0})^j$$
$$\leq \left(\sum_{k=1}^{\infty} 2^{2k-1} \cdot \gamma^{(2^{k-1}-1)\beta_0}\right) \cdot \left(\sum_{j=0}^{\infty} (\gamma^{\beta_0})^j\right) \leq \frac{1}{2(1-\gamma^{\beta_0})} \sum_{k=1}^{\infty} (4^k \cdot \gamma^{(2^{k-1}-1)\beta_0}) < \infty, \quad (31)$$

and Inequalities (29) and (30) can thereby be switched into the following bound, since $C(t) \leq C_{\text{sup}}$ by Inequality (25),

$$\mathbb{E}_{b_\mu}\Big[(a)\Big] \leq C_{\text{sup}} \cdot B(\gamma;\beta_0) \cdot \bar{\eta}(\Upsilon, \mathcal{T}^\pi\Upsilon) \;\; \& \;\; \mathbb{E}_{b_\mu}\Big[(b)\Big] \leq C_{\text{sup}} \cdot \gamma^{t\beta_0}\bar{\eta}(\Upsilon, \Upsilon_\pi).$$

Then starting from Inequality (27), we can obtain

$$\bar{\eta}(\Upsilon, \Upsilon_\pi) \leq 2 \cdot \Big\{\mathbb{E}_{b_\mu}\Big[(a)\Big] + \mathbb{E}_{b_\mu}\Big[(b)\Big]\Big\} \leq 2C_{\text{sup}} \cdot \Big\{B(\gamma;\beta_0) \cdot \bar{\eta}(\Upsilon, \mathcal{T}^\pi\Upsilon) + \gamma^{t\beta_0}\bar{\eta}(\Upsilon, \Upsilon_\pi)\Big\}.$$

Letting $t \to \infty$, we obtain the desired result as follows,

$$\bar{\eta}(\Upsilon, \Upsilon_\pi) \leq 2C_{\text{sup}}B(\gamma;\beta_0) \cdot \bar{\eta}(\Upsilon, \mathcal{T}^\pi\Upsilon).$$

Replacing $B(\gamma;\beta_0)$ with $B_1(\gamma;\beta_0) := \frac{1}{2(1-\gamma^{\beta_0})} \sum_{k=1}^{\infty} (4^k \cdot \gamma^{(2^{k-1}-1)\beta_0})$ as (31) gives us the desired result of Theorem 1.

A.3 Proof that Energy Distance satisfies Properties A.2.1

Property 1 is straightforward from the definition of Energy Distance (8). For an arbitrary $\mathbf{c} \in \mathbb{R}^d$ and $c \in \mathbb{R}$, we have the following that leads to $\beta_0 = 1$ in Property 1,

$$\mathcal{E}\big\{\mathcal{L}(\mathbf{c}+\mathbf{X}), \mathcal{L}(\mathbf{c}+\mathbf{Y})\big\} = \mathcal{E}\big\{\mathcal{L}(\mathbf{X}), \mathcal{L}(\mathbf{Y})\big\} \quad \& \quad \mathcal{E}\big\{\mathcal{L}(c\mathbf{X}), \mathcal{L}(c\mathbf{Y})\big\} = |c| \cdot \mathcal{E}\big\{\mathcal{L}(\mathbf{X}), \mathcal{L}(\mathbf{Y})\big\}.$$

Property 2 is shown by Lemma 3 of Nguyen-Tang et al. (2021).

Property 3 can be verified as follows. Since $\mathcal{E}$ is a squared MMD_k corresponding to the kernel $k(\mathbf{x},\mathbf{y}) = \|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x}-\mathbf{y}\|$, it is a squared form of some metric ρ between two distributions P, Q by Lemma 4 of Gretton et al. (2012),

$$\mathcal{E}(P,Q) = \rho^2(P,Q) \;\; \text{where} \;\; \rho(P,Q) := \|\mu_P - \mu_Q\|_{\mathcal{H}}$$

where $\mu_P, \mu_Q \in \mathcal{H}$ are the mean embeddings of P, Q, and $\mathcal{H}$ is the RKHS corresponding to the suggested kernel k. Based on this, we can derive the so-called Relaxed Triangular Inequality,

$$\begin{aligned}\mathcal{E}(\mathcal{L}(X_0), \mathcal{L}(X_K)) = \rho^2(\mathcal{L}(X_0), \mathcal{L}(X_K)) &\leq \left\{ \sum_{i=0}^{K-1} \rho(\mathcal{L}(X_i), \mathcal{L}(X_{i+1})) \right\}^2 \\ &\leq K \sum_{i=0}^{K-1} \rho^2(\mathcal{L}(X_i), \mathcal{L}(X_{i+1})) = K \sum_{i=0}^{K-1} \mathcal{E}(\mathcal{L}(X_i), \mathcal{L}(X_{i+1})),\end{aligned}$$

where the inequality of the second line used $ab + ba \leq a^2 + b^2$ that leads to $(a_1 + \cdots a_K)^2 \leq K \cdot (a_1^2 + \cdots a_K^2)$. Plugging in $K = 2$ gives us the following special case,

$$\mathcal{E}\{\mathcal{L}(X_0), \mathcal{L}(X_2)\} \leq 2 \cdot [\mathcal{E}\{\mathcal{L}(X_0), \mathcal{L}(X_1)\} + \mathcal{E}\{\mathcal{L}(X_1), \mathcal{L}(X_2)\}]. \tag{32}$$

A.4 Explanation of Subgaussian norm

Subgaussianity can be quantified with subgaussian norm $\|\cdot\|_{\psi_2} : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ or $\mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ (Definition 2.5.6 and Definition 3.4.1 of Vershynin (2018)),

$$\|X\|_{\psi_2} := \inf \left\{ t > 0 : \mathbb{E}(X^2/t^2) \leq 2 \right\} (d = 1) \ \& \ \|\mathbf{X}\|_{\psi_2} := \sup_{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1} \|\langle \mathbf{X}, \mathbf{x} \rangle\|_{\psi_2}. \tag{33}$$

Subgaussian norm is verified to be a valid norm in Exercise 2.5.7 of (Vershynin, 2018). Random variable (vector) $\mathbf{X}$ is called subgaussian if it satisfies $\|\mathbf{X}\|_{\psi_2} < \infty$. A lot of useful inequalilties, such as Dudley's integral inequality and and Hoeffding's inequality (Theorems 4, 5) are based on subgaussianity assumption.

A.5 Why supremum-extended Wasserstein-1 metric satisfies (14)

Let $c \in \mathbb{R}^d$, $\gamma_1, \gamma_2 \in (0, 1]$, and $(s, a), (\tilde{s}, \tilde{a}) \in \mathcal{S} \times \mathcal{A}$ be arbitrary, $Z_i(s, a) \sim \Upsilon_i(s, a)$ with $Z_1(s, a)$, $Z_2(\tilde{s}, \tilde{a})$ and $Z_3(s, a)$, $Z_4(\tilde{s}, \tilde{a})$ being pairwise independent. Letting J_{13} be the possible dependence structures between $Z_1(s, a)$ and $Z_2(s, a)$, and J_{24} be that between $Z_3(\tilde{s}, \tilde{a})$ and $Z_4(\tilde{s}, \tilde{a})$, we have

$$\begin{aligned}&\left| \mathbb{E}\|c + \gamma_1 Z_1(s, a) - \gamma_2 Z_2(\tilde{s}, \tilde{a})\| - \mathbb{E}\|c + \gamma_1 Z_3(s, a) - \gamma_2 Z_4(\tilde{s}, \tilde{a})\| \right| \\ &\quad = \inf_{J_{13}, J_{24}} \left| \mathbb{E}\|c + \gamma_1 Z_1(s, a) - \gamma_2 Z_2(\tilde{s}, \tilde{a})\| - \mathbb{E}\|c + \gamma_1 Z_3(s, a) - \gamma_2 Z_4(\tilde{s}, \tilde{a})\| \right| \\ &\quad \leq \inf_{J_{13}, J_{24}} \mathbb{E}\|c + \gamma_1 Z_1(s, a) - \gamma_2 Z_2(\tilde{s}, \tilde{a}) - c - \gamma_1 Z_3(s, a) + \gamma_2 Z_4(\tilde{s}, \tilde{a})\| \\ &\quad = \inf_{J_{13}} \mathbb{E}\|\gamma_1 Z_1(s, a) - \gamma_1 Z_3(s, a)\| + \inf_{J_{24}} \mathbb{E}\|\gamma_2 Z_2(\tilde{s}, \tilde{a}) - \gamma_2 Z_4(\tilde{s}, \tilde{a})\| \\ &\quad = \gamma_1 \cdot \mathbb{W}_1(Z_1(s, a), Z_3(s, a)) + \gamma_2 \cdot \mathbb{W}_1(Z_2(s, a), Z_4(s, a)) \\ &\quad \leq \gamma_1 \cdot \mathbb{W}_{1,\infty}(\Upsilon_1, \Upsilon_3) + \gamma_2 \cdot \mathbb{W}_{1,\infty}(\Upsilon_2, \Upsilon_4),\end{aligned}$$

where the second last line holds by the definition of Wasserstein-1 metric.

A.6 PROOF OF THEOREM 2

Throughout the proof, we will use $C > 0, C_k > 0$ ($k \in \mathbb{N}$) to denote appropriate universal constants.

Based on the metric $\tilde{\eta}$ in Assumption 5, we will quantify the model complexity by *covering number* (Definition 4.2.2 of Vershynin (2018)). With $N_{\tilde{\eta}}(\theta, t) := \{\theta' \in \Theta : \tilde{\eta}(\theta', \theta) < t\}$ being t-neighborhood of $\theta \in \Theta$, we define the covering number as follows,

$$\mathcal{N}(\Theta, \tilde{\eta}, t) := \min\left\{\tilde{M} \in \mathbb{N} : \exists \theta_1, \cdots, \theta_{\tilde{M}} \text{ s.t. } \Theta \subset \bigcup_{i=1}^{\tilde{M}} N_{\tilde{\eta}}(\theta_i, t)\right\}. \tag{34}$$

Also define the following minimizer of Bellman residual,

$$\theta_* :\in \arg\min_{\theta \in \Theta} \bar{\mathcal{E}}(\Upsilon_\theta, \mathcal{T}^\pi \Upsilon_\theta).$$

Since $\Upsilon_\pi = \Upsilon_{\theta'}$ for some $\theta' \in \Theta$ by Assumption 2, we have $\bar{\mathcal{E}}(\Upsilon_{\theta'}, \mathcal{T}^\pi \Upsilon_{\theta'}) = \bar{\mathcal{E}}(\Upsilon_\pi, \mathcal{T}^\pi \Upsilon_\pi) = 0$, thereby becoming the minimizer of Bellman residual. Then we can let $\theta_* = \theta'$, and have $\bar{\mathcal{E}}(\Upsilon_{\theta_*}, \mathcal{T}^\pi \Upsilon_{\theta_*}) = 0$, that is $\Upsilon_{\theta_*} = \mathcal{T}^\pi \Upsilon_{\theta_*}$.

A.6.1 DECOMPOSITION INTO TWO DISCREPANCIES

Defining Γ_N and Δ_N as

$$\Gamma_N := \sup_{\theta \in \Theta} \bar{\mathcal{E}}(\mathcal{T}^\pi \Upsilon_\theta, \hat{\mathcal{T}}^\pi \Upsilon_\theta) \;\; \& \;\; \Delta_N := \sup_{\theta \in \Theta} \left| \bar{\mathcal{E}}(\Upsilon_\theta, \hat{\mathcal{T}}^\pi \Upsilon_\theta) - \hat{\bar{\mathcal{E}}}(\Upsilon_\theta, \hat{\mathcal{T}}^\pi \Upsilon_\theta) \right|,$$

we can decompose the term $\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \mathcal{T}^\pi \Upsilon_{\hat{\theta}})$ as follows.

$$\begin{aligned}
\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \mathcal{T}^\pi \Upsilon_{\hat{\theta}}) &\le 2 \cdot \left\{ \bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \hat{\mathcal{T}}^\pi \Upsilon_{\hat{\theta}}) + \bar{\mathcal{E}}(\hat{\mathcal{T}}^\pi \Upsilon_{\hat{\theta}}, \mathcal{T}^\pi \Upsilon_{\hat{\theta}}) \right\} \\
&\le 2 \cdot \left\{ \hat{\bar{\mathcal{E}}}(\Upsilon_{\hat{\theta}}, \hat{\mathcal{T}}^\pi \Upsilon_{\hat{\theta}}) + \left| \hat{\bar{\mathcal{E}}}(\Upsilon_{\hat{\theta}}, \hat{\mathcal{T}}^\pi \Upsilon_{\hat{\theta}}) - \bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \hat{\mathcal{T}}^\pi \Upsilon_{\hat{\theta}}) \right| + \bar{\mathcal{E}}(\hat{\mathcal{T}}^\pi \Upsilon_{\hat{\theta}}, \mathcal{T}^\pi \Upsilon_{\hat{\theta}}) \right\} \\
&\le 2 \cdot \left\{ \hat{\bar{\mathcal{E}}}(\Upsilon_{\theta_*}, \hat{\mathcal{T}}^\pi \Upsilon_{\theta_*}) + \Delta_N + \Gamma_N \right\} \\
&\le 2 \cdot \left\{ \bar{\mathcal{E}}(\Upsilon_{\theta_*}, \hat{\mathcal{T}}^\pi \Upsilon_{\theta_*}) + \left| \hat{\bar{\mathcal{E}}}(\Upsilon_{\theta_*}, \hat{\mathcal{T}}^\pi \Upsilon_{\theta_*}) - \bar{\mathcal{E}}(\Upsilon_{\theta_*}, \hat{\mathcal{T}}^\pi \Upsilon_{\theta_*}) \right| + \Delta_N + \Gamma_N \right\} \\
&\le 2 \cdot \left\{ \bar{\mathcal{E}}(\mathcal{T}^\pi \Upsilon_{\theta_*}, \hat{\mathcal{T}}^\pi \Upsilon_{\theta_*}) + 2\Gamma_N + \Delta_N \right\} \\
&\le 4 \cdot (\Gamma_N + \Delta_N).
\end{aligned}$$

Combined with Theorem 1 that requires Assumption 1, it leads to the following bound,

$$\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_\pi) \le 8 C_{\sup} B_1(\gamma) \cdot (\Gamma_N + \Delta_N), \quad \text{where} \tag{35}$$

$$B_1(\gamma) := B_1(\gamma; 1) = \frac{1}{2(1-\gamma)} \sum_{k=1}^{\infty} 4^k \gamma^{2^{k-1}-1} \quad \text{by definition in Theorem 1}$$

since we have verified $\beta_0 = 1$ in A.3. Now it suffices to bound Γ_N and Δ_N, which will be referred to as *Bellman discrepancy* and *state-action discrepancy* to indicate the sources of error, $\mathcal{T}^\pi$ and $b_\mu(s, a)$, respectively. Before we proceed, we list several properties of subgaussian norm (33) that we will utilize in our analysis. Corresponding proofs can be found in Section C.2.1, which are mostly based on Vershynin (2018).

Remark 1. *(Properties of sub-gaussian norm) We have the following properties regarding subgaussian norm,*

1. *For $X \sim Ber(p)$, we have $\|X\|_{\psi_2} \le 1/\sqrt{\log 2}$.*
2. *For a constant $c \in \mathbb{R}$, $\|c\|_{\psi_2} = c/\sqrt{\log 2}$.*

3. *For a random variable* $X \in \mathbb{R}$, $\|\mathbb{E}(X)\|_{\psi_2} \leq \|X\|_{\psi_2}$ *holds.*
4. *For a random vector* $\mathbf{X} \in \mathbb{R}^d$, $\big\|\|\mathbf{X}\|\big\|_{\psi_2} \leq d\big\|\mathbf{X}\big\|_{\psi_2}$ *holds.*
5. *For a random variable* $X \in \mathbb{R}$, $\|X - \mathbb{E}(X)\|_{\psi_2} \leq C\|X\|_{\psi_2}$ *holds.*

A.6.2 Conditioning on Sufficient Sample size for each state-action pair

Prior to bounding Γ_N and Δ_N of (35), we will first condition upon an event where each state-action pair is observed *sufficiently many times*.

Before we proceed, we should note that the given probability space $(\Omega, \Sigma, \mathbb{P})$ can be factorized into two stages. Letting $\mathbf{N} = (N(s,a))_{s,a \in \mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ to be a random vector that indicates the observed number of samples for each state-action pair, we can see that $(\Omega, \Sigma, \mathbb{P})$ consists of two consecutive probability events denoted as follows,

$$\text{Stage 1: } (\Omega_{\mathcal{S}\times\mathcal{A}}, \Sigma_{\mathcal{S}\times\mathcal{A}}, \mathbb{P}_{\mathcal{S}\times\mathcal{A}}) \Rightarrow \text{determines which state-action pairs } S_i, A_i \text{ are sampled,} \quad (36)$$

$$\text{Stage 2: } (\Omega^{(\mathbf{N})}, \Sigma^{(\mathbf{N})}, \mathbb{P}^{(\mathbf{N})}) \Rightarrow \text{conditioned on } (S_i, A_i), \text{ determines } R_i, S_i' \sim p(\cdots|S_i, A_i).$$

This implies that having sufficiently many observations for each s, a is solely associated with probability space of Stage 1. Now let us discuss how "sufficiently large" $N(s,a)$ is characterized (38).

Since we are given with $N \geq 2$, we can divide the data $\mathcal{D} = \{(S_i, A_i, R_i, S_i')\}_{i=1}^N$ into two halves,

$$\mathcal{D}_1 = \{(S_i, A_i, R_i, S_i')\}_{i=1}^{\lfloor N/2 \rfloor} \quad \text{and} \quad \mathcal{D}_2 = \{(S_i, A_i, R_i, S_i')\}_{i=\lfloor N/2 \rfloor + 1}^{N}.$$

Note that we denoted observations (S_i, A_i, R_i, S_i') in capital letters, so as to indicate that they are random objects. Based on this, we define the following notations based on (10),

$$\mathbf{p} = \big(b_\mu(s,a)\big)_{s,a\in\mathcal{S}\times\mathcal{A}} \in [0,1]^{\mathcal{S}\times\mathcal{A}} \;\;\&\;\; \hat{\mathbf{p}} = \big(\hat{b}_\mu(s,a)\big)_{s,a\in\mathcal{S}\times\mathcal{A}} \in [0,1]^{\mathcal{S}\times\mathcal{A}},$$

and it is straightforward to see

$$\hat{\mathbf{p}} = \frac{\lfloor N/2 \rfloor}{N}\hat{\mathbf{p}}_{(1)} + \frac{N - \lfloor N/2 \rfloor}{N}\hat{\mathbf{p}}_{(2)}, \quad (37)$$

where each term in the RHS of (37) is sample mean based on $\mathcal{D}_1$ and $\mathcal{D}_2$,

$$\hat{\mathbf{p}}_{(1)} = \frac{1}{N/2}\sum_{i=1}^{N/2}\mathbf{y}_i \quad \text{and} \quad \hat{\mathbf{p}}_{(2)} = \frac{1}{N - \lfloor N/2 \rfloor}\sum_{i=1}^{N-\lfloor N/2 \rfloor}\mathbf{y}_i,$$

$$\mathbf{y}_i \in \big\{(1,0,\cdots,0)^\intercal, (0,1,\cdots,0)^\intercal, \cdots, (0,0,\cdots,1)^\intercal\big\} \subset [0,1]^{|\mathcal{S}\times\mathcal{A}|}$$

with $\mathbf{y}_i$ being indicators that represent where each observation (S_i, A_i) belongs to. Within Stage 1 probability space (36), we define the following subset with given $\epsilon \in (0,1]$,

$$\Omega_{\mathcal{S}\times\mathcal{A}}^{(\epsilon)} := \left\{\omega \in \Omega_{\mathcal{S}\times\mathcal{A}} \;\middle|\; \|\hat{\mathbf{p}}_{(1)} - \mathbf{p}\| < \frac{1}{2}p_{\min}\cdot\epsilon \quad \text{and} \quad \|\hat{\mathbf{p}}_{(2)} - \mathbf{p}\| < \frac{1}{2}p_{\min}\cdot\epsilon\right\}, \quad (38)$$

under which we can verify that following holds (proofs in C.2.2),

$$\text{Fact 1: } \hat{b}_\mu(s,a) = \frac{N(s,a)}{N} \in \left[\frac{1}{2}b_\mu(s,a), \frac{3}{2}b_\mu(s,a)\right] \text{ for } \forall s,a \in \mathcal{S}\times\mathcal{A}, \quad (39)$$

$$\text{Fact 2: } N(s,a) \geq 2 \text{ for } \forall s,a \in \mathcal{S}\times\mathcal{A},$$

$$\text{Fact 3: } \|\hat{\mathbf{p}} - \mathbf{p}\| < \frac{1}{2}p_{\min}\cdot\epsilon.$$

There is one fact which is crucially important about (38). The conditioned event of observing a plenty of samples for each s, a (38) is not related at all with Stage 2 probability space (36). This implies that regardless of realizations of $\mathbf{N}$, the dependence structure between different samples (conditioned on the same s, a) $(R, S') \sim p(\cdots|s,a)$ remains intact, i.e. R_i, S_i' $(1 \leq i \leq N)$ remain independent with respect to Stage 2 probability measure $\mathbb{P}^{(\mathbf{N})}$ (36).

Throughout the following subsections A.6.3 and A.6.4 where we shall bound Γ_N and Δ_N, we will resort to conditional probability measure $\mathbb{P}^{(\mathbf{N})}(\cdots) := \mathbb{P}(\cdots|\mathbf{N})$ along with its corresponding subgaussian norm $\|\cdot\|_{\psi_2(\mathbf{N})}$. In other words, we will consider $N(s,a)$ to be fixed (non-random), assuming that Facts (39) are satisfied, and later calculate its unconditional probability with $\mathbb{P}$ in A.6.5 by Inequality (75).

A.6.3 BOUNDING BELLMAN DISCREPANCY

Once more, we would like to emphasize that $N(s,a)$ are fixed, and Facts (39) hold. The probability space we are dealing with in this subsection is Stage 2 probability space (36).

Let us define the following stochastic process that can be used in bounding Bellman discrepancy Γ_N:

$$X_\theta := \bar{\mathcal{E}}\big(\hat{\mathcal{T}}^\pi \Upsilon_\theta, \mathcal{T}^\pi \Upsilon_\theta\big) \quad \text{and} \quad X_\theta(s,a) := \mathcal{E}\Big\{\hat{\mathcal{T}}^\pi \Upsilon_\theta(s,a), \mathcal{T}^\pi \Upsilon_\theta(s,a)\Big\},$$

$$\therefore\ \Gamma_N = \sup_{\theta\in\Theta} X_\theta \le \sup_{\theta\in\Theta}\Big|X_\theta - X_{\theta_0}\Big| + X_{\theta_0}, \tag{40}$$

where $\theta_0 \in \Theta$ is a fixed value that will be chosen at the later in the proof.

First, let us handle the supremum term $\sup_{\theta\in\Theta}|X_\theta - X_{\theta_0}|$ of Decomposition (40) with Dudley's integral inequality 4. We have

$$X_\theta = \sum_{s,a} b_\mu(s,a)\cdot X_\theta(s,a), \tag{41}$$

$$\|X_{\theta_1} - X_{\theta_2}\|_{\psi_2(\mathbf{N})} \le \sum_{s,a} b_\mu(s,a)\cdot \|X_{\theta_1}(s,a) - X_{\theta_2}(s,a)\|_{\psi_2(\mathbf{N})}, \tag{42}$$

and therefore we first need to bound the term $\|X_{\theta_1}(s,a) - X_{\theta_2}(s,a)\|_{\psi_2(\mathbf{N})}$. Towards that end, let us first rewrite the term $X_\theta(s,a)$ as follows where $\tilde{\mathbb{E}}(\cdots) = \mathbb{E}(\cdots|\mathcal{D})$ is the conditional expectation used in (11),

$$\begin{aligned} X_\theta(s,a) = 2\cdot\Bigg\{&\tilde{\mathbb{E}}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha;\theta) - \hat{R}_\beta - \gamma Z_\beta(\hat{S}'_\beta, \hat{A}'_\beta;\theta)\| \\ &\qquad - \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha;\theta) - R_\beta - \gamma Z_\beta(S'_\beta, A'_\beta;\theta)\|\Bigg\} \\ -\Bigg\{&\tilde{\mathbb{E}}\|\hat{R}_\alpha + \gamma Z_\alpha(\hat{S}'_\alpha, \hat{A}'_\alpha;\theta) - \hat{R}_\beta - \gamma Z_\beta(\hat{S}'_\beta, \hat{A}'_\beta;\theta)\| \\ &\qquad - \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha;\theta) - R_\beta - \gamma Z_\beta(S'_\beta, A'_\beta;\theta)\|\Bigg\}, \end{aligned}$$

where $Z(s,a;\theta) \sim \Upsilon_\theta(s,a)$, $(R,S') \sim p(\cdots|s,a)$, $(\hat{R},\hat{S}') \sim \hat{p}(\cdots|s,a)$, $A' \sim \pi(\cdot|S')$, $\hat{A}' \sim \pi(\cdot|\hat{S}')$, and having different subscripts (α or β) means they are independent, although they may follow the same distribution(s). That being said, we can simplify it as follows,

$$X_\theta(s,a) = \frac{2}{N(s,a)}\sum_{i=1}^{N(s,a)} W_i^\theta - \frac{1}{N(s,a)^2}\sum_{i=1}^{N(s,a)}\sum_{j=1}^{N(s,a)} W_{ij}^\theta, \tag{43}$$

where W_i^θ and W_{ij}^θ are the random variables that have following realizations,

$$\begin{aligned} w_i^\theta := \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha;\theta) - r_i - \gamma Z_\beta(s'_i, A'_i;\theta)\| \qquad (44)\\ - \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha;\theta) - R_\beta - \gamma Z_\beta(S'_\beta, A'_\beta;\theta)\|, \\ w_{ij}^\theta := \mathbb{E}\|r_i + \gamma Z_\alpha(s'_i, A'_i;\theta) - r_j - \gamma Z_\beta(s'_j, A'_j;\theta)\| \qquad (45)\\ - \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha;\theta) - R_\beta - \gamma Z_\beta(S'_\beta, A'_\beta;\theta)\|. \qquad (46) \end{aligned}$$

Since we have

$$\mathbb{E}(W_i^\theta) = 0,\ \ \mathbb{E}(W_{ij}^\theta) = 0 \text{ if } i \ne j,\ \ \mathbb{E}(W_{ii}^\theta) \ne 0,$$

we should further decompose Equation (43) as following, based on $N(s,a) \ge 2$ by Facts (39),

$$\begin{aligned} X_\theta(s,a) &= \frac{2}{N(s,a)}\sum_{i=1}^{N(s,a)} W_i^\theta - \frac{1}{N(s,a)^2}\cdot\Bigg(\sum_{i\ne j}^{N(s,a)} W_{ij}^\theta + \sum_{i=1}^{N(s,a)} W_i^\theta\Bigg) \qquad (47)\\ &= \frac{2}{N(s,a)}\sum_{i=1}^{N(s,a)} W_i^\theta - \frac{N(s,a)-1}{N(s,a)}\cdot\frac{1}{N(s,a)\cdot(N(s,a)-1)}\sum_{i\ne j}^{N(s,a)} W_{ij}^\theta - \frac{1}{N(s,a)^2}\sum_{i=1}^{N(s,a)} W_{ii}^\theta, \end{aligned}$$

This leads to

$$\begin{aligned}
&\|X_{\theta_1}(s,a) - X_{\theta_2}(s,a)\|_{\psi_2(\mathbf{N})} \\
&\quad \le \frac{N(s,a)-1}{N(s,a)} \cdot \left\| \frac{1}{N(s,a)\cdot(N(s,a)-1)} \sum_{i\neq j}^{N(s,a)} (W_{ij}^{\theta_1} - W_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \\
&\qquad + \left\| \frac{2}{N(s,a)} \sum_{i=1}^{N(s,a)} (W_i^{\theta_1} - W_i^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} + \left\| \frac{1}{N(s,a)^2} \sum_{i=1}^{N(s,a)} (W_{ii}^{\theta_1} - W_{ii}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})}, \quad (48)
\end{aligned}$$

and we will bound each term one by one. Before we begin with the first term, we would like to introduce a useful trick that will be used repetitively throughout the proof. First, it is easy to see that $|w_{12}^{\theta_1} - w_{12}^{\theta_2}|$ can be decomposed into the following two terms,

$$\begin{aligned}
|w_{12}^{\theta_1} - w_{12}^{\theta_2}| \le &\Big| \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', A_1'; \theta_1) - r_2 - \gamma Z_\beta(s_2', A_2'; \theta_1)\| \\
&\qquad - \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', A_1'; \theta_2) - r_2 - \gamma Z_\beta(s_2', A_2'; \theta_2)\| \Big| \\
&+ \Big| \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S_\alpha', A_\alpha'; \theta_1) - R_\beta - \gamma Z_\beta(S_\beta', A_\beta'; \theta_1)\| \\
&\qquad - \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S_\alpha', A_\alpha'; \theta_2) - R_\beta - \gamma Z_\beta(S_\beta', A_\beta'; \theta_2)\| \Big|.
\end{aligned}$$

Since we have the following by Assumption 5,

$$\begin{aligned}
&\Big| \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', a_1'; \theta_1) - r_2 - \gamma Z_\beta(s_2', a_2'; \theta_1)\| \\
&\qquad - \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', a_1'; \theta_2) - r_2 - \gamma Z_\beta(s_2', a_2'; \theta_2)\| \Big| \le 2\gamma \cdot \tilde{\eta}(\theta_1, \theta_2). \quad (49)
\end{aligned}$$

this leads to

$$\begin{aligned}
&\Big| \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', A_1'; \theta_1) - r_2 - \gamma Z_\beta(s_2', A_2'; \theta_1)\| \\
&\qquad - \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', A_1'; \theta_2) - r_2 - \gamma Z_\beta(s_2', A_2'; \theta_2)\| \Big| \\
&= \Big| \sum_{a_1', a_2' \in \mathcal{A}} \pi(a_1'|s_1') \cdot \pi(a_2'|s_2') \cdot \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', A_1'; \theta_1) - r_2 - \gamma Z_\beta(s_2', A_2'; \theta_1)\| \\
&\qquad - \sum_{a_1', a_2' \in \mathcal{A}} \pi(a_1'|s_1') \cdot \pi(a_2'|s_2') \cdot \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', A_1'; \theta_2) - r_2 - \gamma Z_\beta(s_2', A_2'; \theta_2)\| \Big| \\
&\le \sum_{a_1', a_2' \in \mathcal{A}} \pi(a_1'|s_1') \cdot \pi(a_2'|s_2') \cdot \Big| \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', a_1'; \theta_1) - r_2 - \gamma Z_\beta(s_2', a_2'; \theta_1)\| \\
&\qquad - \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', a_1'; \theta_2) - r_2 - \gamma Z_\beta(s_2', a_2'; \theta_2)\| \Big| \\
&\le 2\gamma \cdot \tilde{\eta}(\theta_1, \theta_2) \quad \text{by Inequality (49).} \quad (50)
\end{aligned}$$

Furthermore, by letting $P_{(2)}^{\pi}$ to be the (conditional) probability measure of $(r_1, s_1', a_1', r_2, s_2', a_2')$ with (r_1, s_1', a_1') and (r_2, s_2', a_2') being independent conditioned on (s, a), we can derive the following,

$$\begin{aligned}
&\Big| \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S_\alpha', A_\alpha'; \theta_1) - R_\beta - \gamma Z_\beta(S_\beta', A_\beta'; \theta_1)\| \\
&\qquad - \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S_\alpha', A_\alpha'; \theta_2) - R_\beta - \gamma Z_\beta(S_\beta', A_\beta'; \theta_2)\| \Big|
\end{aligned}$$

$$
\begin{aligned}
&= \Bigg| \int \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', a_1'; \theta_1) - r_2 - \gamma Z_\beta(s_2', a_2'; \theta_1)\| \mathrm{d}P^\pi_{(2)}(r_1, s_1', a_1', r_2, s_2', a_2'|s, a) \\
&\qquad\qquad - \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', a_1'; \theta_2) - r_2 - \gamma Z_\beta(s_2', a_2'; \theta_2)\| \mathrm{d}P^\pi_{(2)}(r_1, s_1', a_1' r_2, s_2', a_2'|s, a) \Bigg| \\
&\leq \int \Bigg| \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', a_1'; \theta_1) - r_2 - \gamma Z_\beta(s_2', a_2'; \theta_1)\| \\
&\qquad\qquad - \mathbb{E}\|r_1 + \gamma Z_\alpha(s_1', a_1'; \theta_2) - r_2 - \gamma Z_\beta(s_2', a_2'; \theta_2)\| \Bigg| \mathrm{d}P^\pi_{(2)}(r_1, s_1', a_1', r_2, s_2', a_2'|s, a) \\
&\leq 2\gamma \cdot \tilde{\eta}(\theta_1, \theta_2) \text{ by Inequality (49).} \qquad (51)
\end{aligned}
$$

Combining the two inequalities (50) and (51), we obtain

$$
|w_{12}^{\theta_1} - w_{12}^{\theta_2}| \leq 4\gamma \cdot \tilde{\eta}(\theta_1, \theta_2), \qquad (52)
$$

which implies that $|W_{ij}^{\theta_1} - W_{ij}^{\theta_2}| \leq 4\gamma \cdot \tilde{\eta}(\theta_1, \theta_2)$ is a bounded random variable. Defining another random variable that satisfies following based on Remark 1,

$$
\tilde{W}_{ij}^{\theta} := \frac{1}{2} \cdot (W_{ij}^{\theta} + W_{ji}^{\theta}) \text{ where } 1 \leq i < j \leq N(s, a), \qquad (53)
$$

$$
\therefore \|\tilde{W}_{12}^{\theta_1} - \tilde{W}_{12}^{\theta_2}\|_{\psi_2(\mathbf{N})} \leq \frac{1}{2}\|W_{12}^{\theta_1} - W_{12}^{\theta_2}\|_{\psi_2(\mathbf{N})} + \frac{1}{2}\|W_{21}^{\theta_1} - W_{21}^{\theta_2}\|_{\psi_2(\mathbf{N})} \leq \frac{4\gamma}{\sqrt{\log 2}} \cdot \tilde{\eta}(\theta_1, \theta_2).
$$

Then we rewrite the first term as follows,

$$
\begin{aligned}
&\left\| \frac{1}{N(s, a) \cdot (N(s, a) - 1)} \sum_{i \neq j}^{N(s,a)} (W_{ij}^{\theta_1} - W_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \\
&\qquad = \left\| \frac{1}{N(s, a) \cdot (N(s, a) - 1)/2} \sum_{i < j}^{N(s,a)} (\tilde{W}_{ij}^{\theta_1} - \tilde{W}_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})}, \qquad (54)
\end{aligned}
$$

and we will apply the following corollary, whose proof is in Section C.2.3.

Corollary 1. *For i.i.d. mean-zero subgaussian random variables $X_1, \cdots, X_n$, we have*

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \right\|_{\psi_2} \leq \frac{C}{\sqrt{n}} \cdot \|X_1\|_{\psi_2}.
$$

The tuples that we add up in Equation (54) are not necessarily independent, so we need to reorganize the terms. Towards that end, we divide our cases into two, when $N(s, a) \geq 2$ is an even number or an odd number. When $N(s, a)$ is even, we can directly use Lemma 6 of C.1 to group $\{(i, j) : 1 \leq i < j \leq N(s, a)\}$ into $(N(s, a) - 1)$ groups G_k $(1 \leq k \leq N(s, a) - 1)$, each of which contains $|G_k| = N(s, a)/2$ pairs of (i, j), with no pair overlapping in any component. Then we can take up Equation (54) as follows,

$$
\begin{aligned}
&\left\| \frac{1}{N(s, a) \cdot (N(s, a) - 1)} \sum_{i \neq j}^{N(s,a)} (W_{ij}^{\theta_1} - W_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \\
&\qquad \leq \left\| \frac{1}{N(s, a) - 1} \sum_{k=1}^{N(s,a)-1} \frac{1}{N(s, a)/2} \sum_{(i,j) \in G_k} (\tilde{W}_{ij}^{\theta_1} - \tilde{W}_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \\
&\qquad \leq \left\| \frac{1}{N(s, a)/2} \sum_{(i,j) \in G_1} (\tilde{W}_{ij}^{\theta_1} - \tilde{W}_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \text{ by Triangular Inequality} \\
&\qquad \leq \frac{C_1}{\sqrt{N(s, a)}} \cdot \|\tilde{W}_{12}^{\theta_1} - \tilde{W}_{12}^{\theta_2}\|_{\psi_2(\mathbf{N})} \text{ by Corollary 1.} \qquad (55)
\end{aligned}
$$

Now let us assume that $N(s,a) \geq 2$ is an odd number, which automatically gives us $N(s,a) \geq 3$. Then we can take up Equation (54) as follows,

$$
\begin{aligned}
&\left\| \frac{1}{N(s,a)\cdot(N(s,a)-1)} \sum_{i\neq j}^{N(s,a)} (W_{ij}^{\theta_1} - W_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \\
&\leq \left\| \frac{N(s,a)-2}{N(s,a)} \cdot \frac{1}{(N(s,a)-1)\cdot(N(s,a)-2)/2} \sum_{i<j}^{N(s,a)-1} (\tilde{W}_{ij}^{\theta_1} - \tilde{W}_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \\
&\qquad + \frac{1}{N(s,a)/2} \cdot \left\| \frac{1}{N(s,a)-1} \sum_{i=1}^{N(s,a)-1} (\tilde{W}_{iN(s,a)}^{\theta_1} - \tilde{W}_{iN(s,a)}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \\
&\leq \frac{N(s,a)-2}{N(s,a)} \cdot \left\| \frac{2}{(N(s,a)-1)\cdot(N(s,a)-2)} \sum_{i<j}^{N(s,a)-1} (\tilde{W}_{ij}^{\theta_1} - \tilde{W}_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \\
&\qquad + \frac{2}{N(s,a)} \cdot \|\tilde{W}_{12}^{\theta_1} - \tilde{W}_{12}^{\theta_2}\|_{\psi_2(\mathbf{N})} \\
&\leq \frac{C_1}{\sqrt{N(s,a)-1}} \cdot \|\tilde{W}_{12}^{\theta_1} - \tilde{W}_{12}^{\theta_2}\|_{\psi_2(\mathbf{N})} + \frac{2}{N(s,a)} \cdot \|\tilde{W}_{12}^{\theta_1} - \tilde{W}_{12}^{\theta_2}\|_{\psi_2(\mathbf{N})} \quad \text{by Inequality (55)} \\
&\leq \frac{C_2}{\sqrt{N(s,a)}} \cdot \|\tilde{W}_{12}^{\theta_1} - \tilde{W}_{12}^{\theta_2}\|_{\psi_2(\mathbf{N})} \quad \because\ N(s,a)-1 \geq \frac{N(s,a)}{2} \text{ since } N(s,a) \geq 3,
\end{aligned} \tag{56}
$$

where the second last line holds since $N(s,a)-1$ is an even number. That being said, we can generalize the following result for $\forall N(s,a) \in \mathbb{N}$ based on (53), regardless of even or odd numbers,

$$
\left\| \frac{1}{N(s,a)\cdot(N(s,a)-1)} \sum_{i\neq j}^{N(s,a)} (W_{ij}^{\theta_1} - W_{ij}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \leq \frac{C_4\gamma}{\sqrt{N(s,a)}} \cdot \tilde{\eta}(\theta_1,\theta_2).
$$

Regarding the second term of Inequality (48), we can apply the same trick of (51) to obtain

$$
\begin{aligned}
&\mathbb{E}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha; \theta_1) - r_1 - \gamma Z_\beta(s'_1, A'_1; \theta_1)\| \\
&\qquad - \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha; \theta_2) - r_1 - \gamma Z_\beta(s'_1, A'_1; \theta_2)\| \\
&\leq 2\gamma \cdot \tilde{\eta}(\theta_1,\theta_2) \ \text{ by Inequality (49).}
\end{aligned}
$$

which leads to following by the same logic of (52),

$$
|w_1^{\theta_1} - w_1^{\theta_2}| \leq 4\gamma \cdot \tilde{\eta}(\theta_1,\theta_2).
$$

This allows us to bound the second term as follows by employing the same logic as when we bounded the first term of (48),

$$
\left\| \frac{2}{N(s,a)} \sum_{i=1}^{N(s,a)} (W_i^{\theta_1} - W_i^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} \leq \frac{C_6\gamma}{\sqrt{N(s,a)}} \cdot \tilde{\eta}(\theta_1,\theta_2).
$$

The third term of Inequality (48) can be bounded as follows,

$$
\begin{aligned}
\left\| \frac{1}{N(s,a)^2} \sum_{i=1}^{N(s,a)} (W_{ii}^{\theta_1} - W_{ii}^{\theta_2}) \right\|_{\psi_2(\mathbf{N})} &\leq \frac{1}{N(s,a)^2} \cdot N(s,a) \cdot \|W_{11}^{\theta_1} - W_{11}^{\theta_2}\|_{\psi_2(\mathbf{N})} \\
&\leq \frac{C_7\gamma}{N(s,a)} \cdot \tilde{\eta}(\theta_1,\theta_2) \ \text{ by the same trick (52).}
\end{aligned}
$$

Finally, we can bound Inequality (48) as follows,

$$
\|X_{\theta_1}(s,a) - X_{\theta_2}(s,a)\|_{\psi_2(\mathbf{N})} \leq \frac{C_8\gamma}{\sqrt{N(s,a)}} \cdot \tilde{\eta}(\theta_1,\theta_2),
$$

which eventually leads to following by Inequality (42), by using $N(s,a) \geq \frac{1}{2} b_\mu(s,a) \cdot N$ based on Fact (39),

$$\|X_{\theta_1} - X_{\theta_2}\|_{\psi_2(\mathbf{N})} \leq \sum_{s,a} b_\mu(s,a) \cdot \frac{C_8 \gamma}{\sqrt{N(s,a)}} \cdot \tilde{\eta}(\theta_1, \theta_2) \leq \frac{C_9 \gamma}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \tilde{\eta}(\theta_1, \theta_2).$$

Without loss of generality, we can assume that separability holds. Therefore, by Assumptions 3 and 4, we can apply Dudley's Integral Inequality 4 to obtain the following for $\forall u > 0$,

$$\mathbb{P}^{(\mathbf{N})}\Bigg[\sup_{\theta \in \Theta} \Big| X_\theta - X_{\theta_0} \Big| \leq \frac{C_{10}\gamma}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \Bigg\{ \int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)} \mathrm{d}t + u \cdot \mathrm{diam}(\Theta; \tilde{\eta}) \Bigg\} \Bigg] \geq 1 - 2\exp(-u^2). \tag{57}$$

The next part is bounding the term X_{θ_0} of Decomposition (40). We first fix a state-action pair $s, a \in \mathcal{S} \times \mathcal{A}$, we can use the last line of Equation (47) to obtain the following decomposition,

$$X_{\theta_0}(s,a) \leq 2 \cdot \Big| \frac{1}{N(s,a)} \sum_{i=1}^{N(s,a)} W_i^{\theta_0} \Big| + \frac{N(s,a)-1}{N(s,a)} \cdot \Big| \frac{1}{N(s,a) \cdot (N(s,a)-1)} \sum_{i \neq j}^{N(s,a)} W_{ij}^{\theta_0} \Big| + \frac{1}{N(s,a)} \cdot \Big| \frac{1}{N(s,a)} \sum_{i=1}^{N(s,a)} W_{ii}^{\theta_0} \Big|. \tag{58}$$

We first select $\epsilon_1 > 0$, and we will bound each term one by one. Starting from the first term of Decomposition (58), we can apply Theorem 5 by Assumptions 3 and 4, to obtain the following, where $\mathbb{E}^{(\mathbf{N})}(\cdots)$ is the conditional expectation that corresponds to the conditional probability $\mathbb{P}^{(\mathbf{N})}(\cdots)$,

$$\mathbb{P}^{(\mathbf{N})}\Bigg(\Big| \frac{1}{N(s,a)} \sum_{i=1}^{N(s,a)} W_i^{\theta_0} \Big| \geq \epsilon_1 \Bigg) \leq 2 \cdot \exp\Bigg\{ \frac{-C_{11} \cdot N(s,a) \cdot \epsilon_1^2}{\|W_1^{\theta_0} - \mathbb{E}^{(\mathbf{N})}(W_1^{\theta_0})\|_{\psi_2(\mathbf{N})}^2} \Bigg\}. \tag{59}$$

Note that we could remove the expectation term in the LHS due to

$$\mathbb{E}^{(\mathbf{N})}(W_i^{\theta_0}) = \mathbb{E}(W_i^{\theta_0}) = 0, \tag{60}$$

which holds since the randomness of $W_i^{\theta_0}$ solely depends on $(R(s,a), S'(s,a)) \sim p(\cdots | s, a)$ for a fixed state-action pair s, a, which is irrelevant (independent) with $\mathbf{N} = (N(s,a))_{s,a \in \mathcal{S} \times \mathcal{A}}$. Then we have the following based on Definition (46),

$$\begin{aligned} \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S'_\alpha, A'_\alpha; \theta_0) - r_1 - \gamma Z_\beta(s'_1, A'_1; \theta_0)\| & \\ \leq \mathbb{E}\|R(s,a)\| + 2\gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + \|r_1\|, & \end{aligned}$$

we can apply Remark 1 to see

$$\begin{aligned} \|W_1^{\theta_0} - \mathbb{E}^{(\mathbf{N})}(W_1^{\theta_0})\|_{\psi_2(\mathbf{N})} &\leq C_{12} \cdot \|W_1^{\theta_0}\|_{\psi_2(\mathbf{N})} \\ &\leq C_{13} \cdot \Big\| \mathbb{E}\|R(s,a)\| + 2\gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + \|R(s,a)\| \Big\|_{\psi_2(\mathbf{N})} \\ &= C_{13} \cdot \Big\| \mathbb{E}\|R(s,a)\| + 2\gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + \|R(s,a)\| \Big\|_{\psi_2} \\ &\leq C_{13} \cdot \Bigg\{ 2 \cdot \Big\| \|R(s,a)\| \Big\|_{\psi_2} + \frac{2\gamma}{\sqrt{\log 2}} \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \Bigg\}, \end{aligned} \tag{61}$$

where the third line holds with the same reason as in (60), that is the realization of $\mathbf{N} = (N(s,a))_{s,a \in \mathcal{S} \times \mathcal{A}}$ plays no role in the distribution of $R(s,a)$ that is conditioned on a fixed state-action pair s, a. Since we have

$$\Big\| \mathbb{E}\|R(s,a)\| \Big\|_{\psi_2} \leq \Big\| \|R(s,a)\| \Big\|_{\psi_2} \leq d \cdot \|R(s,a)\|_{\psi_2} \text{ by Remark 1}, \tag{62}$$

this further leads to following by Inequality (61),

$$\|W_1^{\theta_0} - \mathbb{E}^{(\mathbf{N})}(W_1^{\theta_0})\|_{\psi_2(\mathbf{N})}^2 \leq C_{14} \cdot \Bigg\{ \Big\| \|R(s,a)\| \Big\|_{\psi_2} + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \Bigg\}^2 \tag{63}$$

Since Facts (39) implies $N(s,a)/N \geq \frac{1}{2} b_\mu(s,a) \geq \frac{1}{2} p_{\min}$, this allows us to take up Bound (59) as follows,

$$\mathbb{P}^{(\mathbf{N})}\left(\left|\frac{1}{N(s,a)}\sum_{i=1}^{N(s,a)} W_i^{\theta_0}\right| \geq \epsilon_1\right) \leq 2\cdot\exp\left\{\frac{-C_{15}\cdot N(s,a)\cdot\epsilon_1^2}{\left(\left\|\|R(s,a)\|\right\|_{\psi_2} + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2}\right\}$$
$$\leq 2\cdot\exp\left\{\frac{-C_{16}\cdot p_{\min}\cdot N\cdot\epsilon_1^2}{\left(\left\|\|R(s,a)\|\right\|_{\psi_2} + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2}\right\}. \quad (64)$$

Now we have to bound the second term of Decomposition (58), which can be derived similarly to the first bound (64), but takes one additional step of employing the following lemma that is proved in C.2.4,

Lemma 1. *Given $X_1,\cdots,X_N \sim iid$ $(N\geq 2)$, let $X_{ij} := h(X_i,X_j)$ for some bivariate function $h(\cdot,\cdot)$, and assume that $\|X_{12}\|_{\psi_2} < \infty$ holds. Then we have the following inequality for $\forall\epsilon > 0$,*

$$\mathbb{P}\left\{\left|\frac{1}{N(N-1)}\sum_{i\neq j} X_{ij} - \mathbb{E}(X_{12})\right| \geq \epsilon\right\} \leq 4N\cdot\exp\left\{\frac{-C\cdot N\cdot\epsilon^2}{\|X_{12}-\mathbb{E}(X_{12})\|_{\psi_2}^2}\right\}.$$

By Lemma 1, we have

$$\mathbb{P}^{(\mathbf{N})}\left(\left|\frac{1}{N(s,a)\cdot(N(s,a)-1)}\sum_{i\neq j}^{N(s,a)} W_{ij}^{\theta_0}\right| \geq \epsilon_1\right)$$
$$\leq 4N(s,a)\cdot\exp\left\{\frac{-C_{17}\cdot N(s,a)\cdot\epsilon_1^2}{\|W_{12}^{\theta_0}-\mathbb{E}^{(\mathbf{N})}(W_{12}^{\theta_0})\|_{\psi_2(\mathbf{N})}^2}\right\},$$

and then we can employ the tricks (61) and (63) to obtain

$$\|W_{12}^{\theta_0}-\mathbb{E}^{(\mathbf{N})}(W_{12}^{\theta_0})\|_{\psi_2(\mathbf{N})}^2 \leq C_{18}\cdot\left\{\left\|\|R(s,a)\|\right\|_{\psi_2} + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right\}^2.$$

Using $N(s,a)\geq\frac{1}{2}p_{\min}$ and $N(s,a)\leq\frac{3}{2}b_\mu(s,a)$ that are implied by Facts (39), we can derive the following bound,

$$\mathbb{P}^{(\mathbf{N})}\left(\left|\frac{1}{N(s,a)\cdot(N(s,a)-1)}\sum_{i\neq j}^{N(s,a)} W_{ij}^{\theta_0}\right| \geq \epsilon_1\right)$$
$$\leq 6N\cdot b_\mu(s,a)\cdot\exp\left\{\frac{-C_{19}\cdot p_{\min}\cdot N\cdot\epsilon_1^2}{\left(\left\|\|R(s,a)\|\right\|_{\psi_2} + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2}\right\}. \quad (65)$$

Lastly, we bound the third term of Decomposition (58). Since $\mathbb{E}(W_{ii}^{\theta_0})\neq 0$, we cannot repeat the same procedure that we employed for the first and second terms. Based on Definition (46), we can see that $|W_{11}^{\theta_0}|$ is a bound random variable,

$$\begin{aligned}|w_{11}^{\theta_0}| &= \mathbb{E}\|\gamma Z_\alpha(s_i',A_i';\theta_0) - \gamma Z_\beta(s_j',A_j';\theta_0)\| \\ &\qquad - \mathbb{E}\|R_\alpha + \gamma Z_\alpha(S_\alpha',A_\alpha';\theta_0) - R_\beta - \gamma Z_\beta(S_\beta',A_\beta';\theta_0)\| \\ &\leq 2\sup_{s,a}\mathbb{E}\|R(s,a)\| + 4\gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a,\theta_0)\|,\end{aligned}$$

which leads us to bound the third term of (58) as following based on $N(s,a)\geq\frac{1}{2}p_{\min}\cdot N$,

$$\frac{1}{N(s,a)}\cdot\left|\frac{1}{N(s,a)}\sum_{i=1}^{N(s,a)} W_{ii}^{\theta_0}\right| \leq \frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left\{\mathbb{E}\|R(s,a)\| + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right\}. \quad (66)$$

We can combine Bounds (64), (65), and (66) to form the following bound, based on Decomposition (58),

$$\begin{aligned}
&\mathbb{P}^{(\mathbf{N})}\left\{X_{\theta_0}(s,a) \le 2\epsilon_1 + \frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\mathbb{E}\|R(s,a)\| + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)\right\} \\
&\qquad \ge 1 - 2\cdot\exp\left\{\frac{-C_{16}\cdot p_{\min}\cdot N\cdot\epsilon_1^2}{\left(\left\|\|R(s,a)\|\right\|_{\psi_2} + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2}\right\} \\
&\qquad\qquad - 6N\cdot b_\mu(s,a)\cdot\exp\left\{\frac{-C_{19}\cdot p_{\min}\cdot N\cdot\epsilon_1^2}{\left(\left\|\|R(s,a)\|\right\|_{\psi_2} + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2}\right\}.
\end{aligned}$$

This further leads to

$$\begin{aligned}
&\mathbb{P}^{(\mathbf{N})}\left\{X_{\theta_0} \le 2\epsilon_1 + \frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\sup_{s,a}\mathbb{E}\|R(s,a)\| + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)\right\} \\
&\ge \mathbb{P}^{(\mathbf{N})}\left\{\sup_{s,a}X_{\theta_0}(s,a) \le 2\epsilon_1 + \frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\sup_{s,a}\mathbb{E}\|R(s,a)\| + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)\right\} \\
&\ge 1 - (2|\mathcal{S}\times\mathcal{A}| + 6N)\cdot\exp\left\{\frac{-C_{20}\cdot p_{\min}\cdot N\cdot\epsilon_1^2}{\left(\sup_{s,a}\left\|\|R(s,a)\|\right\|_{\psi_2} + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2}\right\}. \qquad (67)
\end{aligned}$$

Note that we obtained a bound for $\sup_{s,a}X_{\theta_0}(s,a)$, which is one step further than X_{θ_0}. This shall be later used in proof of Lemma 2 for non-realizable scenario, which is suggested in B.2.

Now we can combine the bounds (57) and (67) to take up Decomposition (40) as following for $\forall u>0,\ \epsilon_1>0,\ N\ge 2$,

$$\begin{aligned}
&\mathbb{P}^{(\mathbf{N})}\left\{\Gamma_N \le \frac{C_{10}\gamma}{\sqrt{N}}\cdot\sum_{s,a}\sqrt{b_\mu(s,a)}\cdot\left(\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t + u\cdot\mathrm{diam}(\Theta;\tilde{\eta})\right)\right. \\
&\qquad\qquad \left. + 2\epsilon_1 + \frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\sup_{s,a}\mathbb{E}\|R(s,a)\| + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)\right\} \\
&\ge 1 - 2\exp(-u^2) - (2|\mathcal{S}\times\mathcal{A}| + 6N) \\
&\qquad\qquad \times\exp\left\{\frac{-C_{21}\cdot p_{\min}\cdot N\cdot\epsilon_1^2}{d^2\cdot\left(\sup_{s,a}\|R(s,a)\|_{\psi_2}\right)^2 + \gamma^2\cdot\left(\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2}\right\}, \qquad (68)
\end{aligned}$$

where the second last inequality holds by Inequality (62) and the technique $(a+b)^2\le 2(a^2+b^2)$. We should always remember that we are conditioning on the event $\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}$ defined in Definition (38).

A.6.4 Bounding State-action discrepancy

This time, we will bound state-action discrepancy Δ_N that occurs due to the estimation error of $b_\mu(s,a)$. As warned in the last paragraph of A.6.2, we are still assuming $N(s,a)$ to be fixed, satisfying Facts (39). Accordingly, we only deal with Stage 2 probability space (36) with the conditional probability measure $\mathbb{P}^{(\mathbf{N})}$.

With $\mathbf{p}$ and $\hat{\mathbf{p}}$ defined in A.6.2 and $\epsilon>0$ being the value specified in Definition (38), and defining $\|\mathbf{x}\|_1 := \sum_{i=1}^{p_0}|x_i|$ for $\forall\mathbf{x}=(x_1,\cdots,x_{p_0})^\intercal$, we can derive the following based on $(x_1+\cdots+x_{p_0})^2\le p_0\cdot(x_1^2+\cdots x_{p_0}^2)$,

$$\sum_{s,a}\left|\hat{b}_\mu(s,a) - b_\mu(s,a)\right| = \|\mathbf{p}-\hat{\mathbf{p}}\|_1 \le \sqrt{|\mathcal{S}\times\mathcal{A}|}\cdot\|\mathbf{p}-\hat{\mathbf{p}}\|.$$

Then we have the following extension,

$$\begin{aligned}
\Delta_N &:= \sup_{\theta\in\Theta}\left|\bar{\mathcal{E}}(\Upsilon_\theta,\hat{\mathcal{T}}^\pi\Upsilon_\theta) - \hat{\bar{\mathcal{E}}}(\Upsilon_\theta,\hat{\mathcal{T}}^\pi\Upsilon_\theta)\right| \\
&\le \sum_{s,a}\left|\hat{b}_\mu(s,a) - b_\mu(s,a)\right|\cdot\sup_{\theta\in\Theta}\mathcal{E}\left\{\Upsilon_\theta(s,a),\hat{\mathcal{T}}^\pi\Upsilon_\theta(s,a)\right\}
\end{aligned}$$

$$\leq \sqrt{|\mathcal{S}\times\mathcal{A}|}\cdot\|\mathbf{p}-\hat{\mathbf{p}}\|\cdot\sup_{\theta\in\Theta}\sup_{s,a}\mathcal{E}\left\{\Upsilon_\theta(s,a),\hat{\mathcal{T}}^\pi\Upsilon_\theta(s,a)\right\}$$
$$\leq \frac{\sqrt{p_{\min}}}{2}\cdot\epsilon\cdot\sup_{\theta\in\Theta}\sup_{s,a}\mathcal{E}\left\{\Upsilon_\theta(s,a),\hat{\mathcal{T}}^\pi\Upsilon_\theta(s,a)\right\}\text{ by Facts (39).}\tag{69}$$

where the last line holds due to $p_{\min}\leq 1/|\mathcal{S}\times\mathcal{A}|$ that leads to $|\mathcal{S}\times\mathcal{A}|\leq 1/p_{\min}$, by the definition of minimum probability $p_{\min}>0$ in Assumption 1.

Now let us handle the supremum term of Inequality (69). Letting $R_i(s,a)$ $(1\leq i\leq N(s,a))$ be the reward vectors observed conditioned on s,a, we have the following hold based on the notations introduced in (11),

$$\sup_{\theta\in\Theta}\sup_{s,a}\mathcal{E}\left\{\Upsilon_\theta(s,a),\hat{\mathcal{T}}^\pi\Upsilon_\theta(s,a)\right\}\leq\sup_{\theta\in\Theta}\sup_{s,a}\Big\{2\tilde{\mathbb{E}}\|Z_\alpha(s,a;\theta)-\hat{Z}^{(1)}_\beta(s,a;\theta)\|$$
$$-\tilde{\mathbb{E}}\|Z_\alpha(s,a;\theta)-Z_\beta(s,a;\theta)\|-\tilde{\mathbb{E}}\|\hat{Z}^{(1)}_\alpha(s,a;\theta)-\hat{Z}^{(1)}_\beta(s,a;\theta)\|\Big\}$$
$$\leq\sup_{\theta\in\Theta}\sup_{s,a}\left\{4\tilde{\mathbb{E}}\|Z(s,a;\theta)\|+4\cdot\tilde{\mathbb{E}}\|\hat{Z}^{(1)}(s,a;\theta)\|\right\}$$
$$\leq 4(1+\gamma)\cdot\sup_{\theta\in\Theta}\sup_{s,a}\mathbb{E}\|Z(s,a;\theta)\|+4\cdot\sup_{s,a}\left\{\frac{1}{N(s,a)}\cdot\sum_{i=1}^{N(s,a)}\|R_i(s,a)\|\right\}.\tag{70}$$

The first term can be bounded as follows, using the property of $\tilde{\eta}$ introduced in Assumption 5,

$$\sup_{\theta\in\Theta}\sup_{s,a}\mathbb{E}\|Z(s,a;\theta)\|=\sup_{\theta\in\Theta}\left\{\sup_{s,a}\mathbb{E}\|Z(s,a;\theta)\|-\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right\}+\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|$$
$$\leq\sup_{\theta\in\Theta}\left\{\sup_{s,a}\Big|\mathbb{E}\|Z(s,a;\theta)\|-\mathbb{E}\|Z(s,a;\theta_0)\|\Big|\right\}+\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|$$
$$\leq\sup_{\theta\in\Theta}\tilde{\eta}(\theta,\theta_0)+\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\leq\mathrm{diam}(\Theta;\tilde{\eta})+\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|.\tag{71}$$

Now let $\epsilon_2>0$ be arbitrary. For a fixed $s,a\in\mathcal{S}\times\mathcal{A}$, we have

$$\mathbb{P}^{(\mathbf{N})}\left\{\frac{1}{N(s,a)}\sum_{i=1}^{N(s,a)}\|R_i(s,a)\|\geq\mathbb{E}\|R(s,a)\|+\epsilon_2\right\}$$
$$\leq\mathbb{P}^{(\mathbf{N})}\left\{\left|\frac{1}{N(s,a)}\sum_{i=1}^{N(s,a)}\|R_i(s,a)\|-\mathbb{E}\|R(s,a)\|\right|\geq\epsilon_2\right\}$$
$$\leq 2\cdot\exp\left\{\frac{-C_{22}\cdot N(s,a)\cdot\epsilon_2^2}{\big\|\|R_1(s,a)\|-\mathbb{E}\|R(s,a)\|\big\|^2_{\psi_2}}\right\}\leq 2\cdot\exp\left\{\frac{-C_{23}\cdot p_{\min}\cdot N\cdot\epsilon_2^2}{d^2\cdot\|R(s,a)\|^2_{\psi_2}}\right\}\tag{72}$$

where the last inequality is by Inequality (62) and Remark 1. Note that we used $\mathbb{E}$ and ψ_2 instead of $\mathbb{E}^{(\mathbf{N})}$ and $\psi_2(\mathbf{N})$ in the first two inequalties, by the same reason mentioned beneath Inequalities (59) and (61). We also used Facts (39) that implies $N(s,a)\geq p_{\min}/2\cdot N$ in the last line. Then we can expand the result towards following,

$$\mathbb{P}^{(\mathbf{N})}\left[\sup_{s,a}\left\{\frac{1}{N(s,a)}\sum_{i=1}^{N(s,a)}\|R_i(s,a)\|\right\}\geq\sup_{s,a}\mathbb{E}\|R(s,a)\|+\epsilon_2\right]$$
$$\leq\mathbb{P}^{(\mathbf{N})}\left\{\frac{1}{N(s,a)}\sum_{i=1}^{N(s,a)}\|R_i(s,a)\|\geq\mathbb{E}\|R(s,a)\|+\epsilon_2\ \text{ for }\exists s,a\in\mathcal{S}\times\mathcal{A}\right\}$$
$$\leq\sum_{s,a}\mathbb{P}^{(\mathbf{N})}\left\{\frac{1}{N(s,a)}\sum_{i=1}^{N(s,a)}\|R_i(s,a)\|\geq\mathbb{E}\|R(s,a)\|+\epsilon_2\right\}$$
$$\leq 2|\mathcal{S}\times\mathcal{A}|\cdot\exp\left\{\frac{-C_{23}\cdot p_{\min}\cdot N\cdot\epsilon_2^2}{d^2\cdot\left(\sup_{s,a}\|R(s,a)\|_{\psi_2}\right)^2}\right\}\text{ by Inequality (72).}\tag{73}$$

Plugging (71) and (73) into Inequality (70) and then into Inequality (69) returns

$$\mathbb{P}^{(\mathbf{N})}\Bigg[\Delta_N \leq 2\sqrt{p_{\min}}\cdot\epsilon\cdot\Bigg\{(1+\gamma)\cdot\Bigg(\operatorname{diam}(\Theta;\tilde{\eta})+\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\Bigg) + \sup_{s,a}\mathbb{E}\|R(s,a)\|+\epsilon_2\Bigg\}\Bigg] \geq 1-2|\mathcal{S}\times\mathcal{A}|\cdot\exp\left\{\frac{-C_{23}\cdot p_{\min}\cdot N\cdot\epsilon_2^2}{d^2\cdot\left(\sup_{s,a}\|R(s,a)\|_{\psi_2}\right)^2}\right\}. \tag{74}$$

A.6.5 Finalizing the Bound

Recall that we defined $\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}$ (38) in A.6.2 where the samples are collected sufficiently many for each s,a. Assuming this, we have bounded Γ_N and Δ_N throughout A.6.3 and A.6.4, each in (68) and (74). Simply put, letting $E\subset\Omega$ be the event where Γ_N and Δ_N simulatenously achieve the specified bounds (68) and (74) can be understood as $\mathbb{P}(E|\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})$. Then we get

$$\begin{aligned}\mathbb{P}(E) &\geq \mathbb{P}(E\cap\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}) = \mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})\cdot\mathbb{P}(E|\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})\\ &= \left\{1-(1-\mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}))\right\}\cdot\left\{1-(1-\mathbb{P}(E|\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}))\right\}\\ &\geq 1-(1-\mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}))-(1-\mathbb{P}(E|\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})).\end{aligned} \tag{75}$$

According to (36), it may be more rigorous to denote $\mathbb{P}_{\mathcal{S}\times\mathcal{A}}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})$ instead of $\mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})$ in (75), but we allowed using $\mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})$ since $\mathbb{P}$ is an integrated probability measure of both $\mathbb{P}_{\mathcal{S}\times\mathcal{A}}$ and $\mathbb{P}^{(\mathbf{N})}$.

Term $\mathbb{P}(E|\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})$ can be derived by aggregating two bounds (68) and (74). That is, for $\forall\epsilon_1>0,\ \epsilon_2>0,\ u>0$,

$$\begin{aligned}\mathbb{P}(E|\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}) \geq 1-(2|\mathcal{S}\times\mathcal{A}|+6N)&\\ \times\exp\left\{\frac{-C_{21}\cdot p_{\min}\cdot N\cdot\epsilon_1^2}{d^2\cdot\left(\sup_{s,a}\|R(s,a)\|_{\psi_2}\right)^2+\gamma^2\cdot\left(\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2}\right\}&\\ -2|\mathcal{S}\times\mathcal{A}|\cdot\exp\left\{\frac{-C_{23}\cdot p_{\min}\cdot N\cdot\epsilon_2^2}{d^2\cdot\left(\sup_{s,a}\|R(s,a)\|_{\psi_2}\right)^2}\right\}-2\exp(-u^2).&\end{aligned}$$

Then it remains for us to calculate $\mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})$ in (75), and we should assume $\epsilon\in(0,1]$ as mentioned in Definition (38),

$$\mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}) \geq 1-\mathbb{P}\left(\|\hat{\mathbf{p}}_{(1)}-\mathbf{p}\|\geq\frac{1}{2}p_{\min}\cdot\epsilon\right)-\mathbb{P}\left(\|\hat{\mathbf{p}}_{(2)}-\mathbf{p}\|\geq\frac{1}{2}p_{\min}\cdot\epsilon\right), \tag{76}$$

for which we can utilize the following corollary that we proved in C.2.5, which is a special case of Theorem 6.

Corollary 2. *For* $\mathbf{X}_i\overset{iid}{\sim}\text{Multinomial}(n,\mathbf{p})$ *with* $\mathbf{p}=(p_1,\cdots,p_H)^\intercal$ *with* $\sum_{h=1}^H p_h=1$*, we have the following for* $\forall\epsilon>0$*,*

$$\mathbb{P}\big(\|\hat{\mathbf{p}}-\mathbf{p}\|\geq\epsilon\big)\leq\exp\left(\frac{1}{4}\right)\cdot\exp\left(\frac{-n\cdot\epsilon^2}{32}\right)\ \textit{where}\ \hat{\mathbf{p}}=\frac{1}{n}\sum_{i=1}^n\mathbf{X}_i.$$

If $N\geq 2$ holds (as we assumed in A.6.2), we have $\lfloor N/2\rfloor\geq N/6$ and $(N-\lfloor N/2\rfloor)\geq N/6$. Then applying the above corollary allows us to take up the probability bound (76) as follows,

$$\mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})\geq 1-C_{24}\cdot\exp(-C_{25}\cdot p_{\min}^2\cdot N\cdot\epsilon^2). \tag{77}$$

Then the probability bound (75) can be finalized as follows, with the notation $\epsilon\in(0,1]$ replaced by $\epsilon\in(0,1]$. By putting together Bounds (35), (68), and (74) gives us the following bound. For $\forall\epsilon\in(0,1],\ \epsilon_1>0,\ \epsilon_2>0,\ u>0$, we have

$$\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}},\Upsilon_\pi)\leq 8C_{\sup}B_1(\gamma)\times \tag{78}$$

$$\left[\frac{C_{10}\gamma}{\sqrt{N}}\cdot\sum_{s,a}\sqrt{b_\mu(s,a)}\cdot\left(\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t+u\cdot\operatorname{diam}(\Theta;\tilde{\eta})\right)\right. \tag{79}$$

$$+ 2\epsilon_1 + \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left(\sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)$$
$$+ 2\sqrt{p_{\min}} \cdot \epsilon \cdot \left\{ (1+\gamma) \cdot \left(\mathrm{diam}(\Theta;\tilde{\eta}) + \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right) + \sup_{s,a} \mathbb{E}\|R(s,a)\| + \epsilon_2 \right\} \Bigg],$$

with probability larger than

$$1 - (2|\mathcal{S}\times\mathcal{A}| + 6N) \cdot \exp\left\{ \frac{-C_{21} \cdot p_{\min} \cdot N \cdot \epsilon_1^2}{d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2}\right)^2 + \gamma^2 \cdot \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\|\right)^2} \right\}$$
$$- 2|\mathcal{S}\times\mathcal{A}| \cdot \exp\left\{ \frac{-C_{23} \cdot p_{\min} \cdot N \cdot \epsilon_2^2}{d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2}\right)^2} \right\} - 2\exp(-u^2)$$
$$- C_{24} \cdot \exp(-C_{25} \cdot p_{\min}^2 \cdot N \cdot \epsilon^2). \tag{80}$$

We adjust the existing variables as

$$\epsilon_1 = \sqrt{p_{\min}} \cdot \epsilon,\ \epsilon_2 = \sqrt{p_{\min}} \cdot \epsilon,\ u = \sqrt{N} \cdot p_{\min} \cdot \epsilon. \tag{81}$$

Based on following, which holds due to Cauchy-Schwartz Inequality,

$$\sum_{s,a} \sqrt{b_\mu(s,a)} \le \left(\sum_{s,a} b_\mu(s,a) \right)^{1/2} \cdot \left(\sum_{s,a} 1 \right)^{1/2} = \sqrt{|\mathcal{S}\times\mathcal{A}|} \le \frac{1}{\sqrt{p_{\min}}}, \tag{82}$$

the value within the square bracket $[\cdots]$ of Bound (78) has the following upper bound,

$$\frac{C_{10}\gamma}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \left(\int_0^\infty \sqrt{\log \mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t + \sqrt{N} \cdot p_{\min} \cdot \mathrm{diam}(\Theta;\tilde{\eta}) \cdot \epsilon \right)$$
$$+ \sqrt{p_{\min}} \cdot \epsilon + \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left(\sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)$$
$$+ 2\sqrt{p_{\min}} \cdot \epsilon \cdot \left\{ (1+\gamma) \cdot \left(\mathrm{diam}(\Theta;\tilde{\eta}) + \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right) + \sup_{s,a} \mathbb{E}\|R(s,a)\| + 1 \right\}$$
$$\le \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left(\sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)$$
$$+ \frac{C_{10}\gamma}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \int_0^\infty \sqrt{\log \mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t$$
$$+ C_{26} \cdot \sqrt{p_{\min}} \cdot (1+\gamma) \cdot \left\{ \mathrm{diam}(\Theta;\tilde{\eta}) + \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + \sup_{s,a} \mathbb{E}\|R(s,a)\| + 1 \right\} \cdot \epsilon.$$

The probability bound (80) has the following lower bound with the adjusted variables (81),

$$1 - (2|\mathcal{S}\times\mathcal{A}| + 6N) \cdot \exp\left\{ \frac{-C_{21} \cdot p_{\min}^2 \cdot N \cdot \epsilon^2}{d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2}\right)^2 + \gamma^2 \cdot \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\|\right)^2} \right\}$$
$$- 2|\mathcal{S}\times\mathcal{A}| \cdot \exp\left\{ \frac{-C_{23} \cdot p_{\min}^2 \cdot N \cdot \epsilon^2}{d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2}\right)^2} \right\} - 2\exp(-p_{\min}^2 \cdot N \cdot u^2)$$
$$- C_{24} \cdot \exp(-C_{25} \cdot p_{\min}^2 \cdot N \cdot \epsilon^2)$$
$$\ge 1 - C_{27} \cdot (|\mathcal{S}\times\mathcal{A}| + N) \cdot \exp(-C_{28} \cdot p_{\min}^2 \cdot N \cdot \epsilon^2 / C_{\mathrm{den}})$$
$$\text{with } C_{\mathrm{den}} := d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2}\right)^2 + \gamma^2 \cdot \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\|\right)^2 + 1, \tag{83}$$

Now we can choose $\theta_0 \in \Theta$ in the most favorable way,

$$\theta_0 :\in \arg\min_{\theta\in\Theta} \sup_{s,a} \mathbb{E}\|Z(s,a;\theta)\|. \tag{84}$$

If such θ_0 does not exist for some reason including the case when Θ is not closed, then we can also let it be an arbitrary value.

Then we finally obtain the following result. Under Assumptions 1, 2, 3, 4, and 5, provided that $N \geq 2$, our estimator $\hat{\theta} \in \Theta$ (13) satisfies the following bound for $\forall \epsilon \in (0, 1]$,

$$\begin{aligned}\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_{\pi}) \leq 8C_{\sup}B_1(\gamma) \times \Bigg[& \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left(\sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right) \\ & + \frac{C_{10}\gamma}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_{\mu}(s,a)} \cdot \int_0^{\infty} \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)} \mathrm{d}t \\ & + C_{26} \cdot \sqrt{p_{\min}} \cdot (1+\gamma) \cdot \left\{ \mathrm{diam}(\Theta; \tilde{\eta}) + \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + \sup_{s,a} \mathbb{E}\|R(s,a)\| + 1 \right\} \cdot \epsilon \Bigg],\end{aligned} \tag{85}$$

with probability larger than

$$1 - C_{27} \cdot (|\mathcal{S} \times \mathcal{A}| + N) \cdot \exp(-C_{28} \cdot p_{\min}^2 \cdot N \cdot \epsilon^2 / C_{\mathrm{den}}), \tag{86}$$

where the subscript of C_{den} means the denominator.

A.6.6 Simplifying the Probability term

To simplify our result, we will do some additional algebra. Letting $\epsilon = \epsilon' / \sqrt{p_{\min}}$ where $\epsilon' \in (0, p_{\min}]$, we have

$$\text{Probability (86)} = 1 - C_{27} \cdot (|\mathcal{S} \times \mathcal{A}| + N) \cdot \exp(-C_{28} \cdot p_{\min} \cdot N \cdot \epsilon'^2 / C_{\mathrm{den}})$$

By letting

$$\epsilon' = \sqrt{\frac{C_{\mathrm{den}}}{C_{28} \cdot p_{\min}}} \times \sqrt{\frac{1}{N} \log \left(\frac{C_{27} \cdot (|\mathcal{S} \times \mathcal{A}| + N)}{\delta} \right)},$$

we obtain

$$\text{Probability (86)} = 1 - \delta,$$

but we need an assumption that the sample size N is large enough to satisfy

$$\epsilon' \in (0, p_{\min}].$$

For this reason, we need N to be larger than $N_{(1)}(\delta) \in \mathbb{N}$, where $N_{(1)}(\delta)$ is defined to be the smallest integer such that $N \geq N_{(1)}(\delta)$ implies

$$\frac{1}{N} \log \left(\frac{C_{27} \cdot (|\mathcal{S} \times \mathcal{A}| + N)}{\delta} \right) \leq \frac{C_{28} \cdot p_{\min}^2}{C_{\mathrm{den}}}. \tag{87}$$

Then the bound (85) can be rewritten as

$$\begin{aligned}\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_{\pi}) \leq 8C_{\sup}B_1(\gamma) \times \Bigg[& \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left(\sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right) \\ & + \frac{C_{10}\gamma}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_{\mu}(s,a)} \cdot \int_0^{\infty} \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)} \mathrm{d}t + \sqrt{\frac{1}{N} \log \left(\frac{C_{27} \cdot (|\mathcal{S} \times \mathcal{A}| + N)}{\delta} \right)} \times \\ & C_{26}(1+\gamma) \cdot \left\{ \mathrm{diam}(\Theta; \tilde{\eta}) + \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + \sup_{s,a} \mathbb{E}\|R(s,a)\| + 1 \right\} \cdot \sqrt{\frac{C_{\mathrm{den}}}{C_{28} \cdot p_{\min}}} \Bigg].\end{aligned} \tag{88}$$

Since we have the following by Remark 1 and Inequality (62)

$$\mathbb{E}\|R(s,a)\| = \sqrt{\log 2} \cdot \left\| \mathbb{E}\|R(s,a)\| \right\|_{\psi_2} \leq \sqrt{\log 2} \cdot d \cdot \|R(s,a)\|_{\psi_2},$$

we have

$$\begin{aligned}& \left\{ \mathrm{diam}(\Theta; \tilde{\eta}) + \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + \sup_{s,a} \mathbb{E}\|R(s,a)\| + 1 \right\} \times \sqrt{C_{\mathrm{den}}} \\ & \leq 2 \cdot \sqrt{\mathrm{diam}(\Theta; \tilde{\eta})^2 + \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)^2 + \log 2 \cdot d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2} \right)^2 + 1} \cdot \sqrt{C_{\mathrm{den}}} \\ & \leq 2 \cdot \left\{ \mathrm{diam}(\Theta; \tilde{\eta})^2 + \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)^2 + d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2} \right)^2 + 1 \right\} \quad \text{by (83)}\end{aligned}$$

where we applied $(a+b+c+d)^2 \leq 4 \cdot (a^2+b^2+c^2+d^2)$ in the first inequality. Letting

$$C_{\mathrm{env}}(\Theta) := \mathrm{diam}(\Theta;\tilde{\eta})^2 + \big(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\|\big)^2 + d^2 \cdot \big(\sup_{s,a} \|R(s,a)\|_{\psi_2}\big)^2 + 1., \tag{89}$$

Then we can take up Bound (88) as follows, additionally replacing $\sum_{s,a} \sqrt{b_\mu(s,a)}$ with $\sqrt{|\mathcal{S}\times\mathcal{A}|}$ by (82) in order to enhance interpretability,

$$\begin{aligned}\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_\pi) \leq 8C_{\sup}B_1(\gamma) \times \Bigg[&\frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\sup_{s,a}\mathbb{E}\|R(s,a)\| + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)\\ &+ \frac{C_{10}\gamma}{\sqrt{N}}\cdot\sqrt{|\mathcal{S}\times\mathcal{A}|}\cdot\int_0^\infty \sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t\\ &+ \frac{C_{29}}{\sqrt{p_{\min}}}\cdot(1+\gamma)\cdot C_{\mathrm{env}}(\Theta)\cdot\sqrt{\frac{1}{N}\log\left(\frac{C_{27}\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right)}\Bigg].\end{aligned} \tag{90}$$

We now see that there are three terms all differing in order, $O(1/N)$, $O(1/\sqrt{N})$, and $O(\sqrt{\log(N/\delta)/N})$. Since the last term decays slowest with respect to N, we can further simplify it as follows. First let us define $N_{(2)}(\delta)$ to be the smallest integer such that $N \geq N_{(2)}(\delta)$ implies

$$\begin{aligned}&\frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\sup_{s,a}\mathbb{E}\|R(s,a)\| + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right) + \frac{C_{10}\gamma}{\sqrt{N}}\cdot\sqrt{|\mathcal{S}\times\mathcal{A}|}\times \\ &\int_0^\infty \sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t \leq \frac{C_{29}}{\sqrt{p_{\min}}}\cdot(1+\gamma)\cdot C_{\mathrm{env}}(\Theta)\cdot\sqrt{\frac{1}{N}\log\left(\frac{C_{27}\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right)}.\end{aligned} \tag{91}$$

Then we can rewrite Bound (90) as follows, where we skipped all the complicated calcucations where we used $1+\gamma \leq 2$,

$$\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_\pi) \leq \frac{32C_{29}}{\sqrt{p_{\min}}}\cdot C_{\sup}B_1(\gamma)\cdot C_{\mathrm{env}}(\Theta)\cdot\sqrt{\frac{1}{N}\log\left(\frac{C_{27}\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right)}.$$

A.6.7 Final Statement

Below is the final statement.

Under Assumptions 1, 2, 3, 4, and 5, for arbitrary $\delta \in (0,1)$, given large enough sample size $N \geq \max\{2, N_{(1)}(\delta), N_{(2)}(\delta)\}$, our estimator $\hat{\theta} \in \Theta$ (13) satisfies the following bound with probability larger than $1-\delta$,

$$\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_\pi) \leq \frac{C}{\sqrt{p_{\min}}}\cdot C_{\sup}\cdot B_1(\gamma)\cdot C_{\mathrm{env}}(\Theta)\cdot\sqrt{\frac{1}{N}\log\left(\frac{C_2\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right)}, \tag{92}$$

where $p_{\min} > 0$ is defined in Assumption 1, $B_1(\gamma) > 0$, $C_{\sup} > 0$, in Equations (35), (25), and $C_{\mathrm{env}}(\Theta)$, $N_{(1)}(\delta)$, $N_{(2)}(\delta)$ in Appendix A.6.8.

A.6.8 Recap of Terminologies

Here is a recap of the terminologies that were newly defined, which are repetitions of Equations (35), (83),(84), (87), (91), and (89). First, we have

$$B_1(\gamma) := \frac{1}{2(1-\gamma)}\sum_{k=1}^{\infty} 4^k\gamma^{2^{k-1}-1},$$

$$C_{\mathrm{env}}(\Theta) = \mathrm{diam}(\Theta;\tilde{\eta})^2 + \big(\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\big)^2 + d^2\cdot\big(\sup_{s,a}\|R(s,a)\|_{\psi_2}\big)^2 + 1,$$

$$C_{\mathrm{den}} := d^2\cdot\big(\sup_{s,a}\|R(s,a)\|_{\psi_2}\big)^2 + \gamma^2\cdot\big(\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\big)^2 + 1,$$

$$\theta_0 :\in \arg\min_{\theta\in\Theta}\sup_{s,a}\mathbb{E}\|Z(s,a;\theta)\|.$$

Next, let

$$N(\delta) := \max\{2, N_{(1)}(\delta), N_{(2)}(\delta)\}$$

where $N_{(1)}(\delta)$ and $N_{(2)}(\delta)$ are the smallest integers such that $N \geq N_{(1)}(\delta)$implies

$$\frac{1}{N}\log\left(\frac{C_{27}\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right) \leq \frac{C_{28}\cdot p_{\min}^2}{C_{\text{den}}}.$$

and $N \geq N_{(2)}(\delta)$implies

$$\begin{aligned}
&\frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\sup_{s,a}\mathbb{E}\|R(s,a)\|+\gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right) \\
&\qquad\qquad + \frac{C_{10}\gamma}{\sqrt{N}}\cdot\sqrt{|\mathcal{S}\times\mathcal{A}|}\cdot\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t \\
&\leq \frac{C_{29}}{\sqrt{p_{\min}}}\cdot(1+\gamma)\cdot C_{\text{env}}(\Theta)\cdot\sqrt{\frac{1}{N}\log\left(\frac{C_{27}\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right)}.
\end{aligned}$$

A.7 Special Case for parametric models under realizability

Corollary 3. (Inaccuracy for parametric model in realizable scenario) *Under Assumptions 1, 2, 3, 4, 5, and 7, for arbitrary $\delta \in (0,1)$, given large enough sample size $N \geq \max\{2, N_{(1)}(\delta), N_{(2)}(\delta)\}$ (87),* (95)*, we have following with probability larger than $1-\delta$,*

$$\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_\pi) \leq \frac{C_1}{\sqrt{p_{\min}}}\cdot C_{\sup}\cdot B_1(\gamma)\cdot C'_{\text{env}}(\Theta)\cdot\sqrt{\frac{1}{N}\log\left(\frac{C_2\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right)}. \tag{93}$$

with $B_1(\gamma) > 0$, $C_{\sup} > 0$, $C_{\text{env}}(\Theta)'$ are defined in Equations (35), (25), (94)*, and $p_{\min} > 0$ in Assumption 1.*

A.7.1 Proof

We restart from applying Assumption 7 to (90). We can use $\text{diam}(\Theta;\tilde{\eta}) \leq L\cdot\text{diam}(\Theta;\|\cdot\|)$ and the following remark (proof in C.2.6).

Remark 2. *Under Assumption 7, we have the following,*

$$\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t \leq 6\sqrt{2\pi}\cdot L\sqrt{p}\cdot\text{diam}(\Theta;\|\cdot\|).$$

Then for arbitrary $\delta\in(0,1)$, given large enough sample size $N \geq \max\{2, N_{(1)}(\delta)\}$ (87), we have the following with probability larger than $1-\delta$, with (84)

$$\begin{aligned}
&\bar{\mathcal{E}}(\Upsilon_{\hat{\theta}}, \Upsilon_\pi) \leq 8C_{\sup}B_1(\gamma)\times\left\{\frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\sup_{s,a}\mathbb{E}\|R(s,a)\|+\gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)+\frac{C_1\gamma}{\sqrt{N}}\times\right. \\
&\left.\sqrt{|\mathcal{S}\times\mathcal{A}|}\cdot L\sqrt{p}\cdot\text{diam}(\Theta;\|\cdot\|)+\sqrt{\frac{1}{N}\log\left(\frac{C_3\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right)}\cdot\frac{C_2(1+\gamma)}{\sqrt{p_{\min}}}\cdot C'_{\text{env}}(\Theta)\right\} \\
&\text{with } C'_{\text{env}}(\Theta) := L^2\cdot\text{diam}(\Theta;\|\cdot\|)^2+\left(\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^2+d^2\cdot\left(\sup_{s,a}\|R(s,a)\|_{\psi_2}\right)^2+1.
\end{aligned} \tag{94}$$

Now we will define $N_{(2)}(\delta)$ differently as follows, with θ_0 defined in (84). $N_{(2)}(\delta)$ is the smallest integer such that $N \geq N_{(2)}(\delta)$ implies

$$\begin{aligned}
&\frac{1}{N}\cdot\frac{8}{p_{\min}}\cdot\left(\sup_{s,a}\mathbb{E}\|R(s,a)\|+\gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)+\frac{C_1\gamma}{\sqrt{N}}\cdot\sqrt{|\mathcal{S}\times\mathcal{A}|}\times \\
&L\sqrt{p}\cdot\text{diam}(\Theta;\|\cdot\|)\leq\sqrt{\frac{1}{N}\log\left(\frac{C_3\cdot(|\mathcal{S}\times\mathcal{A}|+N)}{\delta}\right)}\cdot\frac{C_2(1+\gamma)}{\sqrt{p_{\min}}}\cdot C'_{\text{env}}(\Theta)\Big\}.
\end{aligned} \tag{95}$$

Then, assuming $N \geq N_{(2)}(\delta)$ we can take up (94), and it returns the desired bound of Corollary 3.

B PROOFS FOR SECTION 4

As in Appendix 3, we will use $C_k > 0$ ($k \in \mathbb{N}$) to denote appropriate universal constants throughout the proof.

B.1 EXPONENTIAL INCREASING RATE OF TRAJECTORIES

Based on how we defined $\hat{\mathcal{T}}^\pi$ based on $\hat{p}$ (10), $(\hat{\mathcal{T}}^\pi)^m \Upsilon_\theta(s,a)$ utilizes the trajectories of tuples (s,a,r,s') that can occur consecutively under the estimated probability measure $\hat{p}(\cdots|s,a)$ and the target policy $\pi(a|s)$,

$$\big(s, a, r_i^{(1)}, s_i^{(1)}, a_i^{(1)}, r_i^{(2)}, s_i^{(2)}, a_i^{(2)}, r_i^{(3)}, s_i^{(3)}, a_i^{(3)}, \cdots, r_i^{(m)}, s_i^{(m)}\big), \tag{96}$$

$$\text{where } r_i^{(t)}, s_i^{(t)} \sim \hat{p}(\cdots|s_i^{(t-1)}, a_i^{(t-1)}),\ a_i^{(t)} \sim \pi(\cdot|s_i^{(t)}) \text{ for } \forall t \geq 1, \text{ with } (s_i^{(0)}, a_i^{(0)}) = (s,a).$$

Let us first verify how many such trajectories (96) can amount to, which start from a common state-action pair s, a with length $m = 2$.

First there are $N(s,a)$ many tuples that can occur in the first step,

$$(s, a, r_i^{(1)}, s_i^{(1)}) \quad \big(1 \leq i \leq N(s,a)\big).$$

Now fix one observation with index i, and then there can be $|\mathcal{A}|$ many actions at most that can follow $s_i^{(1)}$, giving us the following tuples,

$$(s, a, r_i^{(1)}, s_i^{(1)}, a_1),\ (s, a, r_i^{(1)}, s_i^{(1)}, a_2),\ \cdots,\ (s, a, r_i^{(1)}, s_i^{(1)}, a_{|\mathcal{A}|}) \quad \big(1 \leq i \leq N(s,a)\big).$$

Now we are given with $|\mathcal{A}|$ different state-action pairs, $(s_i^{(1)}, a_1), \cdots, (s_i^{(1)}, a_{|\mathcal{A}|})$, and then the following observations of $(r_i^{(2)}, s_i^{(2)})$ can be as many as

$$\sum_{k=1}^{|\mathcal{A}|} N(s_i^{(1)}, a_k) \leq \sum_{s,a} N(s,a) = N.$$

This eventually gives us at most $N(s,a) \times N$ trajectories of length $m = 2$ starting from the given state-action pair s, a,

$$(s, a, r_i^{(1)}, s_i^{(1)}, a_i^{(1)}, r_i^{(2)}, s_i^{(2)}) \quad \big(1 \leq i \leq N(s,a) \times N\big).$$

Then we can add up for all state-action pairs that we can begin with, which leads to N^2 many trajectories at most,

$$\sum_{s,a} N(s,a) \times N = N^2.$$

We can generalize this result for an arbitrary value of $m \in \mathbb{N}$, which gives us $N(s,a) \times N^{m-1}$ many trajectories for a given state-action pair s, a,

$$(s, a, r_i^{(1)}, s_i^{(1)}, a_i^{(1)}, r_i^{(2)}, s_i^{(2)}, \cdots, s_i^{(m-1)}, a_i^{(m-1)}, r_i^{(m)}, s_i^{(m)}) \quad 1 \leq i \leq N(s,a) \times N^{m-1},$$

which further amounts to N^m many trajectories if we sum them all up for all state-action pairs as the initial point.

B.2 ESTIMATION ERROR OF MULTI-STEP BELLMAN RESIDUAL

Lemma 2. (Convergence of estimated Bellman residual) *Under Assumptions 1, 3–5, 7–8, for arbitrary $\epsilon \in (0,1]$, we have*

$$\sup_{\theta \in \Theta} \left| \hat{F}_m(\theta) - F_m(\theta) \right| \leq C_1 \cdot m^2 \cdot \frac{1}{p_{\min}^2} \cdot C_{\text{env}}^{(m)}(\Theta) \cdot \left(\frac{1}{N^{1/4}} + \sqrt{\epsilon} \right)$$

with probability larger than

$$1 - C_2 \cdot m \cdot (|\mathcal{S} \times \mathcal{A}| + N) \cdot \exp\big\{ - C_3 \cdot p_{\min}^2 \cdot N \cdot \epsilon^2 / C_{\text{den}}(m) \big\}$$

with $p_{\min} > 0$ defined in Assumption 1, $\theta_0 \in \Theta$ in (84), $C_{\text{den}}(m)$ in (132)*, and $C_{\text{env}}^{(m)}(\Theta)$ in* (138)*.*

Although larger values of step level m both increase and decrease some terms, the decreasing parts have a non-zero lower bounds $\gamma^m \sqrt{p} + 1$ and $1 + \gamma^m$ of (138). Thus it can be seen that increased values of step level m eventually leads to looser bound, necessitating larger sample size N.

B.2.1 DECOMPOSITION

We start with the following decomposition,

$$\sup_{\theta\in\Theta}\left|\hat{F}_m(\theta)-F_m(\theta)\right| = \sup_{\theta\in\Theta}\left|\bar{\mathcal{E}}\{\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta\}-\hat{\bar{\mathcal{E}}}\{\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}\right|$$
$$\leq \underbrace{\sup_{\theta\in\Theta}\left|\bar{\mathcal{E}}\{\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta\}-\bar{\mathcal{E}}\{\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}\right|}_{(Term\ 1)}$$
$$+\underbrace{\sup_{\theta\in\Theta}\left|\bar{\mathcal{E}}\{\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}-\hat{\bar{\mathcal{E}}}\{\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}\right|}_{(Term\ 2)}. \tag{97}$$

Unfortunately, we cannot bound $(Term\ 1)$ with $\sup_{\theta\in\Theta}\bar{\mathcal{E}}((\mathcal{T}^\pi)^m\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta)$ by applying triangular inequality, since $\bar{\mathcal{E}}$ is not a metric. Instead, we can devise an alternative (Lemma 3), based on the fact that $\mathcal{E}$ is in fact a squared metric (Property 3), yet we need to pay price by having square-root. Refer to C.2.7 for its proof.

Lemma 3. *For arbitrary $\Upsilon_0,\Upsilon_1,\Upsilon_2\in\mathcal{P}\{\mathcal{S}\times\mathcal{A}\}$, we have*

$$\left|\bar{\mathcal{E}}(\Upsilon_0,\Upsilon_1)-\bar{\mathcal{E}}(\Upsilon_0,\Upsilon_2)\right|\leq 4\cdot\bar{\mathcal{E}}(\Upsilon_1,\Upsilon_2)^{1/2}$$
$$\times\left[\max\left\{\bar{\mathcal{E}}(\Upsilon_0,\Upsilon_1),\bar{\mathcal{E}}(\Upsilon_0,\Upsilon_2)\right\}+\bar{\mathcal{E}}(\Upsilon_1,\Upsilon_2)\right]^{1/2}.$$

Based on the following definition,

$$\Gamma_{N,m}:=\sup_{\theta\in\Theta}\bar{\mathcal{E}}\{(\mathcal{T}^\pi)^m\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\} \tag{98}$$

applying Lemma 3 to $(Term\ 1)$ gives us

$$(Term\ 1)\leq 4\cdot\sup_{\theta\in\Theta}\Bigg(\bar{\mathcal{E}}\{(\mathcal{T}^\pi)^m\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}^{1/2}$$
$$\times\left[\max\left\{\bar{\mathcal{E}}\{\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta\},\bar{\mathcal{E}}\{\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}\right\}+\bar{\mathcal{E}}\{(\mathcal{T}^\pi)^m\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}\right]^{1/2}\Bigg)$$
$$\leq 4\cdot\Gamma_{N,m}^{1/2}\cdot\sup_{\theta\in\Theta}\Bigg[2\cdot\left\{\bar{\mathcal{E}}\{\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta\}+\bar{\mathcal{E}}\{(\mathcal{T}^\pi)^m\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}\right\}$$
$$+\bar{\mathcal{E}}\{(\mathcal{T}^\pi)^m\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\}\Bigg]^{1/2}\qquad\text{by Relaxed Triangle Inequality (32)}$$
$$\leq 8\cdot\Gamma_{N,m}^{1/2}\cdot\left\{\Gamma_{N,m}+\sup_{\theta\in\Theta}\bar{\mathcal{E}}\{\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta\}\right\}^{1/2}$$
$$\leq 8\cdot\Gamma_{N,m}^{1/2}\cdot\left\{\Gamma_{N,m}^{1/2}+\sup_{\theta\in\Theta}\bar{\mathcal{E}}\{\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta\}^{1/2}\right\}, \tag{99}$$

where the last line is due to $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for $x,y\geq 0$.

Now let us deal with $(Term\ 2)$, which can be decomposed as follows,

$$(Term\ 2)=\sup_{\theta\in\Theta}\left|\sum_{s,a}\{b_\mu(s,a)-\hat{b}_\mu(s,a)\}\cdot\mathcal{E}\left\{\Upsilon_\theta(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta(s,a)\right\}\right|$$
$$=\sup_{\theta\in\Theta}\left|\sum_{s,a}\{b_\mu(s,a)-\hat{b}_\mu(s,a)\}\cdot\left[\mathcal{E}\left\{\Upsilon_\theta(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta(s,a)\right\}\right.\right.$$
$$\left.\left.-\mathcal{E}\left\{\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}\right]\right|$$
$$+\left|\sum_{s,a}\{b_\mu(s,a)-\hat{b}_\mu(s,a)\}\cdot\mathcal{E}\left\{\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}\right|. \tag{100}$$

To handle the first line of Decomposition (100), we can first obtain the following bound, where we will use an abuse of notation $\hat{\mathcal{T}}^\pi\theta := \hat{\mathcal{T}}^\pi\Upsilon_\theta$, along with $\hat{Z}^{(m)}(s,a;\theta)$ used in (20), $\tilde{\mathbb{E}}$ in (11), and α, β indicating mutual independence between random variables with different indices (α or β),

$$
\begin{aligned}
&\left|\mathcal{E}\left\{\Upsilon_\theta(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta(s,a)\right\}-\mathcal{E}\left\{\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}\right| \\
&\quad= 2\cdot\left|\tilde{\mathbb{E}}\|Z_\alpha(s,a;\theta)-\hat{Z}_\beta^{(m)}(s,a;\theta)\|-\tilde{\mathbb{E}}\|Z_\alpha(s,a;\theta_0)-\hat{Z}_\beta^{(m)}(s,a;\theta_0)\|\right| \\
&\qquad+\left|\tilde{\mathbb{E}}\|Z_\alpha(s,a;\theta)-Z_\beta(s,a;\theta)\|-\tilde{\mathbb{E}}\|Z_\alpha(s,a;\theta_0)-Z_\beta(s,a;\theta_0)\|\right| \\
&\qquad+\left|\tilde{\mathbb{E}}\|\hat{Z}_\alpha^{(m)}(s,a;\theta)-\hat{Z}_\beta^{(m)}(s,a;\theta)\|-\tilde{\mathbb{E}}\|\hat{Z}_\alpha^{(m)}(s,a;\theta_0)-\hat{Z}_\beta^{(m)}(s,a;\theta_0)\|\right| \qquad (101) \\
&\quad\le 2\cdot\left\{\tilde{\eta}(\theta,\theta_0)+\tilde{\eta}\{(\hat{\mathcal{T}}^\pi)^m\theta,(\hat{\mathcal{T}}^\pi)^m\theta_0\}\right\}+\left\{\tilde{\eta}(\theta,\theta_0)+\tilde{\eta}(\theta,\theta_0)\right\} \\
&\qquad+\left\{\tilde{\eta}\{(\hat{\mathcal{T}}^\pi)^m\theta,(\hat{\mathcal{T}}^\pi)^m\theta_0\}+\tilde{\eta}\{(\hat{\mathcal{T}}^\pi)^m\theta,(\hat{\mathcal{T}}^\pi)^m\theta_0\}\right\}\ \text{based on Trick (49)} \\
&\quad\le 4(1+\gamma^m)\cdot\tilde{\eta}(\theta,\theta_0)\le 4(1+\gamma^m)\cdot\mathrm{diam}(\Theta;\tilde{\eta}) \qquad (102)
\end{aligned}
$$

where the second last line holds, since a Bellman operator is a contraction with respect to $\tilde{\eta}$ by Assumption 8. The second line of Decomposition (100) can be bounded by

$$
\begin{aligned}
&\left|\sum_{s,a}\{b_\mu(s,a)-\hat{b}_\mu(s,a)\}\cdot\mathcal{E}\left\{\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}\right| \\
&\le\sum_{s,a}\left|b_\mu(s,a)-\hat{b}_\mu(s,a)\right|\cdot\sup_{s,a}\mathcal{E}\left\{\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\} \\
&\le\sum_{s,a}\left|b_\mu(s,a)-\hat{b}_\mu(s,a)\right|\times \qquad \text{by Relaxed Triangle Inequality (32)} \\
&\qquad 2\cdot\sup_{s,a}\left[\mathcal{E}\left\{\Upsilon_{\theta_0}(s,a),(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}+\mathcal{E}\left\{(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}\right] \qquad (103)
\end{aligned}
$$

Plugging Inequalities (102) and (103) into (100), we obtain

$$
\begin{aligned}
&(Term\ 2)\le 4\cdot\sum_{s,a}\left|b_\mu(s,a)-\hat{b}_\mu(s,a)\right|\times\left[(1+\gamma^m)\cdot\mathrm{diam}(\Theta;\tilde{\eta})\right. \\
&\left.+\sup_{s,a}\mathcal{E}\left\{\Upsilon_{\theta_0}(s,a),(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}+\sup_{s,a}\mathcal{E}\left\{(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}\right]. \qquad (104)
\end{aligned}
$$

Now we can use the two bounds (99) and (104) to take up Decomposition (97) as follows,

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\hat{F}_m(\theta)-F_m(\theta)\right|\le(Term\ 1)+(Term\ 2) \\
&\le 8\cdot\Gamma_{N,m}^{1/2}\cdot\left\{\Gamma_{N,m}^{1/2}+\sup_{\theta\in\Theta}\bar{\mathcal{E}}\{\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta\}^{1/2}\right\} \qquad (105) \\
&\quad+4\cdot\sum_{s,a}\left|b_\mu(s,a)-\hat{b}_\mu(s,a)\right|\times\left[(1+\gamma^m)\cdot\mathrm{diam}(\Theta;\tilde{\eta})\right. \\
&\quad\left.+\sup_{s,a}\mathcal{E}\left\{\Upsilon_{\theta_0}(s,a),(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}+\sup_{s,a}\mathcal{E}\left\{(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\right\}\right].
\end{aligned}
$$

Here, we can further simplify two terms, $\sup_{s,a}\mathcal{E}\{(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\}$ and $\Gamma_{N,m}$. First let $\big(\hat{S}^{(t)}(s,a),\hat{A}^{(t)}(s,a)\big)$ be the t-th state-action pair that follows the distribution (20), that is the random state-action pair which can be reached by consecutively simulating from the estimated probability $\hat{p}(\cdots|s,a)$ and the target policy $\pi(a|s)$ starting from the initial state-action pair s,a.

Furthermore, let us denote such probability (density) as $\hat{q}^{\pi:t}_{b_\mu}(\cdots|s,a)$ that is conditioned on a fixed initial state-action pair s,a, and denote the marginalized probability as $\hat{q}^{\pi:t}_{b_\mu}(\cdots)$ that treats the initial state-action pair $S,A\sim b_\mu(s,a)$ as random. This aligns with the notation $q^{\pi:t}_{b_\mu}(\cdots)$ defined below Assumption 1. Then we have the following bound using $\tilde{\mathbb{E}}$ (11),

$$
\begin{aligned}
&\sup_{s,a}\mathcal{E}\Big\{(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^m\Upsilon_{\theta_0}(s,a)\Big\}\\
&\le\sup_{s,a}\left[m\cdot\sum_{t=0}^{m-1}\mathcal{E}\Big\{(\hat{\mathcal{T}}^\pi)^{m-t}(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}(s,a),(\hat{\mathcal{T}}^\pi)^{m-t-1}(\mathcal{T}^\pi)^{t+1}\Upsilon_{\theta_0}(s,a)\Big\}\right]\ \text{by (26)}\\
&\le m\cdot\sup_{s,a}\Bigg[\sum_{t=0}^{m-1}\gamma^{m-t-1}\cdot\mathcal{E}\Big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\mathcal{L}\big\{Z(\hat{S}^{(m-t-1)}(s,a),\hat{A}^{(m-t-1)}(s,a);\theta_0)\big\},\\
&\qquad\qquad(\mathcal{T}^\pi)^{t+1}\mathcal{L}\big\{Z(\hat{S}^{(m-t-1)}(s,a),\hat{A}^{(m-t-1)}(s,a);\theta_0)\big\}\Big\}\Bigg]\\
&\le m\cdot\sum_{t=0}^{m-1}\gamma^{m-t-1}\cdot\sup_{s,a}\tilde{\mathbb{E}}_{\hat{s}^{(m-t-1)},\hat{a}^{(m-t-1)}\sim\hat{q}^{\pi:(m-t-1)}_{b_\mu}(\cdots|s,a)}\Big\{\\
&\qquad\qquad\mathcal{E}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}(\hat{s}^{(m-t-1)},\hat{a}^{(m-t-1)}),(\mathcal{T}^\pi)^{t+1}\Upsilon_{\theta_0}(\hat{s}^{(m-t-1)},\hat{a}^{(m-t-1)})\big\}\Big\}\\
&\le m\cdot\sum_{t=0}^{m-1}\gamma^{m-t-1}\cdot\sup_{s,a}\mathcal{E}\Big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}(s,a),\mathcal{T}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}(s,a)\Big\}.
\end{aligned}
\tag{106}
$$

We also have following, where the subscripts of $\tilde{\mathbb{E}}_{\hat{q}^{\pi:m-t-1}_{b_\mu}}$ and $\tilde{\mathbb{E}}_{b_\mu}$ indicate the distribution of S,A,

$$
\begin{aligned}
\Gamma_{N,m}&=\sup_{\theta\in\Theta}\bar{\mathcal{E}}\big\{(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta,(\mathcal{T}^\pi)^m\Upsilon_\theta\big\}\\
&\le\sup_{\theta\in\Theta}\left[m\cdot\sum_{t=0}^{m-1}\bar{\mathcal{E}}\Big\{(\hat{\mathcal{T}}^\pi)^{m-t}(\mathcal{T}^\pi)^t\Upsilon_\theta,(\hat{\mathcal{T}}^\pi)^{m-t-1}(\mathcal{T}^\pi)^{t+1}\Upsilon_\theta\Big\}\right]\\
&=\sup_{\theta\in\Theta}\left[m\cdot\sum_{t=0}^{m-1}\tilde{\mathbb{E}}_{b_\mu}\mathcal{E}\Big\{(\hat{\mathcal{T}}^\pi)^{m-t}(\mathcal{T}^\pi)^t\Upsilon_\theta(S,A),(\hat{\mathcal{T}}^\pi)^{m-t-1}(\mathcal{T}^\pi)^{t+1}\Upsilon_\theta(S,A)\Big\}\right]\\
&=\sup_{\theta\in\Theta}\Bigg[m\cdot\sum_{t=0}^{m-1}\gamma^{m-t-1}\\
&\qquad\times\tilde{\mathbb{E}}_{\hat{q}^{\pi:m-t-1}_{b_\mu}}\mathcal{E}\Big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_\theta(\hat{S}^{(m-t-1)},\hat{A}^{(m-t-1)}),(\mathcal{T}^\pi)^{t+1}\Upsilon_\theta(\hat{S}^{(m-t-1)},\hat{A}^{(m-t-1)})\Big\}\Bigg]\\
&\le\sup_{\theta\in\Theta}\left[m\cdot\sum_{t=0}^{m-1}\gamma^{m-t-1}\cdot\frac{1}{p_{\min}}\cdot\tilde{\mathbb{E}}_{b_\mu}\mathcal{E}\Big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_\theta(S,A),(\mathcal{T}^\pi)^{t+1}\Upsilon_\theta(S,A)\Big\}\right]\\
&\le\frac{m}{p_{\min}}\cdot\sum_{t=0}^{m-1}\gamma^{m-t-1}\cdot\sup_{\theta\in\Theta}\bar{\mathcal{E}}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_\theta,\mathcal{T}^\pi(\mathcal{T}^\pi)^t\Upsilon_\theta\big\}\\
&\le\frac{m}{p_{\min}}\cdot\sum_{t=0}^{m-1}\gamma^{m-t-1}\cdot\Bigg\{\Big|\sup_{\theta\in\Theta}\bar{\mathcal{E}}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_\theta,\mathcal{T}^\pi(\mathcal{T}^\pi)^t\Upsilon_\theta\big\}\\
&\qquad\qquad-\bar{\mathcal{E}}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0},\mathcal{T}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}\big\}\Big|+\bar{\mathcal{E}}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0},\mathcal{T}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}\big\}\Bigg\}\\
&\le\frac{m}{p_{\min}}\cdot\sum_{t=0}^{m-1}\gamma^{m-t-1}\cdot\Bigg[\sup_{\theta\in\Theta}\Big|\bar{\mathcal{E}}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_\theta,\mathcal{T}^\pi(\mathcal{T}^\pi)^t\Upsilon_\theta\big\}\\
&\qquad-\bar{\mathcal{E}}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0},\mathcal{T}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}\big\}\Big|+\sup_{s,a}\mathcal{E}\Big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}(s,a),\mathcal{T}^\pi(\mathcal{T}^\pi)^t\Upsilon_{\theta_0}(s,a)\Big\}\Bigg],
\end{aligned}
\tag{107}
$$

where the fifth line holds, since the Radon-Nikodym derivative is bounded as follows,

$$\sup_{s,a} \frac{\hat{q}_{b_\mu}^{\pi:m-t-1}(s,a)}{b_\mu(s,a)} \leq \frac{1}{p_{\min}}.$$

Let us now define the following terms to make things more simple,

$$Y_t^{(1)} := \sup_{s,a} \mathcal{E}\Big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t \Upsilon_{\theta_0}(s,a), \mathcal{T}^\pi(\mathcal{T}^\pi)^t \Upsilon_{\theta_0}(s,a)\Big\},$$

$$Y_t^{(2)} := \sup_{\theta\in\Theta} \Big|\bar{\mathcal{E}}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t \Upsilon_\theta, \mathcal{T}^\pi(\mathcal{T}^\pi)^t \Upsilon_\theta\big\} - \bar{\mathcal{E}}\big\{\hat{\mathcal{T}}^\pi(\mathcal{T}^\pi)^t \Upsilon_{\theta_0}, \mathcal{T}^\pi(\mathcal{T}^\pi)^t \Upsilon_{\theta_0}\big\}\Big|,$$

$$\bar{\mathcal{E}}_\theta := \bar{\mathcal{E}}\big\{\Upsilon_\theta, (\mathcal{T}^\pi)^m \Upsilon_\theta\big\} \quad \& \quad \mathcal{E}_{\theta_0}(s,a) := \mathcal{E}\Big\{\Upsilon_{\theta_0}(s,a), (\mathcal{T}^\pi)^m \Upsilon_{\theta_0}(s,a)\Big\}.$$

Then we can plug Inequalities (106) and (107) into Decomposition (105), which can then be rewritten as follows,

$$\sup_{\theta\in\Theta} \Big|\hat{F}_m(\theta) - F_m(\theta)\Big| \tag{108}$$

$$\leq 8 \cdot \Big\{\frac{m}{p_{\min}} \cdot \sum_{t=0}^{m-1} \gamma^{m-t-1} \cdot \big(Y_t^{(1)} + Y_t^{(2)}\big)\Big\}^{1/2}$$

$$\times \Big[\sup_{\theta\in\Theta} \bar{\mathcal{E}}_\theta^{1/2} + \Big\{\frac{m}{p_{\min}} \cdot \sum_{t=0}^{m-1} \gamma^{m-t-1} \cdot (Y_t^{(1)} + Y_t^{(2)})\Big\}^{1/2}\Big] \tag{109}$$

$$+ 4 \cdot \sum_{s,a} \Big|b_\mu(s,a) - \hat{b}_\mu(s,a)\Big| \cdot \Big\{(1+\gamma^m) \cdot \mathrm{diam}(\Theta;\tilde{\eta}) + \sup_{s,a} \mathcal{E}_{\theta_0}(s,a)$$

$$+ m \cdot \sum_{t=0}^{m-1} \gamma^{m-t-1} \cdot Y_t^{(1)}\Big\}. \tag{110}$$

B.2.2 Bounding each variable of Bound (108)

Now we can see that there exist three random quantities

$$\sum_{s,a} \Big|b_\mu(s,a) - \hat{b}_\mu(s,a)\Big|, \;\; Y_t^{(1)}, \;\; Y_t^{(2)}, \quad \text{for } \forall t \in \{0,1,2,\cdots,m-1\},$$

and the good thing is that these are very similar to the previous proofs in A.6. We will again assume $\Omega_{\mathcal{S}\times\mathcal{A}}^{(\epsilon)}$ of Definition (38), and utilize the conditional probability $\mathbb{P}^{(\mathbf{N})}(\cdots)$. Under $\Omega_{\mathcal{S}\times\mathcal{A}}^{(\epsilon)}$ ($\epsilon \in (0,1]$), whose probability is larger than

$$\mathbb{P}(\Omega_{\mathcal{S}\times\mathcal{A}}^{(\epsilon)}) \geq 1 - C_1 \cdot \exp(-C_2 \cdot p_{\min}^2 \cdot N \cdot \epsilon^2),$$

we have

$$\sum_{s,a} \Big|b_\mu(s,a) - \hat{b}_\mu(s,a)\Big| \leq \frac{\sqrt{p_{\min}}}{2} \cdot \epsilon \quad \text{for } \epsilon \in (0,1],$$

which can be verified through derivations (69) and (77).

The remaining two terms $Y_t^{(1)}$ and $Y_t^{(2)}$ are merely repetitions of what we showed in A.6.3 that required Assumptions 3, 4, and 5, since they are in fact

$$Y_t^{(1)} = \sup_{s,a} X_{\theta_0}^{(t)}(s,a), \text{ where } X_\theta^{(t)}(s,a) := \mathcal{E}\Big\{\hat{\mathcal{T}}^\pi \Upsilon_\theta^{(t)}(s,a), \mathcal{T}^\pi \Upsilon_\theta^{(t)}(s,a)\Big\}, \;\; \Upsilon_\theta^{(t)} := (\mathcal{T}^\pi)^t \Upsilon_\theta,$$

$$Y_t^{(2)} = \sup_{\theta\in\Theta} \Big|X_\theta^{(t)} - X_{\theta_0}^{(t)}\Big|, \quad \text{where } \; X_\theta^{(t)} := \bar{\mathcal{E}}\big(\hat{\mathcal{T}}^\pi \Upsilon_\theta^{(t)}, \mathcal{T}^\pi \Upsilon_\theta^{(t)}\big) = \sum_{s,a} b_\mu(s,a) \cdot X_\theta^{(t)}(s,a), \tag{111}$$

where the notations align with Definition (40) of X_θ and $X_\theta(s,a)$. The proofs are exactly the same except that W_i^θ, W_{ij}^θ have realizations of the following forms where α, β are defined in the same way as we did above (43),

$$\begin{aligned} w_i^\theta &:= \mathbb{E}\|R_\alpha + \gamma Z_\alpha^{(t)}(S'_\alpha, A'_\alpha; \theta) - r_i - \gamma Z_\beta^{(t)}(s'_i, A'_i; \theta)\| \\ &\qquad - \mathbb{E}\|R_\alpha + \gamma Z_\alpha^{(t)}(S'_\alpha, A'_\alpha; \theta) - R_\beta - \gamma Z_\beta^{(t)}(S'_\beta, A'_\beta; \theta)\|, \\ w_{ij}^\theta &:= \mathbb{E}\|r_i + \gamma Z_\alpha^{(t)}(s'_i, A'_i; \theta) - r_j - \gamma Z_\beta^{(t)}(s'_j, A'_j; \theta)\| \\ &\qquad - \mathbb{E}\|R_\alpha + \gamma Z_\alpha^{(t)}(S'_\alpha, A'_\alpha; \theta) - R_\beta - \gamma Z_\beta^{(t)}(S'_\beta, A'_\beta; \theta)\|, \end{aligned} \tag{112}$$

where

$$Z^{(t)}(s,a;\theta) \sim (\mathcal{T}^\pi)^t \Upsilon_\theta(s,a). \tag{113}$$

One may argue that obtaining the probability bound of $Y_t^{(1)} = \sup_{s,a} X_\theta^{(t)}(s,a)$ should be more difficult than that of $X_\theta^{(t)}$, bound of which we derived in Bound (67). However, we have already derived a stronger bound that bounds $\sup_{s,a} X_\theta^{(t)}(s,a)$, as mentioned right beneath Bound (67). Therefore we can copy the probability bounds (57) and (67). Let us first allow the following abuse of notation $(\mathcal{T}^\pi)^t\Theta$, which we will define as

$$(\mathcal{T}^\pi)^t\Theta := \left\{ (\mathcal{T}^\pi)^t\theta \ : \ \theta \in \Theta \right\} \text{ for } \forall t \in \{0,1,2,\cdots,m-1\}, \text{ where } \mathcal{T}^\pi\theta := \mathcal{T}^\pi \Upsilon_\theta. \tag{114}$$

Then, under $\Omega_{\mathcal{S}\times\mathcal{A}}^{(\epsilon)}$, for a fixed value of $t \in \{0,1,2\cdots,m-1\}$, we have the following for arbitrary $\epsilon \in (0,1]$, $\epsilon_1 > 0$, $u > 0$,

$$\sum_{s,a} \left| b_\mu(s,a) - \hat{b}_\mu(s,a) \right| \le \frac{\sqrt{p_{\min}}}{2} \cdot \epsilon, \tag{115}$$

$$\begin{aligned} &\&\ Y_t^{(1)} \le 2\epsilon_1 + \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left\{ \sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z^{(t)}(s,a;\theta_0)\| \right\}, \\ &\&\ Y_t^{(2)} \le \frac{C_3\gamma}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \left\{ \int_0^\infty \sqrt{\log \mathcal{N}\big((\mathcal{T}^\pi)^t\Theta, \tilde{\eta}, z\big)}\mathrm{d}z + u \cdot \mathrm{diam}\big((\mathcal{T}^\pi)^t\Theta; \tilde{\eta}\big) \right\}, \end{aligned}$$

with probability larger than

$$\begin{aligned} &1 - 2\exp(-u^2) - (2|\mathcal{S}\times\mathcal{A}| + 6N) \\ &\qquad \times \exp\left\{ \frac{-C_4 \cdot p_{\min} \cdot N \cdot \epsilon_1^2}{\left(d \cdot \sup_{s,a} \|R(s,a)\|_{\psi_2} + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z^{(t)}(s,a;\theta_0)\|\right)^2} \right\}, \end{aligned} \tag{116}$$

where the lower bound of probability is derived by combining Inequalities (62) and (67). In order to further bound the metric entropy and the diameter based on (114), we can develop a new metric,

$$\begin{aligned} \tilde{\eta}^{(t)}(\theta_1,\theta_2) := \tilde{\eta}\big\{(\mathcal{T}^\pi)^t\theta_1, (\mathcal{T}^\pi)^t\theta_2\big\} \le \gamma^t \cdot \tilde{\eta}(\theta_1,\theta_2) \quad \text{by Assumption 8,} \\ \therefore\ \mathcal{N}\big((\mathcal{T}^\pi)^t\Theta, \tilde{\eta}, z\big) = \mathcal{N}\big(\Theta, \tilde{\eta}^{(t)}, z\big). \end{aligned}$$

Since it satisfies γ^t-Lipschitz continuity w.r.t. $\tilde{\eta}$, we can apply the logic that we used in Inequality (194) of C.2.6 to obtain the following,

$$\begin{aligned} \int_0^\infty \sqrt{\log \mathcal{N}\big((\mathcal{T}^\pi)^t\Theta, \tilde{\eta}, z\big)}\mathrm{d}z = \int_0^\infty \sqrt{\log \mathcal{N}\big(\Theta, \tilde{\eta}^{(t)}, z\big)}\mathrm{d}z \le \gamma^t \cdot \int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)}\mathrm{d}t, \\ \mathrm{diam}\big((\mathcal{T}^\pi)^t\Theta; \tilde{\eta}\big) = \mathrm{diam}(\Theta; \tilde{\eta}^{(t)}) \le \gamma^t \cdot \mathrm{diam}(\Theta; \tilde{\eta}). \end{aligned}$$

Then we can also bound the new expectation term as follows for $1 \le t \le m-1$ with $Z^{(t)}$ defined in (113),

$$\begin{aligned} \sup_{s,a} \mathbb{E}\|Z^{(t)}(s,a;\theta_0)\| &= \sup_{s,a} \mathbb{E}\|Z^{(t)}(s,a;\theta_0)\| \\ &\le \sum_{k=1}^{t} \gamma^{k-1} \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma^t \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \end{aligned} \tag{117}$$

and this is easily generalized into follows for all $0 \leq t \leq m-1$,

$$\begin{aligned}
&\sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z^{(t)}(s,a;\theta_0)\| \\
&\quad \leq \sum_{k=0}^{t} \gamma^k \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma^{t+1} \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \qquad (118) \\
&\quad \leq d \cdot \sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \|R(s,a)\|_{\psi_2} + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \text{ by Inequality (62).} \qquad (119)
\end{aligned}$$

This eventually allows to rewrite Bound (115) as follows. Under $\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}$, we have the following for arbitrary $\epsilon \in (0,1]$, $\epsilon_1 > 0$, $u > 0$, based on Line (118),

$$\begin{aligned}
&\sum_{s,a} \left| b_\mu(s,a) - \hat{b}_\mu(s,a) \right| \leq \frac{\sqrt{p_{\min}}}{2} \cdot \epsilon, \\
\&\ \ Y_t^{(1)} &\leq 2\epsilon_1 + \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left\{ \sum_{k=0}^{t} \gamma^k \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma^{t+1} \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right\}, \\
\&\ \ Y_t^{(2)} &\leq \frac{C_3 \gamma^{t+1}}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \left\{ \int_0^\infty \sqrt{\log \mathcal{N}\big(\Theta, \tilde{\eta}, z\big)} \mathrm{d}z + u \cdot \mathrm{diam}\big(\Theta; \tilde{\eta}\big) \right\}, \qquad (120)
\end{aligned}$$

with probability larger than the following, based on Line (119),

$$\begin{aligned}
&\text{Probability (116)} \geq 1 - 2\exp(-u^2) - (2|\mathcal{S}\times\mathcal{A}| + 6N) \times \\
&\qquad \exp\left\{ \frac{-C_5 \cdot p_{\min} \cdot N \cdot \epsilon_1^2}{d^2 \cdot \left(\sum_{t=0}^{m-1} \gamma^t\right)^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2}\right)^2 + \gamma^2 \cdot \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\|\right)^2} \right\}. \qquad (121)
\end{aligned}$$

B.2.3 Final Aggregation

Before taking up Decomposition (108), we can further derive the following for $\forall t \in \{0, 1, 2, \cdots, m-1\}$, based on Inequality (120),

$$\begin{aligned}
&m \cdot \sum_{t=0}^{m-1} \gamma^{m-t-1} \cdot Y_t^{(1)} \leq m \cdot \sum_{t=0}^{m-1} \gamma^{m-t-1} \cdot \left[2\epsilon_1 + \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left\{ \sum_{k=0}^{t} \gamma^k \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| \right.\right. \\
&\qquad\qquad \left.\left. + \gamma^{t+1} \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right\} \right] \\
&\leq m \sum_{t=0}^{m-1} \gamma^t \cdot \left[2\epsilon_1 + \frac{1}{N} \cdot \frac{8}{p_{\min}} \cdot \left\{ \sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right\} \right], \qquad (122)
\end{aligned}$$

where the last line holds due to $0 \leq t \leq m-1$, along with

$$\begin{aligned}
&\frac{m}{p_{\min}} \cdot \sum_{t=0}^{m-1} \gamma^{m-t-1} \cdot Y_t^{(2)} \\
&\leq \frac{m}{p_{\min}} \sum_{t=0}^{m-1} \gamma^{m-t-1} \cdot \frac{C_3 \gamma^{t+1}}{\sqrt{N}} \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \times \left\{ \int_0^\infty \sqrt{\log \mathcal{N}\big(\Theta, \tilde{\eta}, z\big)} \mathrm{d}z + u \cdot \mathrm{diam}\big(\Theta; \tilde{\eta}\big) \right\} \\
&= \frac{C_3}{\sqrt{N}} \cdot \frac{m}{p_{\min}} \sum_{t=0}^{m-1} \gamma^m \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \left\{ \int_0^\infty \sqrt{\log \mathcal{N}\big(\Theta, \tilde{\eta}, z\big)} \mathrm{d}z + u \cdot \mathrm{diam}\big(\Theta; \tilde{\eta}\big) \right\} \\
&= \frac{C_3}{\sqrt{N}} \cdot m^2 \gamma^m \cdot \frac{1}{p_{\min}} \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \left\{ \int_0^\infty \sqrt{\log \mathcal{N}\big(\Theta, \tilde{\eta}, z\big)} \mathrm{d}z + u \cdot \mathrm{diam}\big(\Theta; \tilde{\eta}\big) \right\},
\end{aligned}$$

which further leads to the following, where we skipped all the complicated calculations that required $\sum_{s,a}\sqrt{b_\mu(s,a)} \leq 1/\sqrt{p_{\min}}$ (82) and $m\sum_{t=0}^{m-1}\gamma^t \leq m^2$,

$$\begin{aligned}\Gamma_{N,m} &\leq \frac{m}{p_{\min}}\cdot\sum_{t=0}^{m-1}\gamma^{m-t-1}\cdot(Y_t^{(1)}+Y_t^{(2)}) \qquad \text{by Inequality (107) and Definition (111)}\\ &\leq E_1(m)\cdot\epsilon_1+\frac{1}{\sqrt{N}}\cdot E_2(m)\cdot u+\frac{1}{\sqrt{N}}\cdot E_3(m), \end{aligned} \tag{123}$$

where

$$\begin{aligned}E_1(m) &:= \frac{2}{p_{\min}}\cdot m\sum_{t=0}^{m-1}\gamma^t \quad \& \quad E_2(m) := C_4\cdot m^2\gamma^m\cdot\frac{1}{p_{\min}}\sum_{s,a}\sqrt{b_\mu(s,a)}\cdot\operatorname{diam}(\Theta;\tilde{\eta}),\\ E_3(m) &:= C_4\cdot m^2\cdot\frac{1}{p_{\min}^2}\cdot\Bigg\{\gamma^m\cdot\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t+\frac{1}{\sqrt{N}}\cdot\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|\\ &\qquad\qquad +\frac{1}{\sqrt{N}}\cdot\gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\Bigg\}. \end{aligned} \tag{124}$$

Now let us get back to Decomposition (108) by further simplifying Line (109) and (110). Towards that end, we will first adjust the variables as follows for arbitrary $\epsilon_0\in(0,1]$,

$$\begin{aligned}u = \sqrt{N}\cdot p_{\min}\cdot\epsilon_0,\ \ \&\ \ \epsilon_1 = \sqrt{p_{\min}}\cdot\epsilon_0,\ \ \&\\ \epsilon = \frac{\epsilon_0}{d\cdot\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\|R(s,a)\|_{\psi_2}+\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|+1}\in(0,1] \end{aligned} \tag{125}$$

Then Line (109) and Line (122) can be bounded as follows, based on Inequalities (122) and (123) respectively.

$$\begin{aligned}\text{Line (109)} &\leq 8\cdot\Bigg\{E_1(m)\cdot\sqrt{p_{\min}}\cdot\epsilon_0+E_2(m)\cdot p_{\min}\cdot\epsilon_0+\frac{1}{\sqrt{N}}\cdot E_3(m)\Bigg\}^{1/2}\\ &\qquad\times\Bigg[\sup_{\theta\in\Theta}\bar{\mathcal{E}}_\theta^{1/2}+\Bigg\{E_1(m)\cdot\sqrt{p_{\min}}\cdot\epsilon_0+E_2(m)\cdot p_{\min}\cdot\epsilon_0+\frac{1}{\sqrt{N}}\cdot E_3(m)\Bigg\}^{1/2}\Bigg]\\ &= 8\cdot\Bigg\{G(m)\cdot\epsilon_0+\frac{1}{\sqrt{N}}\cdot E_3(m)\Bigg\}^{1/2}\cdot\Bigg[\sup_{\theta\in\Theta}\bar{\mathcal{E}}_\theta^{1/2}+\Bigg\{G(m)\cdot\epsilon_0+\frac{1}{\sqrt{N}}\cdot E_3(m)\Bigg\}^{1/2}\Bigg]\\ &\qquad\qquad\text{where } G(m)\overset{let}{=}E_1(m)\cdot\sqrt{p_{\min}}+E_2(m)\cdot p_{\min}\\ &\leq 8\cdot\Bigg\{\sup_{\theta\in\Theta}\bar{\mathcal{E}}_\theta^{1/2}+\sqrt{E_3(m)}+\sqrt{G(m)}\Bigg\}\cdot\Bigg\{\frac{\sqrt{E_3(m)}}{N^{1/4}}+\sqrt{G(m)}\cdot\sqrt{\epsilon_0}\Bigg\}, \end{aligned}$$

where the last line used $\epsilon_0\in(0,1]$ and $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for $x,y\geq 0$. Next, before we deal with Line (122), we can see

$$\begin{aligned}&\mathcal{E}_{\theta_0}(s,a) = \mathcal{E}\big\{\Upsilon_{\theta_0}(s,a),(\mathcal{T}^\pi)^m\Upsilon_{\theta_0}(s,a)\big\}\\ &= 2\mathbb{E}\|Z_\alpha(s,a;\theta_0)-Z_\beta^{(m)}(s,a;\theta_0)\|-\mathbb{E}\|Z_\alpha(s,a;\theta_0)-Z_\beta(s,a;\theta_0)\|\\ &\qquad\qquad -\mathbb{E}\|Z_\alpha^{(m)}(s,a;\theta_0)-Z_\beta^{(m)}(s,a;\theta_0)\|\\ &\leq 4\mathbb{E}\|Z(s,a;\theta_0)\|+4\mathbb{E}\|Z^{(m)}(s,a;\theta_0)\|\\ &\leq 4\cdot\Bigg\{\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|+(1+\gamma^m)\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\Bigg\}, \end{aligned} \tag{126}$$

which allows us to obtain the following bound,

$$\begin{aligned}
\text{Line (110)} \leq 4 \cdot \frac{\sqrt{p_{\min}}}{2} \cdot \epsilon \cdot \Bigg[& (1+\gamma^m) \cdot \operatorname{diam}(\Theta; \tilde{\eta}) + 4 \sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| \\
& + 4(1+\gamma^m) \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + m \sum_{t=0}^{m-1} \gamma^t \cdot 2\sqrt{p_{\min}} \cdot \epsilon_0 + \frac{1}{N} \cdot m \sum_{t=0}^{m-1} \gamma^t \cdot \frac{8}{p_{\min}} \\
& \times \left\{ \sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| + \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right\} \Bigg] \\
\leq C_6 \cdot \sqrt{p_{\min}} \cdot \epsilon_0 \cdot & \left\{ \operatorname{diam}(\Theta; \tilde{\eta}) + 1 + m \sum_{t=0}^{m-1} \gamma^t + \frac{1}{N} \cdot \gamma \cdot \frac{1}{p_{\min}} \right\} \qquad \text{by Definition (125)} \\
\leq C_7 \cdot \sqrt{p_{\min}} \cdot & \left\{ \operatorname{diam}(\Theta; \tilde{\eta}) + m \sum_{t=0}^{m-1} \gamma^t \cdot \left(1 + \frac{1}{N} \cdot \frac{1}{p_{\min}} \right) \right\} \cdot \epsilon_0.
\end{aligned}$$

This allows us to take up Decomposition (108) as

$$\begin{aligned}
\sup_{\theta \in \Theta} \left| \hat{F}_m(\theta) - F_m(\theta) \right| & \leq \text{Line (109)} + \text{Line (110)} \\
& \leq 8 \cdot \left\{ \sup_{\theta \in \Theta} \bar{\mathcal{E}}_\theta^{1/2} + \sqrt{E_3(m)} + \sqrt{G(m)} \right\} \cdot \left\{ \frac{\sqrt{E_3(m)}}{N^{1/4}} + \sqrt{G(m)} \cdot \sqrt{\epsilon_0} \right\} \\
& \quad + C_7 \cdot \sqrt{p_{\min}} \cdot \left\{ \operatorname{diam}(\Theta; \tilde{\eta}) + m \sum_{t=0}^{m-1} \gamma^t \cdot \left(1 + \frac{1}{N} \cdot \frac{1}{p_{\min}} \right) \right\} \cdot \epsilon_0. \qquad (127)
\end{aligned}$$

Now let us bound $\sup_{\theta \in \Theta} \bar{\mathcal{E}}_\theta := \sup_{\theta \in \Theta} \bar{\mathcal{E}}\{\Upsilon_\theta, (\mathcal{T}^\pi)^m \Upsilon_\theta\}$, based on

$$\sup_{\theta \in \Theta} \bar{\mathcal{E}}_\theta = \sup_{\theta \in \Theta} \left| \bar{\mathcal{E}}_\theta - \bar{\mathcal{E}}_{\theta_0} \right| + \bar{\mathcal{E}}_{\theta_0}.$$

The first term can be bounded with the same trick (102)

$$\sup_{\theta \in \Theta} \left| \bar{\mathcal{E}}_\theta - \bar{\mathcal{E}}_{\theta_0} \right| \leq \sup_{\theta \in \Theta} \left\{ \sum_{s,a} b_\mu(s,a) \cdot \left| \mathcal{E}_\theta(s,a) - \mathcal{E}_{\theta_0}(s,a) \right| \right\} \leq 4(1+\gamma^m) \cdot \operatorname{diam}(\Theta; \tilde{\eta}),$$

The second term can be bounded as following using (126),

$$\bar{\mathcal{E}}_{\theta_0} \leq 4 \cdot \left\{ \sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| + (1+\gamma^m) \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right\}.$$

Then we eventually have

$$\begin{aligned}
\sup_{\theta \in \Theta} \bar{\mathcal{E}}_\theta \leq 4 \cdot \Bigg\{ & (1+\gamma^m) \cdot \operatorname{diam}(\Theta; \tilde{\eta}) + \sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| \\
& + (1+\gamma^m) \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \Bigg\} \overset{let}{=} H(m). \qquad (128)
\end{aligned}$$

This allows us to rewrite Decomposition (127) as

$$\begin{aligned}
\sup_{\theta \in \Theta} \left| \hat{F}_m(\theta) - F_m(\theta) \right| \leq 8 \cdot & \left\{ \sqrt{E_3(m)} + \sqrt{G(m)} + \sqrt{H(m)} \right\} \cdot \left\{ \frac{\sqrt{E_3(m)}}{N^{1/4}} + \sqrt{G(m)} \cdot \sqrt{\epsilon_0} \right\} \\
& + C_7 \cdot \sqrt{p_{\min}} \cdot \left\{ \operatorname{diam}(\Theta; \tilde{\eta}) + m \sum_{t=0}^{m-1} \gamma^t \cdot \left(1 + \frac{1}{N} \cdot \frac{1}{p_{\min}} \right) \right\} \cdot \epsilon_0. \qquad (129)
\end{aligned}$$

The probability bound (121) can be integrated for all $t \in \{0, 1, 2, \cdots, m-1\}$,

$$\begin{aligned}
& 1 - 2m \cdot \exp(-u^2) - m \cdot (2|\mathcal{S} \times \mathcal{A}| + 6N) \qquad (130) \\
& \quad \times \exp\left\{ \frac{-C_5 \cdot p_{\min} \cdot N \cdot \epsilon_1^2}{d^2 \cdot \left(\sum_{t=0}^{m-1} \gamma^t \right)^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2} \right)^2 + \gamma^2 \cdot \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)^2} \right\},
\end{aligned}$$

but still conditioned on $\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}$. That is, denoting the event (129) as E, we have $\mathbb{P}(E|\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}) \geq$(130). Then we can employ the same trick (75) to obtain

$$\begin{aligned}
\mathbb{P}(E) &\geq \mathbb{P}(E \cap \Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}) \geq 1 - (1 - \mathbb{P}(\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})) - (1 - \mathbb{P}(E|\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}})) \qquad (131)\\
&\geq 1 - C_8 \cdot \exp(-C_9 \cdot p_{\min}^2 \cdot N \cdot \epsilon^2) - 2m \cdot \exp(-u^2) - m \cdot (2|\mathcal{S}\times\mathcal{A}| + 6N) \quad \text{by Bound (77)}\\
&\qquad \times \exp\left\{ \frac{-C_5 \cdot p_{\min} \cdot N \cdot \epsilon_1^2}{d^2 \cdot \big(\sum_{t=0}^{m-1} \gamma^t\big)^2 \cdot \big(\sup_{s,a} \|R(s,a)\|_{\psi_2}\big)^2 + \gamma^2 \cdot \big(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\|\big)^2} \right\}\\
&\geq 1 - C_{10} \cdot m \cdot (|\mathcal{S}\times\mathcal{A}| + N) \times \exp\big\{ - C_{11} \cdot p_{\min}^2 \cdot N \cdot \epsilon_0^2 / C_{\text{den}}(m)\big\} \qquad \text{by Equation (125),}
\end{aligned}$$

where

$$C_{\text{den}}(m) := d^2 \cdot \big(\sum_{t=0}^{m-1} \gamma^t\big)^2 \cdot \big(\sup_{s,a} \|R(s,a)\|_{\psi_2}\big)^2 + \gamma^2 \cdot \big(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\|\big)^2 + 1. \qquad (132)$$

Now we can switch the notation ϵ_0 into ϵ. Then so far, we have verified that we have the following bound for $\forall \epsilon \in (0,1]$,

$$\sup_{\theta\in\Theta} \left| \hat{F}_m(\theta) - F_m(\theta) \right| \leq 8 \cdot \left\{ \sqrt{E(m)} + \sqrt{G(m)} + \sqrt{H(m)} \right\} \cdot \left\{ \frac{\sqrt{E(m)}}{N^{1/4}} + \sqrt{G(m)} \cdot \sqrt{\epsilon} \right\} \qquad (133)$$

$$+ C_7 \cdot \sqrt{p_{\min}} \cdot \left\{ \text{diam}(\Theta;\tilde{\eta}) + m \sum_{t=0}^{m-1} \gamma^t \cdot \left(1 + \frac{1}{N} \cdot \frac{1}{p_{\min}} \right) \right\} \cdot \epsilon, \qquad (134)$$

with probability larger than

$$1 - C_{10} \cdot m \cdot (|\mathcal{S}\times\mathcal{A}| + N) \cdot \exp\left\{ - C_{11} \cdot p_{\min}^2 \cdot N \cdot \epsilon^2 / C_{\text{den}}(m) \right\},$$

where we have

$$\begin{aligned}
E(m) &= C_4 \cdot m^2 \cdot \frac{1}{p_{\min}^2} \cdot \left\{ \gamma^m \cdot \int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)} \mathrm{d}t + \frac{1}{\sqrt{N}} \cdot \sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| \right.\\
&\qquad \left. + \frac{1}{\sqrt{N}} \cdot \gamma \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right\}.\\
G(m) &= \frac{2}{\sqrt{p_{\min}}} \cdot m \sum_{t=0}^{m-1} \gamma^t + C_4 \cdot m^2 \gamma^m \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot \text{diam}(\Theta;\tilde{\eta})\\
H(m) &= 4 \cdot \left\{ (1+\gamma^m) \cdot \text{diam}(\Theta;\tilde{\eta}) + \sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| \right.\\
&\qquad \left. + (1+\gamma^m) \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right\}.
\end{aligned}$$

Now let us further simplify Lines (133) and (134), since they are complicated.

$$\begin{aligned}
\text{(Line 133)} &\leq 8 \cdot \left\{ \sqrt{E(m)} + \sqrt{G(m)} + \sqrt{H(m)} \right\} \cdot \left\{ \sqrt{E(m)} + \sqrt{G(m)} \right\} \cdot \left(\frac{1}{N^{1/4}} + \sqrt{\epsilon} \right)\\
&\leq C_{12} \cdot \left\{ E(m) + G(m) + H(m) \right\} \cdot \left(\frac{1}{N^{1/4}} + \sqrt{\epsilon} \right) \qquad (135)
\end{aligned}$$

Since we have

$$\begin{aligned}
E(m) + G(m) &\leq C_{13} \cdot m^2 \cdot \frac{1}{p_{\min}^2} \cdot \left\{ \gamma^m \cdot \int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)} \mathrm{d}t + \gamma^m \cdot \text{diam}(\Theta;\tilde{\eta}) \right.\\
&\qquad \left. + \frac{1}{\sqrt{N}} \cdot \left(\sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| + \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right) + 1 \right\},
\end{aligned}$$

due to $m\sum_{t=0}^{m-1}\gamma^t \le m^2$ and Inequality (82). Then we can get the following,

$$
\begin{aligned}
&E(m)+G(m)+H(m)\\
&\le C_{13}\cdot m^2\cdot\frac{1}{p_{\min}^2}\cdot\Bigg\{\gamma^m\cdot\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t+\gamma^m\cdot\mathrm{diam}(\Theta;\tilde{\eta})\\
&+\frac{1}{\sqrt{N}}\cdot\Bigg(\sum_{t=0}^{m-1}\gamma^t\sup_{s,a}\mathbb{E}\|R(s,a)\|+\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\Bigg)+1\Bigg\}\\
&+4\cdot\Bigg\{(1+\gamma^m)\cdot\mathrm{diam}(\Theta;\tilde{\eta})+\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|+(1+\gamma^m)\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\Bigg\}\\
&\le C_{14}\cdot m^2\cdot\frac{1}{p_{\min}^2}\cdot\Bigg\{\gamma^m\cdot\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t+\mathrm{diam}(\Theta;\tilde{\eta})\\
&\qquad\qquad+\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|+(1+\gamma^m)\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|+1\Bigg\}.
\end{aligned}
$$

which further allows us to incorporate with (134) as follows,

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\hat{F}_m(\theta)-F_m(\theta)\right|\le\text{Line (133)}+\text{Line (134)}\\
&\quad\le C_{14}\cdot m^2\cdot\frac{1}{p_{\min}^2}\cdot\Bigg\{\gamma^m\cdot\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t+\mathrm{diam}(\Theta;\tilde{\eta})\\
&\qquad\qquad+\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|+(1+\gamma^m)\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|+1\Bigg\}\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right)\\
&\qquad+C_7\cdot\sqrt{p_{\min}}\cdot\left\{\mathrm{diam}(\Theta;\tilde{\eta})+m\sum_{t=0}^{m-1}\gamma^t\cdot\left(1+\frac{1}{N}\cdot\frac{1}{p_{\min}}\right)\right\}\cdot\epsilon\\
&\quad\le C_{15}\cdot m^2\cdot\frac{1}{p_{\min}^2}\cdot\Bigg\{\gamma^m\cdot\int_0^\infty\sqrt{\log\mathcal{N}(\Theta,\tilde{\eta},t)}\mathrm{d}t+\mathrm{diam}(\Theta;\tilde{\eta})\qquad(\because\ \epsilon\le\sqrt{\epsilon}\le 1)\\
&\qquad\qquad+\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|+(1+\gamma^m)\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|+1\Bigg\}\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right)\\
&\quad\le C_{16}\cdot m^2\cdot\frac{1}{p_{\min}^2}\cdot\Bigg\{L\cdot(\gamma^m\sqrt{p}+1)\cdot\mathrm{diam}(\Theta;\|\cdot\|)\quad\text{by Remark 2 and Assumption 7}\\
&\qquad\qquad+\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|+(1+\gamma^m)\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|+1\Bigg\}\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right).
\end{aligned}
$$

Now we have the following result for an arbitrary $\epsilon\in(0,1]$,

$$
\sup_{\theta\in\Theta}\left|\hat{F}_m(\theta)-F_m(\theta)\right|\le C_{16}\cdot m^2\cdot\frac{1}{p_{\min}^2}\cdot C_{\mathrm{env}}^{(m)}(\Theta)\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right)\tag{136}
$$

with probability larger than

$$
1-C_{10}\cdot m\cdot(|\mathcal{S}\times\mathcal{A}|+N)\cdot\exp\big\{-C_{11}\cdot p_{\min}^2\cdot N\cdot\epsilon^2/C_{\mathrm{den}}(m)\big\}\tag{137}
$$

where

$$
\begin{aligned}
C_{\mathrm{env}}^{(m)}(\Theta):=&\,L\cdot(\gamma^m\sqrt{p}+1)\cdot\mathrm{diam}(\Theta;\|\cdot\|)+\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|\\
&+(1+\gamma^m)\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|+1
\end{aligned}\tag{138}
$$

This gives us the desired result of Lemma 2.

B.3 OBTAINING THE BOUND OF BOOSTRAP-BASED OBJECTIVE FUNCTION (23)

Our final estimator of the objective function $\hat{F}_m^{(B)}$ (23) is based on bootstrap, not $\hat{F}_m$ (21) covered in Lemma 2. So we shall develop it into following.

Lemma 4. *Under same assumptions of Lemma 2, for a fixed $m \in \mathbb{N}$ and arbitrary $\epsilon, \epsilon' \in (0, 1]$,*

$$\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta) - F_m(\theta)\right| \leq \frac{C_{32}}{p_{\min}^2}\cdot B_{\text{env}}^{(m)}(\Theta)\cdot\left\{m^2\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right)+m\cdot\left(\frac{1}{M^{1/4}}+\sqrt{\epsilon'}\right)\right\}$$

holds with probability larger than

$$\begin{aligned}1-\mathcal{D}(N)-C_1\cdot m\cdot(|\mathcal{S}\times\mathcal{A}|+N)\cdot\exp(-C_2\cdot p_{\min}^2\cdot N\cdot\epsilon^2/C_{\text{den}}(m))\\ -C_1\cdot(|\mathcal{S}\times\mathcal{A}|+M)\cdot\exp(-C_2\cdot p_{\min}^2\cdot M\cdot\epsilon'^2/B_{\text{den}}(m)),\end{aligned}$$

where $C_{\text{den}}(m)$ and $B_{\text{den}}(m)$ are defined in (132) *and* (169)*, and $\mathcal{D}(N)\to 0$ as in* (161)*.*

B.3.1 THREE STAGES OF PROBABILITY SPACE

We can decompose the term $\sup_{\theta\in\Theta}|\hat{F}_m^{(B)}(\theta)-F_m(\theta)|$ as follows,

$$\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-F_m(\theta)\right|\leq\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-\hat{F}_m(\theta)\right|+\sup_{\theta\in\Theta}\left|\hat{F}_m(\theta)-F_m(\theta)\right|. \tag{139}$$

At this point, we should recognize that our probability space (36) is expanded due to bootstrapping procedure reflected in $\hat{F}_m^{(B)}$. Now our probability space $(\Omega,\Sigma,\mathbb{P})$ can be factorized into three stages,

$$\text{Stage 1: } (\Omega_{\mathcal{S}\times\mathcal{A}},\Sigma_{\mathcal{S}\times\mathcal{A}},\mathbb{P}_{\mathcal{S}\times\mathcal{A}})\Rightarrow\text{determines which state-action pairs } S_i,A_i \text{ are sampled,} \tag{140}$$

$$\text{Stage 2: } (\Omega^{(\mathbf{N})},\Sigma^{(\mathbf{N})},\mathbb{P}^{(\mathbf{N})})\Rightarrow\text{conditioned on } (S_i,A_i),\ \text{determines } R_i,S_i'\sim p(\cdots|S_i,A_i),$$

$$\text{Stage 3: } (\Omega_B^{(\mathcal{D})},\Sigma_B^{(\mathcal{D})},\mathbb{P}_B^{(\mathcal{D})})\Rightarrow\text{conditioned on } \mathcal{D},\ \text{determines the bootstrapped trajectories in (22).}$$

We have already bounded $\sup_{\theta\in\Theta}|\hat{F}_m(\theta)-F_m(\theta)|$ of (139) in Lemma 2, which is controlled by Stage 1 and 2 probability spaces (140). Now the remaining term $\sup_{\theta\in\Theta}|\hat{F}_m^{(B)}(\theta)-\hat{F}_m(\theta)|$ of (139) is solely based on Stage 3 probability space (140), conditioned on the observed data $\mathcal{D}=\{(s_i,a_i,r_i,s_i')\}_{i=1}^N$. However, since the bootstrapped probability space (Stage 3) is affected by what was observed in the previous two stages, we will assume some nice properties are satisfied in Stage 1 and Stage 2 probability spaces, which are already mentioned within the proof of Lemma 2 in B.2.

B.3.2 INHERITED RESULTS FROM LEMMA 2

Here we will define two events. The first event can be viewed as an equivalent event to $\Omega_{\mathcal{S}\times\mathcal{A}}^{(\epsilon)}$ (38)

$$E_{1,a}:=\left\{\omega\in\Omega_{\mathcal{S}\times\mathcal{A}}:\text{Facts (39) are satisfied.}\right\}. \tag{141}$$

Next, we will define the second event $E_{1,b}$ where two things are satisfied. We will inherit (123) and modify it according to (125), which leads to following based on Definitions (124),

$$\Gamma_{N,m}\leq E_1(m)\cdot\sqrt{p_{\min}}\cdot\epsilon+E_2(m)\cdot p_{\min}\cdot\epsilon+\frac{1}{\sqrt{N}}\cdot E_3(m). \tag{142}$$

Note that we switched the notation $\epsilon_0\in(0,1]$ with ϵ, as they did right before Line (133). We will also inherit the final result (136),

$$\sup_{\theta\in\Theta}\left|\hat{F}_m(\theta)-F_m(\theta)\right|\leq C_1\cdot m^2\cdot\frac{1}{p_{\min}^2}\cdot C_{\text{env}}^{(m)}(\Theta)\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right). \tag{143}$$

Now we will define a new event

$$E_{1,b}:=\left\{\omega\in\Omega^{(\mathbf{N})}:\text{Both (142) and (143) hold.}\right\}, \tag{144}$$

We have derived in (131) that

$$\begin{aligned}&1-\mathbb{P}(E_{1,a}^c)-\mathbb{P}(E_{1,b}^c|E_{1,a})\\ &\qquad\geq 1-C_2\cdot m\cdot(|\mathcal{S}\times\mathcal{A}|+N)\cdot\exp\left\{-C_3\cdot p_{\min}^2\cdot N\cdot\epsilon^2/C_{\text{den}}(m)\right\},\end{aligned} \tag{145}$$

since E and $\Omega_{\mathcal{S}\times\mathcal{A}}^{(\epsilon)}$ of (131) can be switched into $E_{1,a}$ (141) and $E_{1,b}$ (144).

B.3.3 IMPLICATIONS OF STATEMENTS IN B.3.2

Let us assume that the events $E_{1,a}$ and $E_{1,b}$ both hold, and then bound the term $\sup_{\theta\in\Theta}|\hat{F}_m^{(B)}(\theta) - \hat{F}_m(\theta)|$ of (139). We need to emphasize that at this moment, $\mathcal{D}$ is given, and Stage 3 probability space (140) is the only source of probability. In other words, we can consider $\hat{F}_m$ as our population objective function, which is based upon $\hat{b}_\mu$ (10) and $\hat{p}_m(\cdots|s,a)$. $\hat{p}_m(\cdots|s,a)$ represents the empirical measure of $(\sum_{t=0}^{m-1}\gamma^t\hat{R}^{(t)}, \hat{S}^{(m)})$ conditioned on initial state-action pair s,a that can occur by applying $\hat{p}(r,s'|s,a)$ and $\pi(a|s)$ for m consecutive times (20). In other words, by treating $(\hat{\mathcal{T}}^\pi)^m$ as the population operator and $\mathcal{B}_m$ as its approximation, we can obtain

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta) - \hat{F}_m(\theta)\right| &= \sup_{\theta\in\Theta}\left|\hat{\bar{\mathcal{E}}}\big(\Upsilon_\theta, \mathcal{B}_m\Upsilon_\theta\big) - \hat{\bar{\mathcal{E}}}\big\{\Upsilon_\theta, (\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\big\}\right| \\ &\le 8\cdot\tilde{\Gamma}_{B,m}^{1/2}\cdot\left\{\tilde{\Gamma}_{B,m}^{1/2} + \sup_{\theta\in\Theta}\hat{\bar{\mathcal{E}}}\big\{\Upsilon_\theta, (\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\big\}^{1/2}\right\} \text{ by Derivation (99),}\end{aligned}$$

where we have a new term that we will refer to as *bootstrap discrepancy*

$$\tilde{\Gamma}_{B,m} := \sup_{\theta\in\Theta}\hat{\bar{\mathcal{E}}}\big\{(\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta, \mathcal{B}_m\Upsilon_\theta\big\}. \tag{146}$$

Since the other term can be further bounded as

$$\begin{aligned}\sup_{\theta\in\Theta}\hat{\bar{\mathcal{E}}}\big\{\Upsilon_\theta, (\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\big\} &\le \frac{3}{2}\sup_{\theta\in\Theta}\bar{\mathcal{E}}\big\{\Upsilon_\theta, (\hat{\mathcal{T}}^\pi)^m\Upsilon_\theta\big\} \text{ by Facts (39) in B.3.2} \\ &\le 3\cdot\sup_{\theta\in\Theta}\bar{\mathcal{E}}_\theta + 3\cdot\Gamma_{N,m} \text{ by Relaxed Triangle Inequality (32),}\end{aligned}$$

where $\Gamma_{N,m}$ is defined in (98). Then we obtain

$$\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta) - \hat{F}_m(\theta)\right| \le C_4\cdot\tilde{\Gamma}_{B,m}^{1/2}\cdot\left\{\tilde{\Gamma}_{B,m}^{1/2} + \sup_{\theta\in\Theta}\bar{\mathcal{E}}_\theta^{1/2} + \Gamma_{N,m}^{1/2}\right\}. \tag{147}$$

Let us bound the three terms one by one. First, we can bound the supremum term as follows,

$$\begin{aligned}\sup_{\theta\in\Theta}\bar{\mathcal{E}}_\theta^{1/2} \le 2\cdot\Bigg\{&\sqrt{1+\gamma^m}\cdot\sqrt{L}\cdot\sqrt{\mathrm{diam}(\Theta;\|\cdot\|)} + \left(\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|\right)^{1/2} \\ &+ \left((1+\gamma^m)\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^{1/2}\Bigg\} \text{ by Inequality (128) and Assumption 7.}\end{aligned} \tag{148}$$

Based on what we have in B.3.2, we can further bound Bellman discrepancy as follows,

$$\begin{aligned}\Gamma_{N,m}^{1/2} &\le \left\{E_1(m)\cdot\sqrt{p_{\min}}\cdot\epsilon + E_2(m)\cdot p_{\min}\cdot\epsilon + \frac{1}{\sqrt{N}}\cdot E_3(m)\right\}^{1/2} \text{ by Inequality (142)} \\ &\le \Bigg[\frac{C_6}{\sqrt{p_{\min}}}\cdot m^2\cdot\Big\{1+\gamma^m\cdot L\cdot\mathrm{diam}(\Theta;\|\cdot\|)\Big\}\cdot\epsilon + \frac{C_7}{\sqrt{N}}\cdot\frac{m^2}{p_{\min}^2}\cdot L\sqrt{p}\cdot\gamma^m\cdot\mathrm{diam}(\Theta;\|\cdot\|) \\ &\quad + \frac{C_7}{N}\cdot\frac{m^2}{p_{\min}^2}\cdot\left(\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\| + \gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)\Bigg]^{1/2} \text{ by Equation (124)} \\ &\le C_8\cdot\frac{m}{p_{\min}^{1/4}}\cdot\Big\{1+\gamma^{m/2}\cdot\sqrt{L}\cdot\sqrt{\mathrm{diam}(\Theta;\|\cdot\|)}\Big\}\cdot\sqrt{\epsilon} \quad (\because\ \sqrt{x+y+z}\le\sqrt{x}+\sqrt{y}+\sqrt{z}) \\ &\quad + \frac{C_8}{N^{1/4}}\cdot\frac{m}{p_{\min}}\cdot\sqrt{L}\cdot p^{1/4}\cdot\gamma^{m/2}\cdot\sqrt{\mathrm{diam}(\Theta;\|\cdot\|)} \\ &\quad + \frac{C_8}{\sqrt{N}}\cdot\frac{m}{p_{\min}}\cdot\left\{\left(\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|\right)^{1/2} + \left(\gamma\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right)^{1/2}\right\},\end{aligned} \tag{149}$$

where the second last inequality can be derived by putting together Assumption 7, Inequality (82), and Remark 2. Since Bounds (148) and (149) hold under what we already have in B.3.2, so there is no additional probability term that we have to subtract from the probability (145).

B.3.4 BOUNDING BOOTSTRAP DISCREPANCY

In further bounding (147), bootstrap discrepancy $\tilde{\Gamma}_{B,m}$ is the only term is probabilistic due to Stage 3 probability space (140). Comparing (20) and (22), we can see that $\mathcal{B}_m$ is in fact the single-step estimator of $(\hat{\mathcal{T}}^\pi)^m$ that can be viewed as a new population operator in the new probability space generated by bootstrapping from the already-observed data $\mathcal{D}$. In this regard, the relationship between $\mathcal{B}_m$ and $(\hat{\mathcal{T}}^\pi)^m$ aligns with that between $\hat{\mathcal{T}}^\pi$ and $\mathcal{T}^\pi$, only with a few differences. The reward R is replaced by discounted sum $\sum_{t=1}^m \gamma^{t-1}\hat{R}^{(t)}$, S' is replaced by $\hat{S}^{(m)}$, A' is replaced by $\hat{A}^{(m)}$, and the discount rate γ is replaced by γ^m. In addition, several other quantities are also replaced as follows,

$$
\begin{aligned}
b_\mu(s,a) \leftarrow \hat{b}_\mu(s,a), \quad \mathbb{E}(\cdots) \leftarrow \tilde{\mathbb{E}}(\cdots), \quad \|\cdot\|_{\psi_2} \leftarrow \|\cdot\|_{\tilde{\psi}_2}, \quad N \leftarrow M, \\
p_{\min} \leftarrow \hat{p}_{\min} := \min\{\hat{b}_\mu(s,a) \ : \ \hat{b}_\mu(s,a) > 0\} = \min\{\hat{b}_\mu(s,a)\} \ \text{ by Facts (39).}
\end{aligned} \tag{150}
$$

where $\tilde{\mathbb{E}}(\cdots)$ (11) and $\|\cdot\|_{\tilde{\psi}_2}$ are the expectation and subgaussian norms corresponding to the conditional probability measure $\mathbb{P}(\cdots|\mathcal{D})$. With the replacements by the estimated quantities (that will now be regarded as a new population quantity in Stage 3 probability space), we can replicate the proofs of A.6.3.

Analogous to Bound (68), for arbitrary values of $\epsilon_1' > 0$ and $u' > 0$,

$$
\begin{aligned}
\tilde{\Gamma}_{B,m} \leq \frac{C_9\gamma^m}{\sqrt{M}} \cdot \sum_{s,a} \sqrt{\hat{b}_\mu(s,a)} \cdot \left(\int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)}\mathrm{d}t + u' \cdot \mathrm{diam}(\Theta; \tilde{\eta}) \right) \\
+ 2\epsilon_1' + \frac{1}{M} \cdot \frac{8}{\hat{p}_{\min}} \cdot \left(\sup_{s,a} \tilde{\mathbb{E}} \left\| \sum_{t=1}^m \gamma^{t-1} \hat{R}^{(t)}(s,a) \right\| + \gamma^m \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right) \Bigg\},
\end{aligned} \tag{151}
$$

where the random variable (vector) $\hat{R}^{(t)}(s,a)$ is the same term with $\hat{R}^{(t)}$ defined in Definition (20), with probability larger than

$$
\begin{aligned}
1 - 2\exp(-{u'}^2) - (2|\mathcal{S}\times\mathcal{A}| + 6M)\times \\
\exp\left\{ \frac{-C_{10} \cdot \hat{p}_{\min} \cdot M \cdot {\epsilon_1'}^2}{\left(\sup_{s,a} \left\| \| \sum_{t=1}^m \gamma^{t-1} \hat{R}^{(t)}(s,a) \| \right\|_{\tilde{\psi}_2} + \gamma^m \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)^2} \right\} \\
- C_{11} \cdot \exp(-C_{12} \cdot \hat{p}_{\min}^2 \cdot N \cdot {\epsilon_2'}^2).
\end{aligned} \tag{152}
$$

where Line (152) is added because of conditioning on that each s, a is observed sufficiently many times as initial state-action pairs, which is analogous to $\Omega^{(\epsilon)}_{\mathcal{S}\times\mathcal{A}}$ in Bound (77).

Now adjusting the variables as follows with $\epsilon' \in (0, 1]$,

$$
\begin{aligned}
\epsilon_1' = \sqrt{\hat{p}_{\min}} \cdot \epsilon' \ \ \& \ \ u' = \sqrt{M} \cdot \hat{p}_{\min} \cdot \epsilon' \ \ \& \\
\epsilon_2' = \frac{\epsilon'}{\left(\sup_{s,a} \left\| \| \sum_{t=1}^m \gamma^{t-1} \hat{R}^{(t)}(s,a) \| \right\|_{\tilde{\psi}_2} + \gamma^m \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + 1 \right)^2} \in (0, 1],
\end{aligned}
$$

we can take up Bound (151) as follows,

$$
\begin{aligned}
\tilde{\Gamma}_{B,m} \leq \frac{1}{M} \cdot \frac{8}{\hat{p}_{\min}} \cdot \left(\sum_{t=1}^m \gamma^{t-1} \cdot \sup_{s,a} \tilde{\mathbb{E}}\|\hat{R}(s,a)\| + \gamma^m \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right) \\
+ \frac{C_{13}}{\sqrt{M}} \cdot \gamma^m \cdot \sum_{s,a} \sqrt{\hat{b}_\mu(s,a)} \cdot L\sqrt{p} \cdot \mathrm{diam}(\Theta; \|\cdot\|) + C_{14} \cdot (1 + \gamma^m) \cdot \sqrt{\hat{p}_{\min}} \cdot \epsilon',
\end{aligned}
$$

where we used $\sum_{s,a}\sqrt{\hat{b}_\mu(s,a)} \leq 1/\sqrt{\hat{p}_{\min}}$ (analogous to (82)) and Remark 2. The probability bound (152) has the following lower bound,

$$
\begin{aligned}
&1 - C_{15}\cdot(|\mathcal{S}\times\mathcal{A}| + M) \\
&\qquad \times \exp\left\{ \frac{-C_{16}\cdot \hat{p}_{\min}^2 \cdot M \cdot {\epsilon'}^2}{\left(\sup_{s,a}\left\|\|\sum_{t=1}^m \gamma^{t-1}\hat{R}^{(t)}(s,a)\|\right\|_{\tilde{\psi}_2} + \gamma^m \cdot \sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\| + 1\right)^2}\right\} \\
&\geq 1 - C_{15}\cdot(|\mathcal{S}\times\mathcal{A}| + M) \\
&\qquad \times \exp\left\{ \frac{-C_{16}\cdot \hat{p}_{\min}^2 \cdot M \cdot {\epsilon'}^2}{\left(\sum_{t=1}^m \gamma^{t-1}\cdot\sup_{s,a}\left\|\|\hat{R}^{(t)}(s,a)\|\right\|_{\tilde{\psi}_2} + \gamma^m \cdot \sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\| + 1\right)^2}\right\}.
\end{aligned}
$$

Then we finally achieve the following bound for an arbitrary $\epsilon' \in (0,1]$,

$$
\begin{aligned}
\tilde{\Gamma}_{B,m} \leq \frac{1}{M}\cdot\frac{8}{\hat{p}_{\min}}\cdot\left(\sum_{t=1}^m \gamma^{t-1}\cdot\sup_{s,a}\tilde{\mathbb{E}}\|\hat{R}(s,a)\| + \gamma^m\cdot\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\right) \\
+ \frac{C_{13}}{\sqrt{M}}\cdot\gamma^m\cdot\sum_{s,a}\sqrt{\hat{b}_\mu(s,a)}\cdot L\sqrt{p}\cdot\mathrm{diam}(\Theta;\|\cdot\|) + C_{14}\cdot(1+\gamma^m)\cdot\sqrt{\hat{p}_{\min}}\cdot\epsilon',
\end{aligned} \tag{153}
$$

with probability larger than

$$
\begin{aligned}
&1 - C_{15}\cdot(|\mathcal{S}\times\mathcal{A}| + M) \\
&\qquad \times \exp\left\{ \frac{-C_{16}\cdot \hat{p}_{\min}^2 \cdot M \cdot {\epsilon'}^2}{\left(\sum_{t=1}^m \gamma^{t-1}\cdot\sup_{s,a}\left\|\|\hat{R}^{(t)}(s,a)\|\right\|_{\tilde{\psi}_2} + \gamma^m \cdot \sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\| + 1\right)^2}\right\}.
\end{aligned} \tag{154}
$$

Now we will define the following event, whose probability bound is shown under $E_{1,a}\cap E_{1,b}$.

$$
E_2 := \left\{\omega\in\Omega_B^{(\mathcal{D})} : \text{(153) holds.}\right\}, \quad \mathbb{P}(E_2|E_{1,a}\cap E_{1,b}) \geq \text{(154)}.
$$

B.3.5 Expressing estimated quantities of (153) and (154) with population quantities

The bounds (153) and (154) are not yet useful though, since they are not fully represented with population quantities. This is because we are caring about Stage 3 probability space (140) conditioned upon the observed data $\mathcal{D}$ (that is associated with Stage 1 and 2 probability spaces). So we hope to bound the following terms with the corresponding population quantities,

$$
\sum_{s,a}\sqrt{\hat{b}_\mu(s,a)} \;\&\; \hat{p}_{\min} \;\&\; \sup_{s,a}\tilde{\mathbb{E}}\|\hat{R}(s,a)\| \;\&\; \sup_{s,a}\left\|\|\hat{R}(s,a)\|\right\|_{\tilde{\psi}_2}, \tag{155}
$$

but it comes with a price, that is subtraction of probability.

Let us first condition upon $E_{1,a}$ (141). Then the first term can be bounded readily as follows,

$$
\sum_{s,a}\sqrt{\hat{b}_\mu(s,a)} \leq \sum_{s,a}\sqrt{\frac{3}{2}b_\mu(s,a)} = \sqrt{\frac{3}{2}}\cdot\sum_{s,a}\sqrt{b_\mu(s,a)} \text{ by Facts (39) in B.3.2.}
$$

For the second term, we should obtain both lower bound and upper bound, since it appears in both denominator and numerator of Bound (153). Using Facts (39), we can bound $\hat{p}_{\min}$ (150) as follows, where $\hat{s},\hat{a} \overset{let}{=} \arg\min_{s,a\in\mathcal{S}\times\mathcal{A}}\hat{b}_\mu(s,a)$ and $s_*,a_* \overset{let}{=} \arg\min_{s,a\in\mathcal{S}\times\mathcal{A}} b_\mu(s,a)$,

$$
\begin{aligned}
\text{For } \forall s,a,\hat{b}_\mu(s,a) \geq \frac{1}{2}b_\mu(s,a) \geq \frac{1}{2}p_{\min} \quad \therefore\ \hat{p}_{\min} \geq \frac{1}{2}p_{\min}, \\
\hat{p}_{\min} = \hat{b}_\mu(\hat{s},\hat{a}) \leq \hat{b}_\mu(s_*,a_*) \leq \frac{3}{2}b_\mu(s_*,a_*) = \frac{3}{2}p_{\min}, \\
\therefore\ \frac{1}{2}p_{\min} \leq \hat{p}_{\min} \leq \frac{3}{2}p_{\min}.
\end{aligned}
$$

Unlike the first two terms of (155) that could be bounded as above solely $E_{1,a}$, the remaining two terms cannot be deterministically bounded, necessitiating the derivation of probabilistic bound.

Since we are conditioning on $E_{1,a}$ (141), we should deal with Stage 2 probability space (140), using the conditional probability $\mathbb{P}^{(\mathbf{N})}(\cdots)$ introduced below Definition (38). Based on Derivation (73) for an arbitrary $\epsilon_3' \in (0, 1]$,

$$\mathbb{P}^{(\mathbf{N})}\left[\sup_{s,a}\tilde{\mathbb{E}}\|\hat{R}(s,a)\| \le \sup_{s,a}\mathbb{E}\|R(s,a)\| + \sqrt{p_{\min}}\cdot\epsilon_3'\right] \\ \ge 1 - 2|\mathcal{S}\times\mathcal{A}|\cdot\exp\left\{\frac{-C_{17}\cdot p_{\min}^2\cdot N\cdot {\epsilon_3'}^2}{d^2\cdot\left(\sup_{s,a}\|R(s,a)\|_{\psi_2}\right)^2}\right\}, \tag{156}$$

Now let us bound the forth term of (155). First, let $s, a \in \mathcal{S}\times\mathcal{A}$ be arbitrary. Define two random variables $U(s,a) := \|R(s,a)\|$ and $\hat{U}(s,a) := \|\hat{R}(s,a)\|$, along with the following functions,

$$A_{s,a}(t) := \mathbb{E}\left\{\exp\left(\frac{U(s,a)^2}{t^2}\right)\right\},$$

$$\hat{A}_{s,a}(t) := \tilde{\mathbb{E}}\left\{\exp\left(\frac{\hat{U}(s,a)^2}{t^2}\right)\right\} = \frac{1}{N(s,a)}\sum_{i=1}^{N(s,a)}\exp\left(\frac{U_i(s,a)^2}{t^2}\right),$$

where $U_i(s,a) = \|R_i(s,a)\|$ $(1 \le i \le N(s,a))$ represents the samples. Let $t_0(s,a) > 0$ be the value such that

$$A_{s,a}\big(t_0(s,a)\big) = 1.$$

It is obvious to see $t_0(s,a) > \|\|R(s,a)\|\|_{\psi_2}$, based on that $A_{s,a}(\|\|R(s,a)\|\|_{\psi_2}) = 2$ holds and $A_{s,a}(t)$ is a strictly decreasing function. We can bound its probability term as follows,

$$\begin{aligned}\mathbb{P}^{(\mathbf{N})}\big(\big\|\|\hat{R}(s,a)\|\big\|_{\tilde{\psi}_2} \le t_0(s,a)\big) &= \mathbb{P}^{(\mathbf{N})}\big\{\hat{A}_{s,a}\big(t_0(s,a)\big) \le 2\big\} \quad \text{by Definition (33)} \\ &\ge \mathbb{P}^{(\mathbf{N})}\left\{\left|\hat{A}_{s,a}\big(t_0(s,a)\big) - A_{s,a}\big(t_0(s,a)\big)\right| \le 2 - A_{s,a}\big(t_0(s,a)\big)\right\} \\ &\ge 1 - \mathbb{E}^{(\mathbf{N})}\big|\hat{A}_{s,a}\big(t_0(s,a)\big) - A_{s,a}\big(t_0(s,a)\big)\big| \quad \text{by Markov's Inequality.} \end{aligned} \tag{157}$$

Note that we could apply Markov's Inequality in the last line since $\mathbb{E}|A_{s,a}(t)| < \infty$ by subgaussianity assumption 3 that implies $\|U(s,a)\| = \|\|R(s,a)\|\|_{\psi_2} < \infty$. Now let us shrink the expectation term (157) with the following lemma that is proved in C.2.8,

Lemma 5. *If X_i $(1 \le i \le n)$ are iid with $\mathbb{E}(X_1) = 0$, $\mathbb{E}|X_1| < \infty$, then the expectation of the sample mean shrinks to zero as follows,*

$$\mathbb{E}\big|\bar{X}_n\big| \le \inf_{z>0}\left[\frac{1}{\sqrt{n}}\cdot\left\{\mathbb{E}\big\{X_1^2\cdot\mathbf{1}(|X_1|\le z)\big\}\right\}^{1/2} + \mathbb{E}\big\{|X_1|\cdot\mathbf{1}(|X_1|>z)\big\}\right] \to 0 \quad \textit{as } n\to\infty.$$

Note that we have a deterministic sequence its convergence to zero is guaranteed, however its speed depends on the tail of the distribution X_1.

With the following new notation

$$V(s,a) := \exp\left(\frac{U(s,a)^2}{t_0(s,a)^2}\right) = \exp\left(\frac{\|R(s,a)\|^2}{t_0(s,a)^2}\right),$$

we can apply Lemma 5, we have

$$\begin{aligned}&\mathbb{E}^{(\mathbf{N})}\big|\hat{A}_{s,a}\big(t_0(s,a)\big) - A_{s,a}\big(t_0(s,a)\big)\big| \\ &\le \inf_{z>0}\left[\frac{1}{\sqrt{N(s,a)}}\cdot\left\{\mathbb{E}\big\{V(s,a)^2\cdot\mathbf{1}\big(V(s,a)\le z\big)\big\}\right\}^{1/2} + \mathbb{E}\big\{V(s,a)\cdot\mathbf{1}\big(V(s,a)>z\big)\big\}\right] \\ &\le \sqrt{\frac{2}{p_{\min}}}\cdot\inf_{z>0}\left[\frac{1}{\sqrt{N}}\cdot\left\{\mathbb{E}\big\{V(s,a)^2\cdot\mathbf{1}\big(V(s,a)\le z\big)\big\}\right\}^{1/2} + \mathbb{E}\big\{V(s,a)\cdot\mathbf{1}\big(V(s,a)>z\big)\big\}\right],\end{aligned} \tag{158}$$

where the last line holds by Facts (39). Now defining the following new variable

$$t_0^* := \sup_{s,a} t_0(s,a), \quad \therefore \; A_{s,a}(t_0^*) \leq 1 \;\; \text{for } \forall s,a \in \mathcal{S}\times\mathcal{A}, \tag{159}$$

we have the following,

$$\begin{aligned}
&\mathbb{P}^{(\mathbf{N})}\left\{ \sup_{s,a} \left\|\|\hat{R}(s,a)\|\right\|_{\tilde{\psi}_2} \leq t_0^* \right\} = \mathbb{P}^{(\mathbf{N})}\left\{ \left\|\|\hat{R}(s,a)\|\right\|_{\tilde{\psi}_2} \leq t_0^* \;\; \text{for } \forall s,a \in \mathcal{S}\times\mathcal{A} \right\} \\
&\geq \mathbb{P}^{(\mathbf{N})}\left\{ \left\|\|\hat{R}(s,a)\|\right\|_{\tilde{\psi}_2} \leq t_0(s,a) \;\; \text{for } \forall s,a \in \mathcal{S}\times\mathcal{A} \right\} \\
&\geq 1 - \sum_{s,a} \mathbb{E}^{(\mathbf{N})}\left| \hat{A}_{s,a}\big(t_0(s,a)\big) - A_{s,a}\big(t_0(s,a)\big) \right| \;\; \text{by Bound (157)} \\
&\geq 1 - \sum_{s,a} \sqrt{\frac{2}{p_{\min}}} \cdot \inf_{z>0} \left[\frac{1}{\sqrt{N}} \cdot \left\{ \mathbb{E}\big\{ V(s,a)^2 \cdot \mathbf{1}\big(V(s,a) \leq z\big) \big\} \right\}^{1/2} \right. \\
&\qquad\qquad \left. + \mathbb{E}\big\{ V(s,a) \cdot \mathbf{1}\big(V(s,a) > z\big) \big\} \right] \\
&\geq 1 - |\mathcal{S}\times\mathcal{A}| \cdot \frac{\sqrt{2}}{\sqrt{p_{\min}}} \times \inf_{z>0} \left[\frac{1}{\sqrt{N}} \cdot \sup_{s,a} \left\{ \mathbb{E}\big\{ V(s,a)^2 \cdot \mathbf{1}\big(V(s,a) \leq z\big) \big\} \right\}^{1/2} \right. \\
&\qquad\qquad \left. + \sup_{s,a} \mathbb{E}\big\{ V(s,a) \cdot \mathbf{1}\big(V(s,a) > z\big) \big\} \right] \\
&\geq 1 - |\mathcal{S}\times\mathcal{A}| \cdot \frac{\sqrt{2}}{\sqrt{p_{\min}}} \cdot \inf_{r>2} \left\{ N^{\frac{1}{r}-\frac{1}{2}} + \sup_{s,a} \mathbb{E}\big\{ V(s,a) \cdot \mathbf{1}\big(V(s,a) > N^{1/r}\big) \big\} \right\} \;\; \because \; z \overset{let}{=} N^{1/r} \\
&= 1 - \mathcal{D}(N),
\end{aligned} \tag{160}$$

where the third inequality holds by Inequality (158), and $\mathcal{D}(N)$ is defined as follows,

$$\begin{aligned}
\mathcal{D}(N) &:= |\mathcal{S}\times\mathcal{A}| \cdot \frac{\sqrt{2}}{\sqrt{p_{\min}}} \cdot \inf_{r>2} \left\{ N^{\frac{1}{r}-\frac{1}{2}} + \sup_{s,a} \mathbb{E}\big\{ V(s,a) \cdot \mathbf{1}\big(V(s,a) > N^{1/r}\big) \big\} \right\}, \\
&\therefore \; \mathcal{D}(N) \to 0 \;\; \text{as } N \to \infty, \quad \text{since } \sup_{s,a} \mathbb{E}\big\{ V(s,a) \big\} = \sup_{s,a} A_{s,a}\big(t_0(s,a)\big) = 1 < \infty.
\end{aligned} \tag{161}$$

By letting $\epsilon_3' = \epsilon' \in (0,1]$, we can bound all four estimated quantities (155) at the same time as follows,

$$\begin{aligned}
&\sum_{s,a} \sqrt{\hat{b}_\mu(s,a)} \leq \sqrt{\frac{3}{2}} \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \quad \& \quad \frac{1}{2} p_{\min} \leq \hat{p}_{\min} \leq \frac{3}{2} p_{\min}, \\
&\sup_{s,a} \tilde{\mathbb{E}}\|\hat{R}(s,a)\| \leq \sup_{s,a} \mathbb{E}\|R(s,a)\| + \epsilon' \quad \& \quad \sup_{s,a} \left\|\|\hat{R}(s,a)\|\right\|_{\tilde{\psi}_2} \leq t_0^*,
\end{aligned} \tag{162}$$

with probability larger than

$$1 - 2|\mathcal{S}\times\mathcal{A}| \cdot \exp\left\{ \frac{-C_{17} \cdot p_{\min}^2 \cdot N \cdot {\epsilon'}^2}{d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2} \right)^2} \right\} - \mathcal{D}(N). \tag{163}$$

Let us define a new event

$$E_3 := \left\{ \omega \in \Omega^{(\mathbf{N})} : \text{(162) holds.} \right\}, \quad \mathbb{P}(E_3 | E_{1,a}) \geq \text{(163)}. \tag{164}$$

Now that we have bounded the estimated quantities with its population counterparts (162), we can rewrite the bounds (153) and (154) as follows,

$$\begin{aligned}
\tilde{\Gamma}_{B,m} \leq \frac{1}{M} \cdot \frac{16}{p_{\min}} \cdot \left(\sum_{t=1}^{m} \gamma^{t-1} \cdot \sup_{s,a} (\mathbb{E}\|R(s,a)\| + 1) + \gamma^m \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right) \\
+ \frac{C_{18}}{\sqrt{M}} \cdot \gamma^m \cdot \sum_{s,a} \sqrt{b_\mu(s,a)} \cdot L\sqrt{p} \cdot \operatorname{diam}(\Theta; \|\cdot\|) + C_{19} \cdot (1+\gamma^m) \cdot \sqrt{p_{\min}} \cdot \epsilon',
\end{aligned} \tag{165}$$

with probability larger than

$$1 - C_{20} \cdot (|\mathcal{S} \times \mathcal{A}| + M) \cdot \exp\left\{ \frac{-C_{21} \cdot p_{\min}^2 \cdot M \cdot \epsilon'^2}{\left(\sum_{t=1}^{m} \gamma^{t-1} \cdot t_0^* + \gamma^m \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + 1 \right)^2} \right\}. \quad (166)$$

This means that

$$E_2' := \left\{ \omega \in \Omega_B^{(\mathcal{D})} : (165) \text{ holds.} \right\}, \quad \mathbb{P}(E_2' | E_{1,a} \cap E_{1,b} \cap E_3) \geq (166). \quad (167)$$

Putting together (145), (167), (164), we obtain the following for arbitrary $\epsilon, \epsilon' \in (0, 1]$,

$$\begin{aligned}
&\mathbb{P}(E_{1,a} \cap E_{1,b} \cap E_2' \cap E_3) = \mathbb{P}(E_2' | E_{1,a} \cap E_{1,b} \cap E_3) \cdot \mathbb{P}(E_{1,b} \cap E_3 | E_{1,a}) \cdot \mathbb{P}(E_{1,a}) \\
&= \left\{1 - \mathbb{P}(E_2'^c | E_{1,a} \cap E_{1,b} \cap E_3)\right\} \cdot \left\{1 - \mathbb{P}(E_{1,b}^c | E_{1,a}) - \mathbb{P}(E_3^c | E_{1,a})\right\} \cdot \left\{1 - \mathbb{P}(E_{1,a}^c)\right\} \\
&\geq 1 - \mathbb{P}(E_2'^c | E_{1,a} \cap E_{1,b} \cap E_3) - \mathbb{P}(E_{1,b}^c | E_{1,a}) - \mathbb{P}(E_3^c | E_{1,a}) - \mathbb{P}(E_{1,a}^c) \\
&\geq 1 - \mathcal{D}(N) - C_2 \cdot m \cdot (|\mathcal{S} \times \mathcal{A}| + N) \cdot \exp(-C_3 \cdot p_{\min}^2 \cdot N \cdot \epsilon^2 / C_{\text{den}}(m)) \\
&\quad - 2|\mathcal{S} \times \mathcal{A}| \cdot \exp\left\{ \frac{-C_{17} \cdot p_{\min}^2 \cdot N \cdot 1^2}{d^2 \cdot \left(\sup_{s,a} \|R(s,a)\|_{\psi_2} \right)^2} \right\} - \mathcal{D}(N) \quad (\because \text{ We let } \epsilon_3' = 1.) \\
&\quad - C_{20} \cdot (|\mathcal{S} \times \mathcal{A}| + M) \cdot \exp\left\{ \frac{-C_{21} \cdot p_{\min}^2 \cdot M \cdot \epsilon'^2}{\left(\sum_{t=1}^{m} \gamma^{t-1} \cdot t_0^* + \gamma^m \cdot \sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| + 1 \right)^2} \right\} \\
&\geq 1 - \mathcal{D}(N) - C_{22} \cdot m \cdot (|\mathcal{S} \times \mathcal{A}| + N) \cdot \exp(-C_{23} \cdot p_{\min}^2 \cdot N \cdot \epsilon^2 / C_{\text{den}}(m)) \\
&\quad - C_{22} \cdot (|\mathcal{S} \times \mathcal{A}| + M) \cdot \exp(-C_{23} \cdot p_{\min}^2 \cdot M \cdot \epsilon'^2 / B_{\text{den}}(m)). \quad (168)
\end{aligned}$$

where

$$B_{\text{den}}(m) := \left(\sum_{t=0}^{m-1} \gamma^t \right)^2 \cdot t_0^{*2} + \gamma^{2m} \cdot \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)^2 + 1. \quad (169)$$

B.3.6 SUMMARIZATION

Now all three terms of Inequality (147) are bounded in Inequalities (148), (149), (165), with probability larger than (168). Since the square-root of bootstrap discrepancy can be bounded as

$$\begin{aligned}
\tilde{\Gamma}_{B,m}^{1/2} \leq \frac{1}{\sqrt{M}} \cdot \frac{4}{\sqrt{p_{\min}}} \cdot \Bigg\{ &\left(\sum_{t=0}^{m-1} \gamma^t \cdot \sup_{s,a} \mathbb{E}\|R(s,a)\| \right)^{\frac{1}{2}} + \gamma^{m/2} \cdot \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)^{\frac{1}{2}} \\
&+ 1 \Bigg\} + \frac{C_{24}}{M^{1/4}} \cdot \gamma^{m/2} \cdot \left(\sum_{s,a} \sqrt{b_\mu(s,a)} \right)^{1/2} \cdot \sqrt{L} \cdot p^{1/4} \cdot \sqrt{\text{diam}(\Theta; \|\cdot\|)} \\
&+ C_{25} \cdot \sqrt{1 + \gamma^m} \cdot p_{\min}^{1/4} \cdot \sqrt{\epsilon'}, \quad (170)
\end{aligned}$$

Skipping all the messy calculations, we can obtain the following bound using $M, N \geq 1$,

$$\begin{aligned}
&\tilde{\Gamma}_{B,m}^{1/2} + \sup_{\theta \in \Theta} \bar{\mathcal{E}}_\theta^{1/2} + \Gamma_{N,m}^{1/2} \\
&\leq C_{26} \cdot \frac{m}{p_{\min}^{1/4}} \cdot \left\{ 1 + \gamma^{m/2} \cdot \sqrt{L} \cdot p^{1/4} \cdot \sqrt{\text{diam}(\Theta; \|\cdot\|)} \right\} \cdot (\sqrt{\epsilon} + \sqrt{\epsilon'}) \\
&\quad + C_{27} \cdot \frac{m}{p_{\min}} \cdot \Bigg\{ \sqrt{L} \cdot p^{1/4} \cdot \sqrt{\text{diam}(\Theta; \|\cdot\|)} \\
&\qquad + \left(\sum_{t=0}^{m-1} \gamma^t \right)^{1/2} \cdot \left(\sup_{s,a} \mathbb{E}\|R(s,a)\| \right)^{1/2} + \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)^{1/2} + 1 \Bigg\} \\
&\leq C_{28} \cdot \frac{m}{p_{\min}} \cdot \Bigg\{ \sqrt{L} \cdot p^{1/4} \cdot \sqrt{\text{diam}(\Theta; \|\cdot\|)} \quad (\because \epsilon, \epsilon' \leq 1) \\
&\qquad + \left(\sum_{t=0}^{m-1} \gamma^t \right)^{1/2} \cdot \left(\sup_{s,a} \mathbb{E}\|R(s,a)\| \right)^{1/2} + \left(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\| \right)^{1/2} + 1 \Bigg\}
\end{aligned}$$

Then we can go back to Inequality (147), and obtain the following bound using Inequalities (170) and (82),

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-\hat{F}_m(\theta)\right| \\
&\leq C_{29}\cdot\frac{m}{p_{\min}}\cdot\Bigg\{\sqrt{L}\cdot p^{1/4}\cdot\sqrt{\operatorname{diam}(\Theta;\|\cdot\|)} \\
&\qquad\qquad +\Big(\sum_{t=0}^{m-1}\gamma^t\Big)^{1/2}\cdot\Big(\sup_{s,a}\mathbb{E}\|R(s,a)\|\Big)^{1/2}+\Big(\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\Big)^{1/2}+1\Bigg\} \\
&\times\Bigg[\frac{1}{\sqrt{M}}\cdot\frac{1}{\sqrt{p_{\min}}}\cdot\Bigg\{\Bigg(\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|\Bigg)^{1/2}+\gamma^{m/2}\cdot\Bigg(\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|\Bigg)^{1/2}+1\Bigg\} \\
&+\frac{1}{M^{1/4}}\cdot\gamma^{m/2}\cdot\Bigg(\sum_{s,a}\sqrt{b_\mu(s,a)}\Bigg)^{1/2}\cdot\sqrt{L}\cdot p^{1/4}\cdot\sqrt{\operatorname{diam}(\Theta;\|\cdot\|)}+\sqrt{1+\gamma^m}\cdot p_{\min}^{1/4}\cdot\sqrt{\epsilon'}\Bigg] \\
&\leq\frac{C_{31}}{p_{\min}^{3/2}}\cdot m\cdot B_{\text{env}}^{(m)}(\Theta)\cdot\left(\frac{1}{M^{1/4}}+\sqrt{\epsilon'}\right)
\end{aligned}
$$

where

$$
B_{\text{env}}^{(m)}(\Theta):=L\sqrt{p}\cdot\operatorname{diam}(\Theta;\|\cdot\|)+\sum_{t=0}^{m-1}\gamma^t\cdot\sup_{s,a}\mathbb{E}\|R(s,a)\|+\sup_{s,a}\mathbb{E}\|Z(s,a;\theta_0)\|+1 \quad (171)
$$

Now the final task is to incorporate this with the bound of Lemma 2, based on Decomposition (139). As we mentioned in B.3.2, we already inherited the bound, so we do not have to subtract any additional probability from the current bound (168). Then we have

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-F_m(\theta)\right|\leq\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-\hat{F}_m(\theta)\right|+\sup_{\theta\in\Theta}\left|\hat{F}_m(\theta)-F_m(\theta)\right| \\
&\leq C_1\cdot\frac{1}{p_{\min}^2}\cdot m^2\cdot C_{\text{env}}^{(m)}(\Theta)\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right)+\frac{C_{31}}{p_{\min}^{3/2}}\cdot m\cdot B_{\text{env}}^{(m)}(\Theta)\cdot\left(\frac{1}{M^{1/4}}+\sqrt{\epsilon'}\right) \\
&\leq\frac{C_{32}}{p_{\min}^2}\cdot B_{\text{env}}^{(m)}(\Theta)\cdot\left\{m^2\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right)+m\cdot\left(\frac{1}{M^{1/4}}+\sqrt{\epsilon'}\right)\right\}\ \text{by (138) and (171),}
\end{aligned}
$$

with probability larger than

$$
\begin{aligned}
&1-\mathcal{D}(N)-C_{22}\cdot m\cdot(|\mathcal{S}\times\mathcal{A}|+N)\cdot\exp(-C_{23}\cdot p_{\min}^2\cdot N\cdot\epsilon^2/C_{\text{den}}(m)) \\
&\quad-C_{22}\cdot(|\mathcal{S}\times\mathcal{A}|+M)\cdot\exp(-C_{23}\cdot p_{\min}^2\cdot M\cdot\epsilon'^2/B_{\text{den}}(m)).
\end{aligned}
$$

B.4 PROOF OF THEOREM 3

B.4.1 INACCURACY OF PARAMETER ESTIMATION

Our idea is that larger N, M, m will lead to tighter (probabilistic) bound of $\sup_{\theta\in\Theta}|\hat{F}_m^{(B)}(\theta)-F(\theta)|$, which can be decomposed as follows,

$$\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-F(\theta)\right| \le \sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-F_m(\theta)\right| + \sup_{\theta\in\Theta}\left|F_m(\theta)-F(\theta)\right|.$$

Note that the first term of RHS is the probabilistic term that we bounded in Lemma 4, and the second term is a deterministic term that can be bounded based on following (proof in C.2.9)

$$\sup_{\theta\in\Theta}\left|F_m(\theta)-F(\theta)\right| \le 4\gamma^m\cdot C_{\text{bias}} \quad \text{where} \quad C_{\text{bias}} := \tilde{\eta}(\tilde{\theta},\pi) + L\cdot\text{diam}(\Theta;\|\cdot\|), \tag{172}$$

$$\text{with} \quad \tilde{\eta}(\theta,\pi) := \tilde{\eta}\big\{\Upsilon_\theta,\Upsilon_\pi\big\}. \tag{173}$$

Then combined with Lemma 4, we have

$$\begin{aligned}\mathbb{P}\bigg\{\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-F(\theta)\right| \le A(m,N,M,\epsilon,\epsilon') + 4\gamma^m\cdot C_{\text{bias}}\bigg\} \qquad (174)\\ \ge 1-\mathcal{D}(N) - C_1\cdot m\cdot(|\mathcal{S}\times\mathcal{A}|+N)\cdot\exp(-C_2\cdot p_{\min}^2\cdot N\cdot\epsilon^2/C_{\text{den}}(m))\\ - C_1\cdot(|\mathcal{S}\times\mathcal{A}|+M)\cdot\exp(-C_2\cdot p_{\min}^2\cdot M\cdot\epsilon'^2/B_{\text{den}}(m)),\end{aligned}$$

where

$$A(m,N,M,\epsilon,\epsilon') := \frac{C_3}{p_{\min}^2}\cdot B_{\text{env}}^{(m)}(\Theta)\times\left\{m^2\cdot\left(\frac{1}{N^{1/4}}+\sqrt{\epsilon}\right)+m\cdot\left(\frac{1}{M^{1/4}}+\sqrt{\epsilon'}\right)\right\}. \tag{175}$$

Now in order to relate the bound (174) to the estimation inaccuracy of $\hat{\theta}_m^{(B)}$ (23), we shall use the function $\psi(\cdot)$ introduced by Example 1.3 of Sen (2018),

$$\psi(\delta) := \inf_{\theta\in\Theta:\|\theta-\tilde{\theta}\|\ge\delta} F(\theta)-F(\tilde{\theta}) \quad \text{and} \quad \psi^{-1}(y) := \inf\left\{\delta>0:\psi(\delta)\ge y\right\}. \tag{176}$$

Depending on whether Θ includes any element in the outermost boundary, $\psi(\cdot)$ can be defined either for $0\le\delta<\sup_{\theta\in\Theta}\|\theta-\tilde{\theta}\|$ or $0\le\delta\le\sup_{\theta\in\Theta}\|\theta-\tilde{\theta}\|$. We will extend the function in the following trivial way of extending horizontally from the rightmost point,

$$\psi(\delta) := \begin{cases}\psi(\sup_{\theta\in\Theta}\|\theta-\tilde{\theta}\|) & \text{if } \psi(\sup_{\theta\in\Theta}\|\theta-\tilde{\theta}\|) \text{ is defined,}\\ \sup_{0\le\delta'<\sup_{\theta\in\Theta}\|\theta-\tilde{\theta}\|}\psi(\delta') & \text{otherwise.}\end{cases}$$

There are several important properties of $\psi(\cdot)$ that are proved in C.2.10 based on Assumption 6.

Remark 3. *The function ψ (176) satisfies the following properties.*

1. *$\psi^{-1}(\cdot)$ is an increasing function such that $\psi^{-1}(y)\to 0$ as $y\to 0$ and $\psi^{-1}(y)=\infty$ for $y>\sup_{\delta>0}\psi(\delta)$.*
2. *$\lim_{\epsilon\to 0+}\psi\{\psi^{-1}(y)+\epsilon\}\ge y$ holds for all $y\in[0,\psi(\sup_{\theta\in\Theta}\|\theta-\tilde{\theta}\|)]$.*
3. *Let $\hat{F}:\Theta\subset\mathbb{R}^p\to\mathbb{R}$ be an arbitrary estimate of F (21) that have minimizer(s) within Θ. For an arbitrary value $\delta>0$, if there exists a minimizer $\exists\hat{\theta}\in\arg\min_{\theta\in\Theta}\hat{F}(\theta)$ such that $\|\hat{\theta}-\tilde{\theta}\|>\delta$, then $\sup_{\theta\in\Theta}|\hat{F}(\theta)-F(\theta)|\ge\frac{1}{2}\lim_{\delta'\to\delta+}\psi(\delta')$ holds.*

Based on the Remark 3 (3rd statement), we have the following bound for $\psi_+(\delta):=\lim_{\delta'\to\delta+}\psi(\delta')$,

$$\mathbb{P}\bigg\{\exists\hat{\theta}_m^{(B)}\in\Theta \text{ such that } \|\hat{\theta}_m^{(B)}-\tilde{\theta}\|\ge\delta\bigg\} \le \mathbb{P}\bigg\{\sup_{\theta\in\Theta}\left|\hat{F}_m^{(B)}(\theta)-F(\theta)\right|\ge\frac{1}{2}\psi_+(\delta)\bigg\}.$$

Now letting $\delta = \psi^{-1}\{2 \cdot A(m, N, M, \epsilon, \epsilon') + 8\gamma^m \cdot C_{\text{bias}}\}$, we have

$$\mathbb{P}\left\{\exists \hat{\theta}_m^{(B)} \in \Theta \text{ such that } \|\hat{\theta}_m^{(B)} - \tilde{\theta}\| \geq \psi^{-1}\left(2 \cdot A(m, N, M, \epsilon, \epsilon') + 8\gamma^m \cdot C_{\text{bias}}\right)\right\}$$
$$\leq \mathbb{P}\left\{\sup_{\theta \in \Theta}\left|\hat{F}_m^{(B)}(\theta) - F(\theta)\right| \geq A(m, N, M, \epsilon, \epsilon') + 4\gamma^m \cdot C_{\text{bias}}\right\} \quad \text{by Remark 3 (2nd Statement)}$$
$$\leq \mathcal{D}(N) + C_1 \cdot m \cdot (|\mathcal{S} \times \mathcal{A}| + N) \cdot \exp(-C_2 \cdot p_{\min}^2 \cdot N \cdot \epsilon^2 / C_{\text{den}}(m))$$
$$+ C_1 \cdot (|\mathcal{S} \times \mathcal{A}| + M) \cdot \exp(-C_2 \cdot p_{\min}^2 \cdot M \cdot \epsilon'^2 / B_{\text{den}}(m)) \qquad \text{by (174).} \tag{177}$$

B.4.2 Simplifying the Probability term

Let $\delta_1, \delta_2 \in (0, 1)$ be arbitrary. Now let us simplify the result (177) by letting

$$\epsilon = \sqrt{\frac{C_{\text{den}}(m)}{C_2 \cdot p_{\min}^2}} \cdot \sqrt{\frac{1}{N} \cdot \log\left(\frac{C_1 \cdot m \cdot (|\mathcal{S} \times \mathcal{A}| + N)}{\delta_1}\right)}, \tag{178}$$
$$\epsilon' = \sqrt{\frac{B_{\text{den}}(m)}{C_2 \cdot p_{\min}^2}} \cdot \sqrt{\frac{1}{M} \cdot \log\left(\frac{C_1 \cdot (|\mathcal{S} \times \mathcal{A}| + M)}{\delta_2}\right)},$$

Recall that we have to ensure $\epsilon, \epsilon' \in (0, 1]$. Furthermore, since $A(m, N, M, \epsilon, \epsilon')$ in (175) contains $(N^{1/4} + \sqrt{\epsilon})$ and $(M^{1/4} + \sqrt{\epsilon'})$. Since ϵ and ϵ' (178) decays in a slower rate than $N^{1/4}$ and $M^{1/4}$, we can ignore them when the sample size is sufficiently large. To ensure these, we will assume $N \geq N_m(\delta_1),\ M \geq M_m(\delta_2)$ where $N_m(\delta_1)$ and $M_m(\delta_2)$ are the smallest integers such that $N \geq N_m(\delta_1),\ M \geq M_m(\delta_2)$ implies the following, with $C_{\text{den}}(m)$ (132) and $B_{\text{den}}(m)$ (169),

$$\frac{1}{N} \leq \frac{C_{\text{den}}(m)}{C_2 \cdot p_{\min}^2} \cdot \frac{1}{N} \cdot \log\left(\frac{C_1 \cdot m \cdot (|\mathcal{S} \times \mathcal{A}| + N)}{\delta_1}\right) \leq 1, \tag{179}$$
$$\frac{1}{M} \leq \frac{B_{\text{den}}(m)}{C_2 \cdot p_{\min}^2} \cdot \frac{1}{M} \cdot \log\left(\frac{C_1 \cdot (|\mathcal{S} \times \mathcal{A}| + M)}{\delta_2}\right) \leq 1.$$

Then let us bound the term $A(m, N, M, \epsilon, \epsilon')$ of Equation (175) as follows, using the values of ϵ, ϵ' specified in Equation (178) and assuming (179). Skipping the calculation details, we can derive

$$A(m, N, M, \epsilon, \epsilon') \leq C_{\text{model}} \times$$
$$\left[m^2 \cdot \left\{\frac{1}{N} \cdot \log\left(\frac{C_1 \cdot m \cdot (|\mathcal{S} \times \mathcal{A}| + N)}{\delta_1}\right)\right\}^{\frac{1}{4}} + m \cdot \left\{\frac{1}{M} \cdot \log\left(\frac{C_1 \cdot (|\mathcal{S} \times \mathcal{A}| + M)}{\delta_2}\right)\right\}^{\frac{1}{4}}\right]$$
$$\stackrel{let}{=} A_1(m, N, M, \delta_1, \delta_2). \tag{180}$$

where

$$C_{\text{model}} := \frac{C_4}{p_{\min}^{5/2}} \cdot \left\{L^2 p \cdot \text{diam}(\Theta; \|\cdot\|)^2 + \left(\frac{1}{1-\gamma}\right)^2 \cdot \max\left\{d \cdot \sup_{s,a} \|R(s, a)\|_{\psi_2}, t_0^*\right\}^2\right.$$
$$\left. + \left(\sup_{s,a} \mathbb{E}\|Z(s, a; \theta_0)\|\right)^2 + 1\right\}^{3/4}, \tag{181}$$

Now we can rewrite Bound (177) as follows by Remark 3 (1st Statement),

$$\mathbb{P}\left\{\forall \hat{\theta}_m^{(B)} \in \arg\min_{\theta \in \Theta} \hat{F}_m^{(B)}(\theta),\ \ \|\hat{\theta}_m^{(B)} - \tilde{\theta}\| \leq \psi^{-1}\left(2 \cdot A_1(m, N, M, \delta_1, \delta_2) + 8\gamma^m \cdot C_{\text{bias}}\right)\right\}$$
$$\geq 1 - \mathcal{D}(N) - \delta_1 - \delta_2.$$

Next, we can analyze the convergence rate of our estimated distribution towards the best approximation $\Upsilon_{\tilde{\theta}}$ in Energy Distance, based on following relationship with Euclidean distance,

$$\mathcal{E}\left\{\Upsilon_{\theta_1}(s, a), \Upsilon_{\theta_2}(s, a)\right\} = \left\{\mathbb{E}\|Z_\alpha(s, a; \theta_1) - Z_\beta(s, a; \theta_2)\| - \mathbb{E}\|Z_\alpha(s, a; \theta_1) - Z_\beta(s, a; \theta_1)\|\right\}$$
$$+ \left\{\mathbb{E}\|Z_\alpha(s, a; \theta_1) - Z_\beta(s, a; \theta_2)\| - \mathbb{E}\|Z_\alpha(s, a; \theta_2) - Z_\beta(s, a; \theta_2)\|\right\}$$
$$\leq \tilde{\eta}(\theta_1, \theta_1) + \tilde{\eta}(\theta_1, \theta_2) + \tilde{\eta}(\theta_1, \theta_2) + \tilde{\eta}(\theta_2, \theta_2) \quad \text{by Trick (49)}$$
$$\leq 2\tilde{\eta}(\theta_1, \theta_2) \leq 2L \cdot \|\theta_1 - \theta_2\|, \quad \text{by Assumption 7}$$

which leads to

$$\bar{\mathcal{E}}\big(\Upsilon_{\hat{\theta}_m^{(B)}}, \Upsilon_{\tilde{\theta}}\big) = \sum_{s,a} b_\mu(s,a) \cdot \mathcal{E}\Big\{ \Upsilon_{\hat{\theta}_m^{(B)}}(s,a), \Upsilon_{\tilde{\theta}}(s,a) \Big\} \leq 2L \cdot \|\hat{\theta}_m^{(B)} - \tilde{\theta}\|,$$

B.4.3 Finite Sample Error Bound

Under Assumptions 1, 3–8, for a fixed step level $m \in \mathbb{N}$ and arbitrary $\delta_1, \delta_2 \in (0,1)$, given that $N \geq \max\{N_m(\delta_1), 2\},\ M \geq \max\{M_m(\delta_2), 2\}$ defined in (179), we have the following bound with probability larger than $1 - \mathcal{D}(N) - \delta_1 - \delta_2$, all values of $\hat{\theta}_m^{(B)} \in \arg\min_{\theta\in\Theta} \hat{F}_m^{(B)}(\theta)$ satisfy

$$\|\hat{\theta}_m^{(B)} - \tilde{\theta}\| \leq \psi^{-1}\Bigg(2C_{\text{model}} \cdot \Bigg[m^2 \cdot \Bigg\{\frac{1}{N} \cdot \log\Bigg(\frac{C_1 \cdot m \cdot (|\mathcal{S}\times\mathcal{A}| + N)}{\delta_1}\Bigg)\Bigg\}^{\frac{1}{4}} + m \cdot \Bigg\{\frac{1}{M} \cdot \log\Bigg(\frac{C_1 \cdot (|\mathcal{S}\times\mathcal{A}| + M)}{\delta_2}\Bigg)\Bigg\}^{\frac{1}{4}}\Bigg] + 4\gamma^m \cdot C_{\text{bias}}\Bigg), \tag{182}$$

along with

$$\bar{\mathcal{E}}\big(\Upsilon_{\hat{\theta}_m^{(B)}}, \Upsilon_{\tilde{\theta}}\big) \leq 2L \cdot \|\hat{\theta}_m^{(B)} - \tilde{\theta}\|, \tag{183}$$

where $\mathcal{D}(N) \to 0$ as $N \to \infty$ (184), and ψ^{-1} (176) is an increasing function that ensures $\psi^{-1}(y) \to 0$ as $y \to 0$, as stated in Remark 3 (1st statement).

Here is the recap of definitions of the terms that we used in Equations (161), (172), (181), (159),

$$\mathcal{D}(N) := |\mathcal{S}\times\mathcal{A}| \cdot \frac{\sqrt{2}}{\sqrt{p_{\min}}} \cdot \inf_{r>2}\Bigg\{ N^{\frac{1}{r}-\frac{1}{2}} + \sup_{s,a} \mathbb{E}\big\{V(s,a)\cdot \mathbf{1}\big(V(s,a) > N^{1/r}\big)\big\}\Bigg\}$$
$$\to 0 \ \text{ as } N \to \infty \ \text{ where } V(s,a) := \exp\Bigg(\frac{\|R(s,a)\|^2}{t_0(s,a)^2}\Bigg),$$

$$C_{\text{bias}} := \tilde{\eta}(\tilde{\theta}, \pi) + L \cdot \text{diam}(\Theta; \|\cdot\|),$$

$$C_{\text{model}} := \frac{C_4}{p_{\min}^{5/2}} \cdot \Bigg\{ L^2 p \cdot \text{diam}(\Theta;\|\cdot\|)^2 + \Bigg(\frac{1}{1-\gamma}\Bigg)^2 \cdot \max\big\{d \cdot \sup_{s,a} \|R(s,a)\|_{\psi_2}, t_0^*\big\}^2 + \big(\sup_{s,a} \mathbb{E}\|Z(s,a;\theta_0)\|\big)^2 + 1\Bigg\}^{3/4},$$

$$t_0^* := \sup_{s,a} t_0(s,a) \ \text{ where } t_0(s,a) > 0 \text{ are values such that } \ \mathbb{E}\Bigg\{\exp\Bigg(\frac{\|R(s,a)\|^2}{t_0(s,a)^2}\Bigg)\Bigg\} = 1. \tag{184}$$

Just for a brief note, minimizer(s) of the estimated objective function $\hat{\theta}_m^{(B)} \in \arg\min_{\theta\in\Theta} \hat{F}_m^{(B)}(\theta)$ always exists due to continuity of $\hat{F}_m^{(B)}$ (proof in C.2.11).

B.4.4 Asymptotic Setting

Based on the finite-sample error bound provided in B.4.3, we will now assume N, M, m are large enough to satisfy the assumptions $N \geq \max\{N_m(\delta_1), 2\},\ M \geq \max\{M_m(\delta_2), 2\}$. We should also assume the following holds as $N, M, m \to \infty$ (terms defined in (180) and (172)),

$$2 \cdot A_1(m, N, M, \delta_1, \delta_2) + 8\gamma^m \cdot C_{\text{bias}} \to 0 \tag{185}$$

where the LHS is exactly the term inside $\psi^{-1}(\cdot)$ in (182). This condition is necessary to ensure that the RHS of Bound (182) to have a finite value by Remark 3 (1st statement). It should be verified that these conditions hold in the asymptotic sense, which we will discuss in the following section B.4.5 where we choose the actual growing speed of M and m with respect to N.

By Assumption 6, we could derive the following within the proof of Remark 3 in C.2.10 (1st statement),

$$\psi^{-1}(y) \leq \frac{1}{{c_q}^{1/q}} \cdot y^{1/q} \quad \text{for } \forall y \in [0, \sup_{\delta>0} \psi(\delta)) \quad \text{by (197).}$$

Based on this, by letting $\delta_1 = \delta_2 = \delta/2$ in (182), we have following with probability larger than $1 - \mathcal{D}(N) - \delta$,

$$\|\hat{\theta}_m^{(B)} - \tilde{\theta}\| \lesssim \Bigg[\underbrace{m^2 \cdot \left\{ \frac{1}{N} \cdot \log\left(\frac{2mN}{\delta} \right) \right\}^{\frac{1}{4}}}_{\text{data}} + \underbrace{m \cdot \left\{ \frac{1}{M} \log\left(\frac{2M}{\delta} \right) \right\}^{\frac{1}{4}}}_{\text{bootstrap}} + \underbrace{\gamma^m}_{\text{bias}} \Bigg]^{\frac{1}{q}}, \tag{186}$$

where $\lesssim$ means bounded by the given bound (RHS) multiplied by a positive number that does not depend on N, M, m. Each of the three terms (186) corresponds to the inaccuracy associated with the observed data $\{(s_i, a_i, r_i, s_i')\}_{i=1}^N$, the resampled trajectories (22), and the extent of non-realizability. Each can be reduced by increasing N, M, and m, however larger m makes the first two terms more challenging for to shrink, so it resembles bias-variance trade-off.

Accordingly to (186), the conditions of N, M, m (185) can be rewritten as following,

$$N \geq \max\{N_m(\delta/2), 2\}, \quad M \geq \max\{M_m(\delta/2), 2\}, \tag{187}$$

based on Definitions (179), along with the following based on Definition (180),

$$2C_{\text{model}} \cdot \Bigg[m^2 \cdot \left\{ \frac{1}{N} \cdot \log\left(\frac{2C_1 \cdot m \cdot (|\mathcal{S} \times \mathcal{A}| + N)}{\delta} \right) \right\}^{\frac{1}{4}} + m \cdot \left\{ \frac{1}{M} \cdot \log\left(\frac{2C_1 \cdot (|\mathcal{S} \times \mathcal{A}| + M)}{\delta} \right) \right\}^{\frac{1}{4}} \Bigg] + 8\gamma^m \cdot C_{\text{bias}} \to 0. \tag{188}$$

B.4.5 Optimal Convergence Rate

Now we will assume an asymptotic case where the sample size grows to infinity $N \to \infty$, and M, m grow accordingly with a chosen rate. Towards that end, we have to achieve two goals. First, we have to ensure that (187) and (188) are satisfied as $N \to \infty$. Second, we have to make (186) shrink in the fastest possible rate. As long as M (which we can let to be arbitrarily large) grows in the same rate with (or faster than) N,

$$M = \lfloor C_5 \cdot N \rfloor \quad \text{for arbitrary} \quad C_5 > 0, \tag{189}$$

the second term of (186) becomes ignorable in the asymptotic sense.

As noted below (186), increasing m has trade-off effect, so we shall derive an appropriate speed of m, and then verify that it can satisfy (187) with large enough N. Using (189), we can take up (186) as follows for sufficiently large N,

$$\|\hat{\theta}_m^{(B)} - \tilde{\theta}\| \lesssim \left[m^2 \cdot \left\{ \frac{1}{N} \cdot \log\left(\frac{mN}{\delta} \right) \right\}^{\frac{1}{4}} + \gamma^m \right]^{\frac{1}{q}}.$$

Now we choose the optimal level of m that makes the two terms converge in the same rate,

$$\gamma^m \approx m^2 \cdot \frac{1}{N^{1/4}} \cdot \big(\log(mN) \big)^{1/4}.$$

However, this relationship is very intricate, so we could not calculat m that makes both sides perfectly match. So we alternatively solved an easier equation that gives us the following relationship,

$$\gamma^m \approx C_6 \cdot \left(\frac{\log N}{N} \right)^{1/4}, \quad \therefore\ m \overset{let}{=} \left\lfloor \frac{1}{4} \cdot \log_{\frac{1}{\gamma}} \left(\frac{C_7 \cdot N}{\log N} \right) \right\rfloor. \tag{190}$$

Note that the values of $C_6, C_7 > 0$ can be arbitrary, as long as $C_7 = C_6^{-4}$ holds. Skipping the calculation details, it can be ascertained that the orders (189) and (190) ensures (187) and (188) to hold as $N \to \infty$. This can be verified based on the fact that $\sup_{m \in \mathbb{N}} C_{\text{den}}(m) < \infty$ and $\sup_{m \in \mathbb{N}} B_{\text{den}}(m) < \infty$, which are defined in (132) and (169).

Furthermore, it allows us to achieve

$$\left\{ \frac{m^8}{N} \cdot \log\left(\frac{mN}{\delta} \right) \right\}^{\frac{1}{4}} \lesssim \frac{1}{N^{\frac{1}{4}}} \cdot \left\{ \log_{\frac{1}{\gamma}} \left(\frac{C_8 N}{\log N} \right)^{\frac{1}{4}} \right\}^2 \cdot \left(\log \left[N \cdot \left\{ \log_{\frac{1}{\gamma}} \left(\frac{C_8 N}{\log N} \right)^{\frac{1}{4}} \right\} \Big/ \delta \right] \right)^{\frac{1}{4}},$$

which eventually leads to following when combined with (183),

$$\bar{\mathcal{E}}\big(\Upsilon_{\hat{\theta}_m^{(B)}}, \Upsilon_{\tilde{\theta}} \big) \leq O_p \left[\frac{1}{N^{1/(4q)}} \cdot \left\{ \log_{\frac{1}{\gamma}} \left(\frac{N}{\log N} \right) \right\}^{2/q} \cdot \left[\log \left\{ N \cdot \log_{\frac{1}{\gamma}} \left(\frac{N}{\log N} \right) \right\} \right]^{1/(4q)} \right],$$

which gives us desired result of Theorem 3.

C SUPPORTING RESULTS

C.1 BORROWED RESULTS

Theorem 4. *(Dudley's integral inequality) Assume the following two conditions.*

1. *(Subgaussian process) The stochastic process $(X_t)_{t\in\mathcal{T}} \in \mathbb{R}$ lies in the domain $(\mathcal{T}, \eta)$ which is a metric space w.r.t. the metric η with constant K_0, that is $\|X_t - X_s\|_{\psi_2} \leq K_0 \cdot \eta(t, s)$. To elaborate, $\eta : \mathcal{T} \times \mathcal{T} \to \mathbb{R}^+$ satisfies the following properties, (Theorem 8.1.6 of Vershynin (2018))*
   (a) $\eta(x, x) = 0$ *for* $\forall x \in \mathcal{T}$,
   (b) $\eta(x, y) > 0$ *if* $x \neq y$,
   (c) $\eta(x, y) = \eta(y, x)$ *for* $\forall x, y \in \mathcal{T}$,
   (d) $\eta(x, y) \leq \eta(x, z) + \eta(z, y)$ *for* $\forall x, y \in \mathcal{T}$.
2. *(Separability) There exists a set $\exists\tilde{\mathcal{T}} \subset \mathcal{T}$ and $\exists\tilde{\Omega} \in \Sigma$,where $(\Omega, \Sigma, \mathbb{P})$ is the corresponding probability space, such that (Definition 4.4 of Sen (2018))*
   (a) $\mathbb{P}(\tilde{\Omega}) = 1$,
   (b) $\tilde{\mathcal{T}}$ *is countable,*
   (c) $\forall w \in \tilde{\Omega}$, $\forall t \in \mathcal{T}$, *there is* $\exists(t_n)_{n\in\mathbb{N}} \in \tilde{\mathcal{T}}$ *such that* $\lim_{n\to\infty} X_{t_n}(w) = X_t(w)$ *holds.*

Then letting $\mathcal{N}(E, \eta, \epsilon)$ be the covering number (34) of set E w.r.t. η with $\epsilon > 0$, $\operatorname{diam}(\mathcal{T}; \eta)$ be the diameter of $\mathcal{T}$ w.r.t. η, and $C_1 > 0$ be a certain universal constant, for $\forall t_0 \in \mathcal{T}$, we have the following with probability bigger than $1 - 2\exp(-u^2)$,

$$\sup_{t\in\mathcal{T}} |X_t - X_{t_0}| \leq K_0 C \left\{ \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{T}, \eta, \epsilon)} \mathrm{d}\epsilon + u \cdot \operatorname{diam}(\mathcal{T}; \eta) \right\} \text{ for } \forall u > 0.$$

Theorem 5. *(Hoeffding's inequality for iid cases) For $X_1, \cdots, X_n$ that are iid distributed from some subgaussian distribution, for $\forall \epsilon \geq 0$, we have the following (Theorem 2.6.2 of Vershynin (2018)),*

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right) \leq 2 \cdot \exp\left(\frac{-C \cdot n \cdot \epsilon^2}{\|X_1 - \mu\|_{\psi_2}^2} \right).$$

Theorem 6. *(Adapted version of vector Bernstein inequality in Lemma 18 of Kohler & Lucchi (2017)) Let $\mathbf{X}_1, \cdots \mathbf{X}_n \in \mathbb{R}^d$ be independent random vectors that satisfy*

$$\mathbb{E}(\mathbf{X}_i) = 0,\ \|\mathbf{X}_i\| \leq \mu,\ \mathbb{E}\{\|\mathbf{X}_i\|^2\} \leq \sigma^2 \text{ for } \exists \mu, \sigma > 0.$$

Then we have the following for $\forall \epsilon \in (0, \sigma^2/\mu)$,

$$\mathbb{P}\left(\left\| \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i \right\| \geq \epsilon \right) \leq \exp\left\{ -N \cdot \frac{\epsilon^2}{8\sigma^2} + \frac{1}{4} \right\}.$$

Lemma 6. *(Lemma S4 of Wang et al. (2022)) When m is even, we can decompose $\{(j, j') : 1 \leq j < j' \leq n\}$ into $(n-1)$ groups, each of which contains $n/2$ pairs of (j, j') that share no repeated components at all.*

Proposition 1. *(Proposition 2.6.1 of Vershynin (2018)) Let $X_1, \cdots, X_n$ be independent mean-zero subgaussian random variables. Then we have*

$$\left\| \sum_{i=1}^n X_i \right\|_{\psi_2}^2 \leq C \cdot \sum_{i=1}^n \|X_i\|_{\psi_2}^2.$$

C.2 SUBPROOFS WITHIN THE MAIN PROOF

C.2.1 PROOF OF REMARK 1

Properties 1 and 2 are mentioned in Example 2.5.8 of Vershynin (2018).

Property 3 can be verified by following,

$$\big\|\mathbb{E}(X)\big\|_{\psi_2} = \inf\left\{t>0 : \mathbb{E}\left(\frac{\mathbb{E}(X)^2}{t^2}\right) \le 2\right\} = \inf\left\{t>0 : \frac{\mathbb{E}(X)^2}{t^2} \le 2\right\}.$$

Since the following holds for $\forall t > 0$

$$\frac{\mathbb{E}(X)^2}{t^2} \le \mathbb{E}\left(\frac{X^2}{t^2}\right),$$

this directly proves $\|\mathbb{E}(X)\|_{\psi_2} \le \|X\|_{\psi_2}$.

Property 4 can be verified as follows. Let $\mathbf{x} = (x_1, \cdots, x_d)^\intercal \in \mathbb{R}^d$ with unit norm $\|\mathbf{x}\| = 1$ be arbitrary. Denoting the canonical vectors as $\mathbf{e}_1, \cdots, \mathbf{e}_d$, we have

$$\begin{aligned}
\big|\langle \mathbf{X}, \mathbf{x}\rangle\big| &= \big|\langle \mathbf{X}, x_1\mathbf{e}_1\rangle + \cdots + \langle \mathbf{X}, x_d\mathbf{e}_d\rangle\big| \\
&\le |x_1| \cdot \big|\langle \mathbf{X}, \mathbf{e}_1\rangle\big| + \cdots |x_d| \cdot \big|\langle \mathbf{X}, \mathbf{e}_d\rangle\big| \\
&\le |x_1| \cdot \|\mathbf{X}\|_{\psi_2} + \cdots + |x_d| \cdot \|\mathbf{X}\|_{\psi_2} && \text{by the definition in Equation (33)} \\
&\le d\|\mathbf{X}\|_{\psi_2} && \because |x_j| \le 1 \text{ for } \forall j \in \{1, \cdots, d\}.
\end{aligned}$$

Property 5 is verified in Exercise 2.7.10 of Vershynin (2018).

C.2.2 Proof of Facts (39)

Letting $\hat{b}^{(1)}_\mu(s,a)$ and $\hat{b}^{(2)}_\mu(s,a)$ be the components of $\hat{\mathbf{p}}_{(1)}$ and $\hat{\mathbf{p}}_{(2)}$ (37) corresponding to (s,a), we have

$$\hat{b}^{(1)}_\mu(s,a) = \frac{N_1(s,a)}{\lfloor N/2 \rfloor} \quad \& \quad \hat{b}^{(2)}_\mu(s,a) = \frac{N_2(s,a)}{N - \lfloor N/2 \rfloor},$$

$$\text{with } N_1(s,a) = \sum_{i=1}^{\lfloor N/2 \rfloor} \mathbf{1}\big\{(S_i, A_i) = (s,a)\big\} \quad \& \quad N_2(s,a) = \sum_{i=\lfloor N/2 \rfloor + 1}^{N} \mathbf{1}\big\{(S_i, A_i) = (s,a)\big\}.$$

Let us first prove Fact 1. Letting $s, a \in \mathcal{S} \times \mathcal{A}$ be arbitrary, we have the following,

$$\begin{aligned}
\left|\hat{b}^{(1)}_\mu(s,a) - b_\mu(s,a)\right| &\le \sup_{s,a}\left|\hat{b}^{(1)}_\mu(s,a) - b_\mu(s,a)\right| = \|\hat{\mathbf{p}}_{(1)} - \mathbf{p}\|_\infty \\
&\le \|\hat{\mathbf{p}}_{(1)} - \hat{\mathbf{p}}\| \le \frac{1}{2} p_{\min} \cdot \epsilon \le \frac{1}{2} b_\mu(s,a) \cdot \epsilon \qquad (191) \\
&\le \frac{1}{2} b_\mu(s,a) \;\; (\because \epsilon < 1).
\end{aligned}$$

With the same logic, we also have

$$\left|\hat{b}^{(2)}_\mu(s,a) - b_\mu(s,a)\right| \le \frac{1}{2} b_\mu(s,a),$$

which gives us

$$\hat{b}^{(1)}_\mu(s,a),\ \hat{b}^{(2)}_\mu(s,a) \in \left[\frac{1}{2} b_\mu(s,a), \frac{3}{2} b_\mu(s,a)\right], \qquad (192)$$

that further implies Fact 1 based on Equation (37).

Regarding Fact 2, we have $N_1(s,a) \ge 1$ by following based on Equation (192),

$$\frac{N_1(s,a)}{N/2} = \frac{1}{2} b_\mu(s,a) \ge \frac{1}{2} p_{\min} > 0.$$

We also have $N_2(s,a) \ge 1$ by the same logic, so this leads to $N(s,a) = N_1(s,a) + N_2(s,a) \ge 2$.

Showing Fact 3 is straightforward by Equation (37) and Definition (38),

$$\|\hat{\mathbf{p}} - \mathbf{p}\| \le \frac{\lceil N/2 \rceil}{N} \|\hat{\mathbf{p}}_{(1)} - \mathbf{p}\| + \frac{N - \lceil N/2 \rceil}{N} \|\hat{\mathbf{p}}_{(2)} - \mathbf{p}\| < \frac{1}{2} p_{\min} \cdot \epsilon.$$

C.2.3 PROOF OF COROLLARY 1

Based on Proposition 1, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|_{\psi_2}^2 \le C\cdot\sum_{i=1}^{n}\left\|\frac{X_i}{n}\right\|_{\psi_2}^2 \le \frac{C}{n^2}\cdot n\cdot\|X_1\|_{\psi_2}^2 = \frac{C}{n}\|X_1\|_{\psi_2}^2,$$

which directly implies the desired result.

C.2.4 PROOF OF LEMMA 1

Let us temporarily assume that $N\in\mathbb{N}$ is an even number. Newly define $Y_{ij}:=(X_{ij}+X_{ji})/2$ for $1\le i<j\le N$. Then by Lemma 6, we can group them into $N-1$ groups $G_1,\cdots,G_{N-1}$, each of which contains $N/2$ pairs of (i,j) where no pairs overlap in any components. Based on this, we can obtain the following bound,

$$\begin{aligned}
\mathbb{P}\left\{\left|\frac{1}{N(N-1)}\sum_{i\neq j}X_{ij}-\mathbb{E}(X_{12})\right|\ge\epsilon\right\} &\le \mathbb{P}\left\{\left|\frac{1}{N(N-1)/2}\sum_{i<j}Y_{ij}-\mathbb{E}(Y_{12})\right|\ge\epsilon\right\}\\
&\le \mathbb{P}\left\{\frac{1}{N-1}\sum_{k=1}^{N-1}\left|\frac{1}{N/2}\sum_{(i,j)\in G_k}Y_{ij}-\mathbb{E}(Y_{12})\right|\ge\epsilon\right\}\\
&\le \mathbb{P}\left\{\left|\frac{1}{N/2}\sum_{(i,j)\in G_k}Y_{ij}-\mathbb{E}(Y_{12})\right|\ge\epsilon \text{ for } \exists k\in\{1,\cdots,N-1\}\right\}\\
&\le (N-1)\cdot\mathbb{P}\left\{\left|\frac{1}{N/2}\sum_{(i,j)\in G_1}Y_{ij}-\mathbb{E}(Y_{12})\right|\ge\epsilon\right\}\\
&\le 2N\cdot\exp\left\{\frac{-C_2\cdot N\cdot\epsilon^2}{\|X_{12}-\mathbb{E}(X_{12})\|_{\psi_2}^2}\right\}, \text{ by Theorem 5} \qquad (193)
\end{aligned}$$

where the last line holds because $\|Y_{12}-\mathbb{E}(Y_{12})\|_{\psi_2}=\|\frac{1}{2}(X_{12}-\mathbb{E}(X_{12}))+\frac{1}{2}(X_{21}-\mathbb{E}(X_{21}))\|_{\psi_2}\le\|X_{12}-\mathbb{E}(X_{12})\|_{\psi_2}$.

Now let us expand our result towards odd numbers $N\in\mathbb{N}$. Since we are assuming $N\ge 2$ in our assumption, N being odd implies that $N\ge 3$. Then we can obtain the following result,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{N(N-1)}\sum_{i\neq j}X_{ij}-\mathbb{E}(X_{12})\right|\ge\epsilon\right\} = \mathbb{P}\left\{\left|\frac{1}{N(N-1)/2}\sum_{i<j}Y_{ij}-\mathbb{E}(Y_{12})\right|\ge\epsilon\right\}\\
&\le \mathbb{P}\left\{\left|\frac{N-2}{N}\cdot\frac{1}{(N-1)(N-2)/2}\sum_{i<j}^{N-1}\left\{Y_{ij}-\mathbb{E}(Y_{12})\right\}\right|\ge\frac{\epsilon}{2} \qquad \text{by technique (56)}\right.\\
&\qquad\qquad \left. \text{or } \left|\frac{1}{N/2}\cdot\frac{1}{N-1}\sum_{i=1}^{N-1}\left\{Y_{iN}-\mathbb{E}(Y_{12})\right\}\right|\ge\frac{\epsilon}{2}\right\}\\
&\le \mathbb{P}\left\{\left|\frac{1}{(N-1)(N-2)/2}\sum_{i<j}^{N-1}\left\{Y_{ij}-\mathbb{E}(Y_{12})\right\}\right|\ge\frac{\epsilon}{2}\right\}\\
&\qquad\qquad + \mathbb{P}\left\{\left|Y_{iN}-\mathbb{E}(Y_{12})\right|\ge\frac{N}{4}\epsilon \text{ for } \exists i\in\{1,\cdots,N-1\}\right\}\\
&\le 2(N-1)\cdot\exp\left\{\frac{-C_2\cdot N\cdot(\epsilon/2)^2}{\|X_{12}-\mathbb{E}(X_{12})\|_{\psi_2}^2}\right\}+(N-1)\cdot\mathbb{P}\left\{\left|Y_{12}-\mathbb{E}(Y_{12})\right|\ge\frac{N}{4}\epsilon\right\} \qquad \text{by (193)}\\
&\le 4N\cdot\exp\left\{\frac{C_3\cdot N\cdot\epsilon^2}{\|X_{12}-\mathbb{E}(X_{12})\|_{\psi_2}^2}\right\}, \qquad \text{by Theorem 5}
\end{aligned}$$

where the third last line holds since $N-1$ is an even number.

C.2.5 PROOF FOR COROLLARY 2

It is trivial for $\epsilon > 1$ since the probability term in the LHS shall be 0, so we will assume $\epsilon \in (0, 1)$. Let $\mathbf{Y}_i := \mathbf{X}_i - \mathbf{p}$, and we can see that

$$\mathbb{E}(\mathbf{Y}_1) = \mathbf{0} \in \mathbb{R}^H,$$

$$\|\mathbf{Y}_1\| = \|\mathbf{X_1} - \mathbf{p}\| \leq \|\mathbf{X}_1\| + \|\mathbf{p}\| \leq 2 \ \left(\because \ \|\mathbf{p}\| = \sqrt{\sum_{h=1}^{H} p_h^2} \leq \sqrt{\sum_{h=1}^{H} p_h} = 1\right).$$

Therefore we can let $\mu = 2$ and $\sigma^2 = 4$, and applying Theorem 6 gives us the desired result for $\epsilon \in (0, 2)$, so it validates the result for $\epsilon \in (0, 1)$.

C.2.6 PROOF OF REMARK 2

Let $t > 0$ be arbitrarily chosen. Under Assumption 7, $\|\theta_1 - \theta_2\| \leq t/L$ implies $\tilde{\eta}(\theta_1, \theta_2) \leq t$. Letting $M_0 = \mathcal{N}(\Theta, \|\cdot\|, t/L)$ defined in Equation (34), and $\theta_1, \cdots, \theta_{M_0}$ to be such centers, we have

$$\Theta \subset \bigcup_{i=1}^{M_0} N_{\|\cdot\|}(\theta_i, t/L) \subset \bigcup_{i=1}^{M_0} N_{\tilde{\eta}}(\theta_i, t), \quad \therefore \ \mathcal{N}(\Theta, \tilde{\eta}, t) \leq M_0 = \mathcal{N}(\Theta, \|\cdot\|, t/L),$$

which leads to

$$\begin{aligned}\int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)}\mathrm{d}t &\leq \int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, t/L)}\mathrm{d}t \leq L \cdot \int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, t)}\mathrm{d}t. \\ &\leq L \cdot \int_0^{\mathrm{diam}(\Theta; \|\cdot\|)} \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, t)}\mathrm{d}t, \end{aligned} \tag{194}$$

where the last line holds since we have $\mathcal{N}(\Theta, \|\cdot\|, t) = 1$ for $\forall t \geq \mathrm{diam}(\Theta; \|\cdot\|)$. Here we can use the well-known Volume Comparison Lemma that states the following for $\mathcal{B}_r(p) = \{\theta \in \mathbb{R}^p : \|\theta\| \leq r\}$ and for $\forall t > 0$,

$$\left(\frac{r}{t}\right)^p \leq \mathcal{N}(\mathcal{B}_r(d), \|\cdot\|, t) \leq \left(1 + \frac{2r}{t}\right)^p.$$

By applying this, we can take up from Inequality (194) as follows,

$$\begin{aligned}\int_0^\infty \sqrt{\log \mathcal{N}(\Theta, \tilde{\eta}, t)}\mathrm{d}t &\leq L \cdot \int_0^{\mathrm{diam}(\Theta; \|\cdot\|)} \sqrt{p \cdot \log\left\{1 + \frac{2 \cdot \mathrm{diam}(\Theta; \|\cdot\|)}{t}\right\}}\mathrm{d}t \\ &\leq L \cdot \int_0^{\mathrm{diam}(\Theta; \|\cdot\|)} \sqrt{p \cdot \log\left\{\frac{3 \cdot \mathrm{diam}(\Theta; \|\cdot\|)}{t}\right\}}\mathrm{d}t \\ &= 3 \cdot \mathrm{diam}(\Theta; \|\cdot\|) \cdot L\sqrt{p} \cdot \int_0^{1/3} \sqrt{\log\left(\frac{1}{t}\right)}\mathrm{d}t \\ &\leq 6\sqrt{2\pi} \cdot L\sqrt{p} \cdot \mathrm{diam}(\Theta; \|\cdot\|), \end{aligned} \tag{195}$$

where the last line holds since

$$\begin{aligned}\int_0^{1/3} \sqrt{\log\left(\frac{1}{t}\right)}\mathrm{d}t &\leq \int_\infty^{\sqrt{\log 3}} z \cdot (-2z\exp(-z^2))\mathrm{d}t \ \text{ where } z = \sqrt{\log\left(\frac{1}{t}\right)} \\ &\leq 2\int_{-\infty}^{\infty} z^2 \exp(-z^2)\mathrm{d}t \\ &= 2\sqrt{2\pi}\int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}}\exp(-z^2)\mathrm{d}t \\ &= 2\sqrt{2\pi}. \end{aligned}$$

C.2.7 PROOF OF LEMMA 3

Letting $\mu_0, \mu_1, \mu_2 \in \mathcal{P}$ be arbitrary, we have the following decomposition, where ρ is the metric such that $\mathcal{E}(P, Q) = \rho^2(P, Q)$ as mentioned in Property 3,

$$\begin{aligned}\mathcal{E}(\mu_0, \mu_1) &= \rho^2(\mu_0, \mu_1) \leq \big\{\rho(\mu_0, \mu_1) + \rho(\mu_2, \mu_1)\big\}^2 \\ &= \rho^2(\mu_0, \mu_2) + 2 \cdot \rho(\mu_0, \mu_1) \cdot \rho(\mu_2, \mu_1) + \rho^2(\mu_2, \mu_1) \\ &= \mathcal{E}(\mu_0, \mu_2) + \sqrt{\mathcal{E}(\mu_1, \mu_2)} \cdot \big\{2\sqrt{\mathcal{E}(\mu_0, \mu_2)} + \sqrt{\mathcal{E}(\mu_1, \mu_2)}\big\}.\end{aligned}$$

Now let us extend this result towards $\bar{\mathcal{E}}$ as follows,

$$\begin{aligned}\bar{\mathcal{E}}(\Upsilon_0, \Upsilon_1) &\leq \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) + \sum_{s,a} b_\mu(s,a) \cdot \sqrt{\mathcal{E}\{\Upsilon_1(s,a), \Upsilon_2(s,a)\}} \\ &\qquad \times \left\{2\sqrt{\mathcal{E}\{\Upsilon_0(s,a), \Upsilon_2(s,a)\}} + \sqrt{\mathcal{E}\{\Upsilon_1(s,a), \Upsilon_2(s,a)\}}\right\} \\ &= \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) + \sum_{s,a} \left[\sqrt{b_\mu(s,a) \cdot \mathcal{E}\{\Upsilon_1(s,a), \Upsilon_2(s,a)\}}\right. \\ &\qquad \left. \times \sqrt{b_\mu(s,a)} \cdot \left\{2\sqrt{\mathcal{E}\{\Upsilon_0(s,a), \Upsilon_2(s,a)\}} + \sqrt{\mathcal{E}\{\Upsilon_1(s,a), \Upsilon_2(s,a)\}}\right\}\right] \\ &\leq \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) + \left(\sum_{s,a} b_\mu(s,a) \cdot \mathcal{E}\{\Upsilon_1(s,a), \Upsilon_2(s,a)\}\right)^{1/2} \\ &\qquad \times \left[\sum_{s,a} b_\mu(s,a) \cdot \left\{2\sqrt{\mathcal{E}\{\Upsilon_0(s,a), \Upsilon_2(s,a)\}} + \sqrt{\mathcal{E}\{\Upsilon_1(s,a), \Upsilon_2(s,a)\}}\right\}^2\right]^{1/2} \\ &\leq \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) + \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)^{1/2} \\ &\qquad \times \left[\sum_{s,a} b_\mu(s,a) \cdot 2 \cdot \left\{4\mathcal{E}\{\Upsilon_0(s,a), \Upsilon_2(s,a)\} + \mathcal{E}\{\Upsilon_1(s,a), \Upsilon_2(s,a)\}\right\}\right]^{1/2} \\ &\leq \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) + 4 \cdot \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)^{1/2} \cdot \big\{\bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) + \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)\big\}^{1/2},\end{aligned}$$

where the third line used Cauchy-Schwartz inequality, the forth line is based on $(x+y)^2 \leq 2(x^2 + y^2)$. Eventually we have

$$\bar{\mathcal{E}}(\Upsilon_0, \Upsilon_1) - \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) \leq 4 \cdot \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)^{1/2} \cdot \big\{\bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) + \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)\big\}^{1/2},$$

and by the symmetry, we also have

$$\bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2) - \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_1) \leq 4 \cdot \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)^{1/2} \cdot \big\{\bar{\mathcal{E}}(\Upsilon_0, \Upsilon_1) + \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)\big\}^{1/2}.$$

This eventually leads to

$$\begin{aligned}&\left|\bar{\mathcal{E}}(\Upsilon_0, \Upsilon_1) - \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2)\right| \\ &\qquad \leq 4 \cdot \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)^{1/2} \cdot \left[\max\left\{\bar{\mathcal{E}}(\Upsilon_0, \Upsilon_1), \bar{\mathcal{E}}(\Upsilon_0, \Upsilon_2)\right\} + \bar{\mathcal{E}}(\Upsilon_1, \Upsilon_2)\right]^{1/2}.\end{aligned}$$

C.2.8 PROOF OF LEMMA 5

Let $z > 0$ be arbitrary, and let $E_i = X_i \cdot \mathbf{1}\big\{|X_i| \leq z\big\}$ and $F_i = X_i \cdot \mathbf{1}\big\{|X_i| > z\big\}$. Since we have $X_i = E_i + F_i$, we have

$$\mathbb{E}\left|\sum_{i=1}^n X_i\right| \leq \mathbb{E}\left|\sum_{i=1}^n E_i\right| + \mathbb{E}\left|\sum_{i=1}^n F_i\right|.$$

Note that each term satisfies

$$\mathbb{E}\Big|\sum_{i=1}^{n} U_i\Big| \leq \mathbb{E}\Big\{\big(\sum_{i=1}^{n} E_i\big)^2\Big\}^{1/2} = \sqrt{\mathbb{V}\Big(\sum_{i=1}^{n} E_i\Big)} \quad \because \mathbb{E}(E_1) = 0$$
$$\leq \sqrt{n}\cdot \mathbb{V}(E_1)^{1/2} = \sqrt{n}\cdot \big[\mathbb{E}\big\{X_1^2 \cdot \mathbf{1}\big(|X_1| \leq z\big)\big\}\big]^{1/2}$$
$$\mathbb{E}\Big|\sum_{i=1}^{n} V_i\Big| \leq n\cdot \mathbb{E}|V_1| = n\cdot \mathbb{E}\big\{|X_1|\cdot \mathbf{1}\big(|X_1| > z\big)\big\},$$

which leads to

$$\mathbb{E}\big|\bar{X}_n\big| \leq \frac{1}{\sqrt{n}}\cdot \Big\{\mathbb{E}\big\{X_1^2\cdot \mathbf{1}(|X_1|\leq z)\big\}\Big\}^{1/2} + \mathbb{E}\big\{|X_1|\cdot \mathbf{1}(|X_1| > z)\big\}.$$

Since $z > 0$ was arbitrary, we eventually have

$$\mathbb{E}\big|\bar{X}_n\big| \leq \inf_{z>0}\left[\frac{1}{\sqrt{n}}\cdot \Big\{\mathbb{E}\big\{X_1^2\cdot \mathbf{1}(|X_1|\leq z)\big\}\Big\}^{1/2} + \mathbb{E}\big\{|X_1|\cdot \mathbf{1}(|X_1| > z)\big\}\right] \to 0 \text{ as } n\to\infty.$$

C.2.9 Proof of Equation (172)

Since we have

$$\big|F_m(\theta) - F(\theta)\big| = \big|\bar{\mathcal{E}}\big\{\Upsilon_\theta, (\mathcal{T}^\pi)^m\Upsilon_\theta\big\} - \bar{\mathcal{E}}\big\{\Upsilon_\theta, \Upsilon_\pi\big\}\big|$$
$$\leq \sum_{s,a} b_\mu(s,a)\cdot \left|\mathcal{E}\Big\{\Upsilon_\theta(s,a), (\mathcal{T}^\pi)^m\Upsilon_\theta(s,a)\Big\} - \mathcal{E}\Big\{\Upsilon_\theta(s,a), \Upsilon_\pi(s,a)\Big\}\right|,$$

we can further bound it using the technique shown in Line (101). With a new notation $Z(s,a;\pi) = Z_\pi(s,a)$ and the abuse of notation $(\mathcal{T}^\pi)^m\theta$ introduced in Definition (114), we can derive

$$\left|\mathcal{E}\Big\{\Upsilon_\theta(s,a), (\mathcal{T}^\pi)^m\Upsilon_\theta(s,a)\Big\} - \mathcal{E}\Big\{\Upsilon_\theta(s,a), \Upsilon_\pi(s,a)\Big\}\right|$$
$$\leq 2\cdot \tilde{\eta}\big\{(\mathcal{T}^\pi)^m\theta, \pi\big\} + \tilde{\eta}\big\{(\mathcal{T}^\pi)^m\theta, \pi\big\} + \tilde{\eta}\big\{(\mathcal{T}^\pi)^m\theta, \pi\big\} \text{ by Assumption 5}$$
$$\leq 4\gamma^m\cdot \tilde{\eta}(\theta,\pi) \qquad \text{by Assumption 8 ,}$$
$$\leq 4\gamma^m\cdot \big\{\tilde{\eta}(\tilde{\theta},\pi) + \sup_\theta \tilde{\eta}(\theta,\tilde{\theta})\big\} \leq 4\gamma^m\cdot \big\{\tilde{\eta}(\tilde{\theta},\pi) + L\cdot \mathrm{diam}(\Theta;\|\cdot\|)\big\}, \tag{196}$$

where the last line used Assumption 7. Again, as we did in Trick (102), we used the contractive property (Assumption 8) of Bellman operator with respect to $\tilde{\eta}$ in the third inequality.

C.2.10 Proof of Remark 3

Let us show the first fact. $\psi^{-1}(\cdot)$ is an increasing function, since the following holds for arbitrary y_1, y_2 ($y_1 \leq y_2$),

$$\psi^{-1}(y_1) = \inf_{\delta>0}\big\{\psi(\delta)\geq y_1\big\} \leq \inf_{\delta>0}\big\{\psi(\delta)\geq y_2\big\} = \psi^{-1}(y_2).$$

If $y > \sup_{\delta>0}\psi(\delta)$, then $\psi^{-1}(y) = \inf(\varnothing) := \infty$ by definition of infimum. To prove $\lim_{y\to 0}\psi^{-1}(y) = 0$, it suffices to show right-side convergence. Towards that end, we let $\delta > 0$ be sufficiently small, that is $\delta < \sup_{\theta\in\Theta}\|\theta - \tilde{\theta}\|$, which leads to following by Definition (176),

$$\psi(\delta) = \inf_{\theta\in\Theta:\|\theta-\tilde{\theta}\|\geq\delta} F(\theta) - F(\tilde{\theta}) \geq c_q\cdot\delta^q \quad \text{by Assumption 6,}$$
$$\therefore\ \psi^{-1}(y) \leq \frac{1}{{c_q}^{1/q}}\cdot y^{1/q} \quad \text{for } \forall y\in[0, \sup_{\delta>0}\psi(\delta)). \tag{197}$$

This gives us $\psi^{-1}(y)\to 0$ as $y\to 0$.

The second fact can be shown as follows. Let $y \in [0, \psi(\sup_{\theta\in\Theta} \|\theta - \tilde{\theta}\|)]$ be arbitrary ($\psi^{-1}(y) < \infty$), and let $\epsilon > 0$ be arbitrarily small. Letting $\delta_0 := \inf_{\delta>0} \{\psi(\delta) \geq y\}$, we have following,

$$\psi\{\psi^{-1}(y) + \epsilon\} = \psi\left[\inf_{\delta>0}\left\{\psi(\delta) \geq y\right\} + \epsilon\right] = \psi(\delta_0 + \epsilon) \geq y,$$
$$\therefore \lim_{\epsilon\to 0+} \psi\{\psi^{-1}(y) + \epsilon\} \geq y.$$

The third fact can be validated by extending the proof of Example 1.3. of Sen (2018). Suppose that there exists a minimizer $\hat{\theta} \in \arg\min_{\theta\in\Theta} \hat{F}(\theta)$ such that $\|\hat{\theta} - \theta_0\| \geq \delta$. Now we temporarily make new notations $G = -F$ and $\hat{G} = -\hat{F}$, which leads to $\hat{\theta} \in \arg\max_{\theta\in\Theta} \hat{G}(\theta)$ and $\psi(\delta) = G(\tilde{\theta}) - \sup_{\theta\in\Theta:\|\theta-\tilde{\theta}\|\geq\delta} G(\theta)$. Then we have $\hat{G}(\tilde{\theta}) \leq \sup_{\theta\in\Theta:\|\theta-\tilde{\theta}\|\geq\delta} \hat{G}(\theta)$, which leads to

$$\sup_{\theta\in\Theta:\|\theta-\tilde{\theta}\|\geq\delta} \hat{G}(\theta) - \hat{G}(\tilde{\theta}) + \psi(\delta) \geq \psi(\delta),$$
$$\therefore \sup_{\theta\in\Theta:\|\theta-\tilde{\theta}\|\geq\delta} \{(\hat{G}(\theta) - G(\theta)) - (\hat{G}(\theta) - G(\tilde{\theta}))\} \geq \psi(\delta),$$

from which we can derive $\sup_{\theta\in\Theta} |\hat{F}(\theta) - F(\theta)| \geq \frac{1}{2}\psi(\delta)$. Up to this point we have derived

$$\exists\hat{\theta} \in \arg\min_{\theta\in\Theta} \hat{F}(\theta) \text{ such that } \|\hat{\theta} - \theta_0\| \geq \delta \quad \rightarrow \quad \sup_{\theta\in\Theta} |\hat{F}(\theta) - F(\theta)| \geq \frac{1}{2}\psi(\delta). \tag{198}$$

Now let us assume that there exists a value $\hat{\theta} \in \Theta$ such that $\|\hat{\theta}-\tilde{\theta}\| > \delta$. Then we have $\|\hat{\theta}-\tilde{\theta}\| = \delta+\epsilon_0$ for some $\epsilon_0 > 0$. By (198), we have $\sup_{\theta\in\Theta} |\hat{F}(\theta)-F(\theta)| \geq \frac{1}{2}\psi(\delta+\epsilon_0) \geq \frac{1}{2}\lim_{\epsilon\to 0+}\psi(\delta+\epsilon)$, by using the fact that $\psi(\cdot)$ is an increasing function. This gives us the desired result.

C.2.11 Continuity of bootstrap-based objective function (23)

By replicating the trick (102), we have

$$\begin{aligned}
&\left|\mathcal{E}\left\{\Upsilon_{\theta_1}(s,a), \mathcal{B}_m\Upsilon_{\theta_1}(s,a)\right\} - \mathcal{E}\left\{\Upsilon_{\theta_2}(s,a), \mathcal{B}_m\Upsilon_{\theta_2}(s,a)\right\}\right| \\
&\quad \leq 4\cdot\tilde{\eta}(\theta_1,\theta_2) + 4\cdot\tilde{\eta}\{\mathcal{B}_m\Upsilon_{\theta_1}, \mathcal{B}_m\Upsilon_{\theta_2}\} \leq 4(1+\gamma^m)\cdot\tilde{\eta}(\theta_1,\theta_2) \quad \text{by Assumption 8} \\
&\quad \leq 8L\cdot\|\theta_1 - \theta_2\| \quad \text{by Assumption 7,}
\end{aligned}$$

which further leads to $|\hat{F}_m^{(B)}(\theta_1) - \hat{F}_m^{(B)}(\theta_2)| \leq 8L\cdot\|\theta_1 - \theta_2\|$, implying Lipschitz continuity.

C.3 Further Discussion

C.3.1 Rate Mismatch between Theorem 2 and B.4.3

Below Theorem 3, we have mentioned that the theoretical result for nonrealizable scenario B.4.3 does not degenerate to Theorem 2 under realizable setting with $m = 1$. There are two factors that slow down the convergence rate throughout the proof of Appendix B.

First, we are forming multi-step trajectories by resampling from the collected data $\{(s_i, a_i, r_i, s_i')\}_{i=1}^N$, as mentioned before (22). This practice of re-using the samples no longer guarantees the independence of the data that we use. Then we cannot make use of common tricks for showing sample mean converging to the population mean (e.g. Theorems 4, 5, 6 employed in A.6.3 and A.6.4), since since they require independence of samples. Therefore, we resort to another trick, that is Lemma 3 that resembles triangular inequality, but slows down the rate by $1/2$.

Second, in non-realizable cases, the converging target is no longer the true distribution. Instead, we compromise our goal into the "best approximation" $\Upsilon_{\tilde{\theta}}$ defined in (21). That being said, Theorem 2, which only required us to derive the convergence rate of estimated Bellman residual (our objective function), is not enough now. Therefore, unlike realizable scenario, it requires one additional procedure, which is obtaining convergence of the minimizing parameter. This leads us to take into account the function $\psi(\cdot)$ in (176), which further slows down the rate by $1/q$.

Due to the aforementioned reasons, the rate in non-realizable scenario is slower than realizable scenario by $1/(2q)$, that is $\sqrt{\log N/N}$ verses $\sqrt[4q]{\log N/N}$.

C.3.2 REGARDING LINEAR MDP

Linear MDP may be useful in expanding our method into continuous state-action space by expressing the transition probability and reward distribution as linear combinations of multiple features $\phi(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$. To elaborate, it may allow us to construct an unbiased estimate of energy distance for continuous state-action space, where a single s, a cannot be observed twice or more (almost surely). In such cases, our current estimation in Equation (10) leads to a biased estimate of the third term of expansion of $\mathcal{E}\{\Upsilon_\theta(s,a), \mathcal{T}^\pi\Upsilon_\theta(s,a)\}$ shown in Equation (9), which is $\mathbb{E}\|R_\alpha + \gamma \cdot Z_\alpha(S'_\alpha, A'_\alpha; \theta) - R_\beta - \gamma \cdot Z_\beta(S'_\beta, A'_\beta; \theta)\|$. Here, we have $(R_\alpha, S'_\alpha), (R_\beta, S'_\beta) \sim p(\cdots|s,a)$ and $A'_\alpha \sim \pi(\cdot|S'_\alpha)$, $A'_\beta \sim \pi(\cdot|S'_\beta)$, where α, β are for indicating independent copies. According to our current way of estimation (10), our estimate becomes $\mathbb{E}\|\gamma \cdot Z_\alpha(s', A'_1; \theta) - \gamma \cdot Z_\beta(s', A'_2; \theta)\|$ with $A'_1, A'_2 \sim \pi(\cdot|s')$ being independent. This is because the terms $\hat{R}_\alpha = r$ and $\hat{R}_\beta = r$ have the same value, and are thereby cancelled out when there is only a single observation (s, a, r, s'). This bias might be prevented if we can estimate the distribution of terms $\hat{R}_\alpha, \hat{R}_\beta \sim \hat{p}(\cdots|s,a)$ (for s, a that was observed only once) by expressing them with features. In addition, leveraging the structure of the linear MDP may improve the efficiency of our method. It will be interesting to see what the structure of return distribution induced by the linear MDP is, which requires further investigation.

D Other Materials

D.1 Comparison between DRL methods in details

In Table 3, we specified the "contractive distance," which is the distance that makes Bellman operator contractive, or that compined with projected operators $\Pi_{\mathcal{C}}, \Pi_{\mathbb{W}_1} : \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}} \to \mathcal{P}(\mathbb{R}^d)^{\mathcal{S}\times\mathcal{A}}$. Each of these is the projection towards categorized support items (Bellemare et al., 2017a) and with respect to Wassserstein-1 metric (Dabney et al., 2018b), respectively. Then we compared them with the objective functions, and most methods had misalignment between these two. That is, the contraction and objective function are either based upon different distances, or their ways of extension (expecation or supremum) are different. FLE (Wu et al., 2023) and EBRM are the only two methods that could overcome this issue, which allowed them to prove convergence of the estimation towards some target with a certain rate (Table 4).

Table 3: Comparison between Contractive Distances and Objective Functions

Method	**Operator & Contractive Distance**	**Objective Function**
Categorical algorithm (Bellemare et al., 2017b) (Rowland et al., 2018)	$\mathcal{T}^\pi$: Wasserstein-1 (supremum-extended) $\Pi_{\mathcal{C}}\mathcal{T}^\pi$: Cramer distance (supremum-extended)	Cross Entropy (expectation-extended)
QRTD/QRDQN (Dabney et al., 2018a) IQN (Dabney et al., 2018a) FQF (Yang et al., 2019)	$\Pi_{\mathbb{W}_1}\mathcal{T}^\pi$: Wasserstein-$\infty$ (supremum-extended)	quantile Huber Loss (expectation-extended)
EDRL (Rowland et al., 2019)	no additional result about contraction	Expectile Regression Loss (expectation-extended)
MMDRL (Nguyen-Tang et al., 2021)	$\mathcal{T}^\pi$: supremum-extended MMD_k (unrectified kernel) (supremum-extended)	MMD_k^2 (expectation-extended)
SinkhornDRL (Sun et al., 2022)	$\mathcal{T}^\pi$: Sinkhorn Divergence (supremum-extended)	Sinkhorn Divergence (expectation-extended)
MD3QN (Zhang et al., 2021)	$\mathcal{T}^\pi$: Wasserstein-p ($p \geq 1$) (supremum-extended)	MMD_k^2 (gaussian kernel) (expectation-extended)
FLE (Wu et al., 2023)	$\mathcal{T}^\pi$: Squared Wasserstein-p ($p \geq 1$) (expectation-extended)	log-Likelihood (expectation-extended)
EBRM (our method)	$\mathcal{T}^\pi$, $(\mathcal{T}^\pi)^m$ ($m \geq 2$) : Energy Distance (expectation-extended) (not exactly contraction: Theorem 1)	Energy Distance (expectation-extended)

Table 4: Convergence towards some target distribution

Method	Convergence towards Target
Categorical algorithm (Bellemare et al., 2017a) (Rowland et al., 2018)	$(\Pi_C \mathcal{T}^\pi)^m \Upsilon_0 \to \Upsilon_C \neq \Upsilon_\pi$ as $m \to \infty$ Convergence rate not suggested (Assumption: bounded reward)
QRTD/QRDQN (Dabney et al., 2018a) IQN (Dabney et al., 2018a) FQF (Yang et al., 2019)	NA
EDRL (Rowland et al., 2019)	only implies convergence of expectation, not the distribution.
MMDRL (Nguyen-Tang et al., 2021)	NA
SinkhornDRL (Sun et al., 2022)	NA
MD3QN (Zhang et al., 2021)	NA
FLE (Wu et al., 2023)	$(\mathcal{T}^\pi)^m \Upsilon_0 \to \Upsilon_\pi$ as $m \to \infty$ Convergence rate suggested (Assumption: bounded reward, completeness)
EBRM (our method)	Convergence towards truth / best approximation Convergence rate suggested

D.2 Realizable and Nonrealizable Models

D.2.1 Introduction of Simulation Settings

With the state-action space in Section 5, let us assume that s, a is given. The agent moves by one in the direction of $a \in \{-1, 1\}$, that is $s' = s + a$ (value of which fully determines the reward distribution (199)). If the agent is already blocked by the direction, that is $(s, a) = (1, -1)$ or $(s, a) = (30, 1)$, it stays at the same position $s' = s$. With given values of $A_0 > 0, p_0 \in (0, 1), \sigma_0^2 > 0$, our transition $p(r, s'|s, a)$ is characterized by:

$$\text{Conditioned on } S + A = k,\ R \sim N(\mu_k, \sigma_0^2) \quad \text{where} \quad \mu_k = \begin{cases} A_0 \cdot p_0^k & (k = 0, \cdots, 30) \\ 0 & (k = 31) \end{cases} \tag{199}$$
$$\text{and } S' = k \text{ if } k \in \{1, \cdots, 30\}, \quad S' = 30 \text{ if } k = 31, \quad S' = 1 \text{ if } k = 0.$$

We assume infinite-horizontal setting. Following the environment (199) and target policy (24), we have $\Upsilon_\pi(s, a)$ to be normal distributions, with expectation and variance as follows,

$$\mathbb{E}\big\{Z_\pi(i, -1)\big\} = A_0 \cdot p_0^{i-1} \cdot \frac{1 - (\gamma p_0)^{32-i}}{1 - \gamma p_0} \ \ (i \geq 2), \quad \mathbb{V}\big\{Z_\pi(i, \pm 1)\big\} = \frac{\sigma_0^2}{1 - \gamma^2} \ \ (i \geq 1), \tag{200}$$

$$\mathbb{E}\big\{Z_\pi(i, 1)\big\} = A_0 \cdot p_0^{i+1} \cdot \frac{1 - (\gamma p_0)^{30-i}}{1 - \gamma p_0} \ \ (i \geq 1), \quad \mathbb{E}\big\{Z_\pi(1, -1)\big\} = A_0 + \gamma \cdot \mathbb{E}\big\{Z_\pi(1, 1)\big\}.$$

We always let $A_0 = 100$, $p_0 = 0.9$ throughout the simulations.

First, we assumed a realizable scenario where the correct model (200) is known (Appendix D.2.2), only not knowing the values of A_0, p_0. Here, we always assumed $\gamma = 0.99$ and tried two settings with $\sigma_0^2 = 20$ and $\sigma_0^2 = 5000$. Second, we always tried the non-realizable scenario where there is a model misspecification (201), as will be demonstrated in Appendix D.2.3. Here, we always assumed $\sigma^2 = 20$, trying $\gamma = 0.50$ and $\gamma = 0.99$.

D.2.2 REALIZABLE SCENARIO

In the realizable scenario, we assume that Equations (200) are known, except the values of $A_0 = 100$ and $p_0 = 0.9$. The distributions can be plotted as Figure 2, each for $\sigma_0^2 = 20$ and $\sigma_0^2 = 5000$.

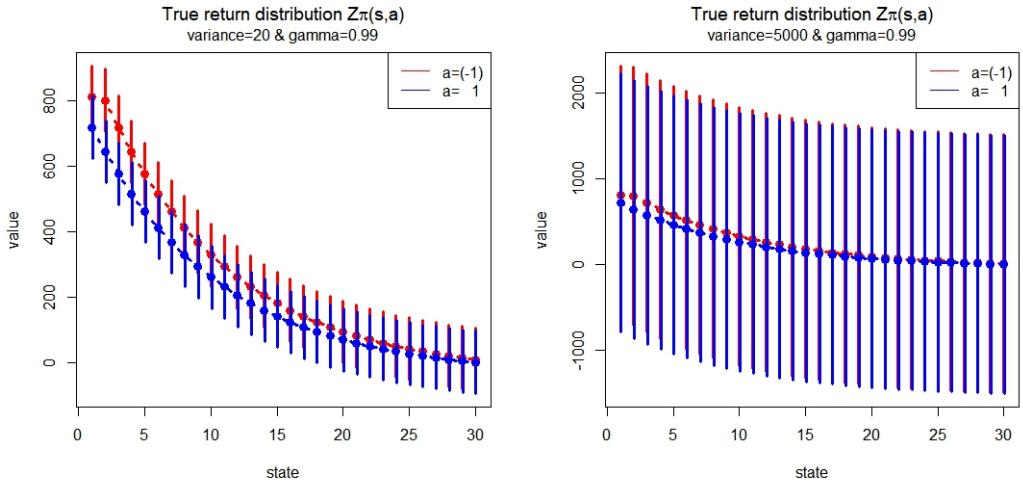

Figure 2: Red and blue represent the distributions of $Z_\pi(s,-1)$ and $Z_\pi(s,1)$ respectively. The dots indicate the expectation values and the vertical bars include (Mean$\pm$ 3·SD).

D.2.3 NON-REALIZABLE SCENARIO

On the other hand, in the non-realizable scenario, we assume that we are not aware of the true model (200). Instead, we assume that we are only aware of the decreasing trend demonstrated in Figure 2. That being said, we apply the following linear model that holds for all $1 \leq i \leq 30$. with four different parameters $\beta_L, \beta_R, \beta_1 \in \mathbb{R}$, $\sigma^2 > 0$,

$$\mathbb{E}\big(Z_\pi(i,-1)\big) = \beta_L + \beta_1 \cdot i, \;\; \mathbb{E}\big(Z_\pi(i,1)\big) = \beta_R + \beta_1 \cdot i, \;\; \mathbb{V}\big(Z_\pi(i,\pm 1)\big) = \frac{\sigma^2}{1-\gamma^2}. \tag{201}$$

This means that the distributions (conditioned on each s, a) have common variance, common slope in expectations, but different y-intercepts in expectations.

We always assumed $\sigma_0^2 = 20$ and tried two different settings, $\gamma = 0.50$ and $\gamma = 0.99$. Denoting the parameter as $\theta = (\beta_L, \beta_R, \beta_1, \sigma^2)$ and candidate space as $\Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}^- \times \mathbb{R}^+$, the best approximation value $\tilde{\theta}$ that minimize the inaccuracy $\bar{\mathcal{E}}(\Upsilon_\theta, \Upsilon_\pi)$ are calculated in Table 5. They are visualized in Figure 3 (true distributions Υ_π on left and best approximations $\Upsilon_{\tilde{\theta}}$ on right).

Table 5: Best approximation values and minimum inaccuracy

Scenario	β_L	β_R	β_0	σ^2	**Minimum $\bar{\mathcal{E}}$-inaccuracy**
$\gamma = 0.50$	126.216	116.614	-4.571	203.099	13.238
$\gamma = 0.99$	610.970	562.782	-23.246	149.866	63.216

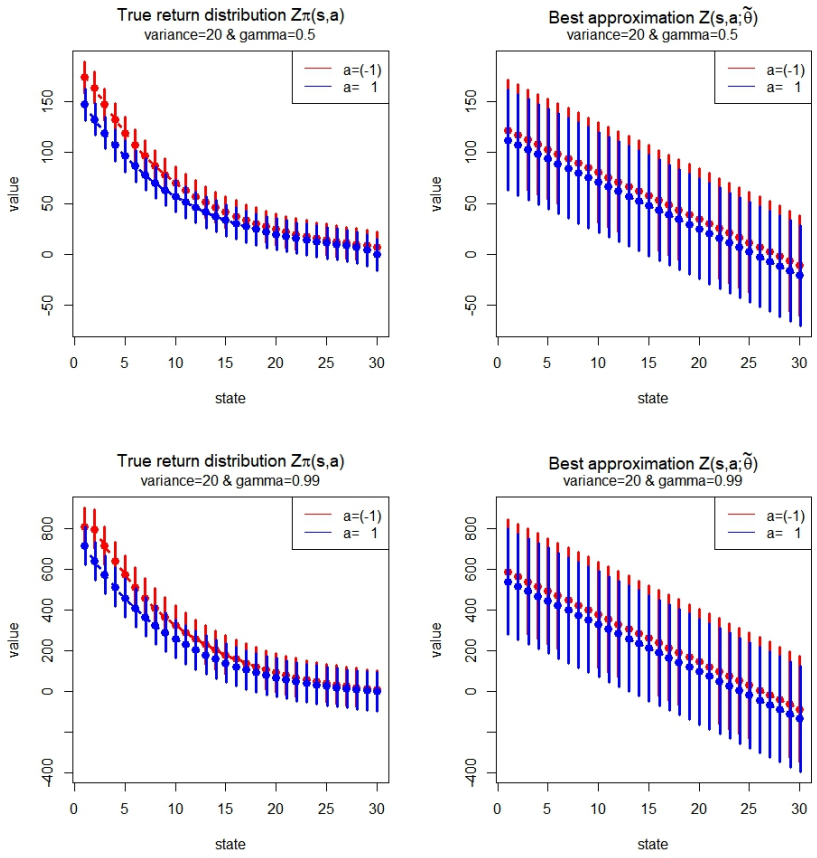

Figure 3: Red and blue represent the distributions of $Z_\pi(s,-1)$ and $Z_\pi(s,1)$, or $Z(s,-1;\tilde{\theta})$ and $Z(s,1;\tilde{\theta})$, respectively. The dots indicate the expectation values and the vertical bars include (Mean$\pm$ 3·SD).

D.3 TUNING PARAMETERS OF EACH METHOD

D.3.1 EBRM

Energy distance (8) is calculated via numerical integration given the densities of the probability measures. Here is the algorithm of choosing the step level m, solely based on the observed data, as mentioned in 4.4. The basic skeleton is based on SLOPE suggested by Su et al. (2020).

Algorithm 3 Lepski's rule of selecting step level m

Require: $1 = m_0 < m_1 < \cdots m_K$
  Input: $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$, J, M, $(m_1, m_2, \cdots, m_K)$
  Output: m_k
  Estimate $\hat{\theta}$ with single-step estimation (13).
  $k \leftarrow K+1$, $O_{K+1} \leftarrow [-\infty, \infty]$.
  while $O_k \neq \varnothing$ **do**
    $k \leftarrow k-1$.
    if $k \neq 0$ **then**
      for $j = 1, \cdots, J$ **do**
        Estimate $\hat{\theta}^{(B)}_{m_k,j}$ with multi-step estimation (23) of step level $m = m_k$.
        Calculate $\hat{e}_{k,j} := \hat{\bar{\mathcal{E}}}(\Upsilon_{\hat{\theta}}, \Upsilon_{\hat{\theta}^{(B)}_{m_k,j}})$.
      end for
      Calculate the sample mean ($\hat{\mu}_k$) and variance ($\hat{s}_k$) of $\hat{e}_{k,j}$ ($j = 1, \cdots, J$).
      Calculate $I_k := [\hat{\mu}_k \pm 1.96 \cdot \hat{s}_k]$.
      $O_k \leftarrow O_{k+1} \cap I_k$.
    else
      $O_k \leftarrow \varnothing$.
    end if
  end while

Throughout multiple simulations in each setting (D.2.2 and D.2.3) for each sample size N (demonstrated in D.4), it is rigorous to pick its own optimal step level m. However, since they do not differ significantly, we picked the step level m via Algorithm 3 based on the first simulated data, and used the same value of m throughout the remaining simulations, in order to save computational burden. We applied the algorithm with $M = N$ and $J = 50$. Obviously, we had to try larger values of $(m_1, \cdots, m_K)$ for non-realizable scenario with $\gamma = 0.99$ than $\gamma = 0.50$. However, to avoid numerical issues in integration caused by extremely small γ^m, we limited the choice of step levels into $m \leq 4$ ($\gamma = 0.50$) and $m \leq 250$ ($\gamma = 0.99$). The corresponding intervals I_k are visualized in Figures 4 and 5, and the selected step level m_* is specified in Table 6.

Table 6: Selected step level m_*

Realizable	$N = 500$	$N = 1000$	$N = 2000$	$N = 5000$	$N = 10000$	$N = 20000$
$\sigma^2 = 20$	1	1	1	1	1	1

Realizable	$N = 2000$	$N = 5000$	$N = 10000$	$N = 20000$	$N = 50000$	$N = 10^5$
$\sigma^2 = 5000$	1	1	1	1	1	1

Non-realizable	$N = 2000$	$N = 3000$	$N = 5000$	$N = 10000$
$\gamma = 0.50$	1	1	1	2
$\gamma = 0.99$	100	160	200	250

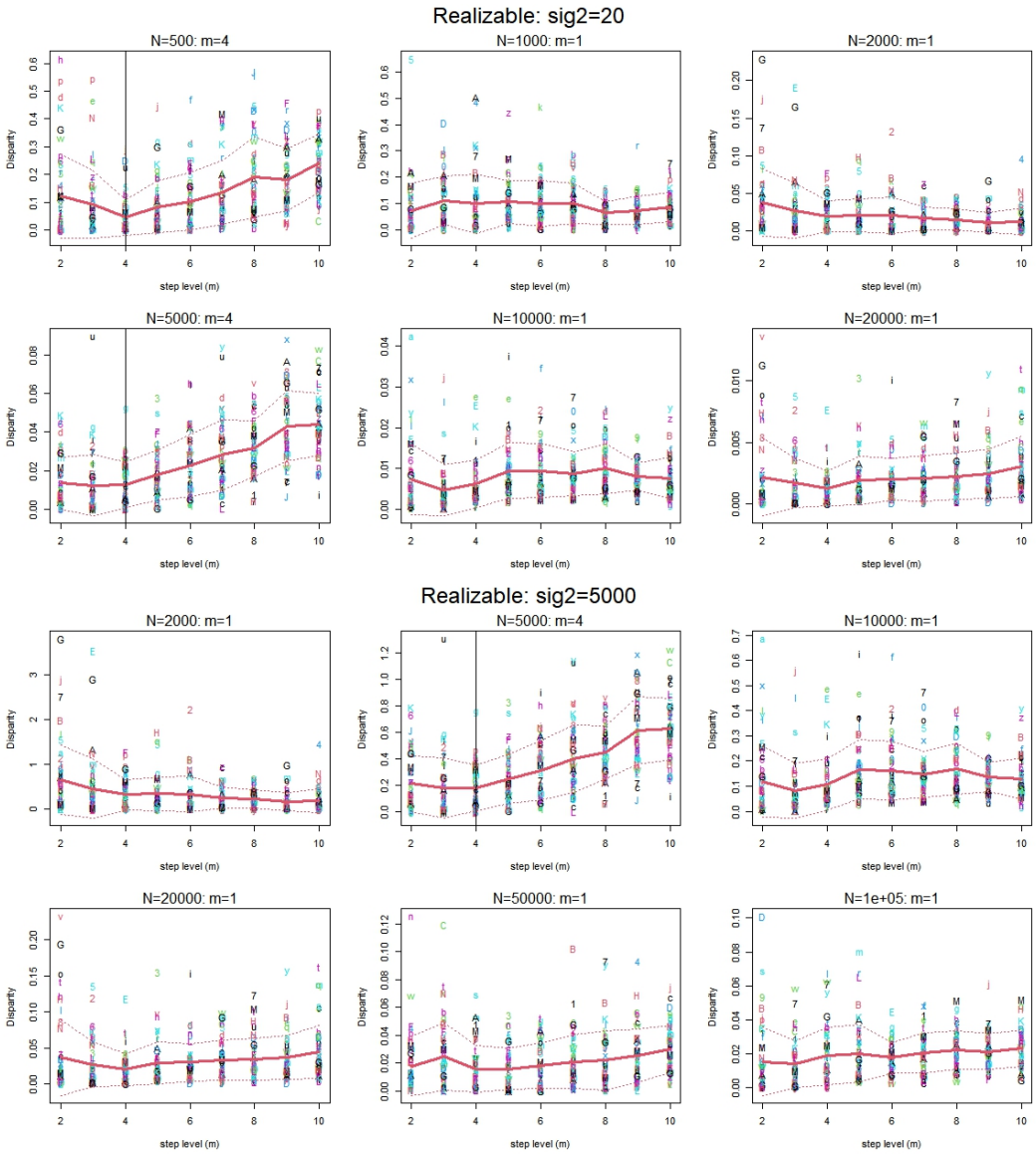

Figure 4: Realizable, $\sigma_0^2 = 20$ (top 6 figures) $\sigma_0^2 = 5000$ (bottom 6 figures)

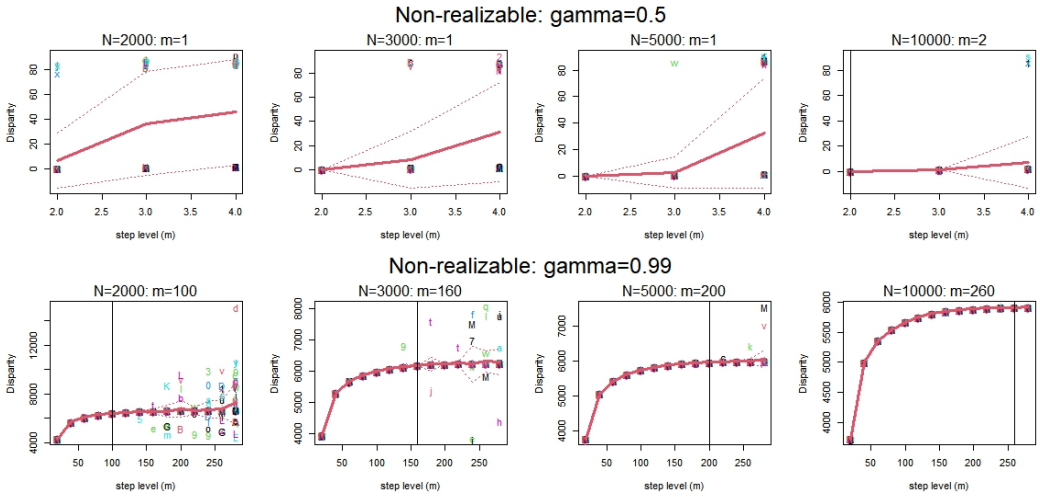

Figure 5: Non-realizable, $\gamma = 0.50$ (top 4 figures) and $\gamma = 0.99$ (bottom 4 figures). To prevent numerical issues, we set an upper limit $m \leq 250$ in $\gamma = 0.99$, so we have $m_* = 250$ for $N = 10000$.

D.3.2 FLE

Wu et al. (2023) did not officially suggest a rule of selecting the number of partitions T and their sizes $|\mathcal{D}_1|, \cdots, |\mathcal{D}_T|$, so we utilized its asymptotic result (Corollary 4.14 of (Wu et al., 2023)) to construct the following heuristic rule based on pre-determined values of N_0, T_0, and $l > 0$,

$$T(N) \overset{let}{=} \log_{\left(\frac{1}{\gamma}\right)^{1-\frac{1}{2l}}} \left\{ \frac{1}{C} \cdot \left(\frac{N}{\log N} \right)^{\frac{1}{2l}} \right\}, \quad \text{where } C_0 > 0 \text{ satisfies } T(N_0) = T_0. \tag{202}$$

Note that larger value of l slows down the increasing speed of $T(N)$. In addition, we prevent the number of partition T from becoming too small, we put a lower bound $\tilde{T}$

$$T^*(N) = \max \left\{ \tilde{T}, \lfloor T(N) \rfloor \right\}.$$

Then each partition has $|\mathcal{D}_t| = \lfloor N/T^*(N) \rfloor$, but may have some remaining observations when N is not divisble by the chosen T. In this case, we included all the remaining observations into the last partition $\mathcal{D}_T$. The tuning parameters are chosen after multiple numerical experiments, and the following choice (Table 7) seemed to work best.

Table 7: Tuning parameters of FLE

Sample size	l	N_0	T_0	$\tilde{T}$
realizable, $\sigma = 50$	10	2000	25	15
realizable, $\sigma = 5000$	10	20000	25	15
non-realizable, $\gamma = 0.50$	0.7	3000	10	10
non-realizable, $\gamma = 0.99$	10	2000	25	15

D.3.3 QRTD

QRTD suggested in Equation (12) of Dabney et al. (2018b) was originally designed for updating value distributions only conditioned on the state $s \in \mathcal{S}$, not the action $a \in \mathcal{A}$. However, we could readily develop it towards the action-value distributions conditioned on $s, a \in \mathcal{S} \times \mathcal{A}$. That is, with $\tau_i \in [0, 1]$ $(1 \leq i \leq N_\tau)$ being cdf values and $\theta_i(s, a)$ being the corresponding quantile value, they model the distribution to be uniform across $\theta_i(s, a)$,

$$Z_\theta(s, a) = \frac{1}{N_\tau} \sum_{i=1}^{N_\tau} \delta_{\theta_i(s,a)}.$$

In analogy to Equation (12) of Dabney et al. (2018b), we update it as follows, whenever a single new observation (s, a, r, s') is collected,

$$\text{Sample } i \sim \text{Unif}\{1, \cdots, N_\tau\} \tag{203}$$
$$\theta_i(s, a) \leftarrow \theta_i(s, a) + \alpha_0 \cdot \{\tau_i - \mathbf{1}(r + \gamma z' < \theta_i(s, a))\} \quad \text{where } z' \sim Z_\theta(s', a'),\ a' \sim \pi(\cdot|s').$$

The number of quantiles is chosen to be 99, that is $\tau_i = i/100$ $(1 \leq i \leq 99)$. As mentioned beneath Equation (12) of Dabney et al. (2018b), we repeat the procedure (203) for multiple times within each iteration, which we let to be the same as $N_\tau = 99$ times in our case. We chose $\alpha_0 = 5$ for $\sigma^2 = 20$ and $\alpha_0 = 2$ for $\sigma^2 = 5000$ of realizable cases, since they worked fine empirically.

D.4 SIMULATION RESULTS

D.4.1 REALIZABLE SCENARIO

The following is the simulation result in the realizable setting (Section D.2.2), where we tried a more variety of sample sizes than Table 2. With the tuning parameters specified in D.3 for each method, we compared EBRM, FLE, and QRTD. EBRM showed the lowest inaccuracy in all cases (Table 8).

Table 8: Mean $\bar{\mathcal{E}}$-inaccuracy (standard deviation in parenthesis) over 100 simulations under realizability ($\gamma = 0.99$) for $\sigma_0^2 = 20$ (top) versus $\sigma_0^2 = 5000$ (bottom). Smallest inaccuracy values are in boldface.

Sample size	500	1000	2000	5000	10000	20000
EBRM	**0.164**	**0.066**	**0.046**	**0.019**	**0.008**	**0.005**
	(0.227)	(0.087)	(0.060)	(0.022)	(0.010)	(0.007)
FLE	17.729	8.802	5.533	2.385	1.220	0.761
	(15.438)	(9.175)	(6.448)	(2.883)	(1.618)	(0.888)
QRTD	149.338	64.259	48.679	46.032	49.402	49.965
	(25.221)	(23.160)	(34.323)	(30.909)	(34.617)	(31.458)
Sample size	2000	5000	10000	20000	50000	100000
EBRM	**0.728**	**0.301**	**0.128**	**0.074**	**0.028**	**0.018**
	(0.920)	(0.354)	(0.167)	(0.105)	(0.034)	(0.022)
FLE	24.603	14.482	6.528	5.062	2.662	1.522
	(25.768)	(16.101)	(7.814)	(6.007)	(3.386)	(1.985)
QRTD	105.274	75.173	70.483	74.398	73.533	77.358
	(11.728)	(21.515)	(33.965)	(52.039)	(70.004)	(62.997)

As was mentioned in Section 5, we also measured the estimation inaccuracy with expectation-extended (6) Wasserstein-1 metric $\overline{\mathbb{W}}_1(\Upsilon_1, \Upsilon_2)$ for fairness. This could be approximated with R package `transport` using randomly generated samples. The superiority of EBRM remains unchanged.

Table 9: Mean $\overline{\mathbb{W}}_1$-inaccuracy (standard deviation in parenthesis) over 100 simulations under realizability ($\gamma = 0.99$) for $\sigma_0^2 = 20$ (top) versus $\sigma_0^2 = 5000$ (bottom). Smallest inaccuracy values are in boldface.

Sample size	500	1000	2000	5000	10000	20000
EBRM	**2.176**	**1.523**	**1.339**	**0.985**	**0.782**	**0.706**
	(1.442)	(0.864)	(0.651)	(0.388)	(0.227)	(0.171)
FLE	22.912	15.755	12.374	8.036	5.694	4.590
	(12.774)	(9.229)	(7.843)	(5.091)	(3.773)	(2.856)
QRTD	105.561	64.290	56.739	54.397	57.145	57.953
	(12.418)	(13.936)	(23.716)	(22.259)	(24.314)	(22.252)
Sample size	2000	5000	10000	20000	50000	100000
EBRM	**21.221**	**15.532**	**12.371**	**11.178**	**9.971**	**9.694**
	(10.337)	(6.117)	(3.595)	(2.717)	(1.143)	(0.802)
FLE	101.232	79.628	53.745	49.426	35.453	27.493
	(58.586)	(46.772)	(33.948)	(29.198)	(21.370)	(16.038)
QRTD	274.405	236.383	223.537	223.399	218.028	224.539
	(11.003)	(22.376)	(38.935)	(63.145)	(82.134)	(76.002)

Lastly, we also used $\mathbb{W}_1$ to compare the marginal distributions $\Upsilon^{marginal}$ which is defined as the mixture of $\{\Upsilon(s,a) : s, a \in \mathcal{S} \times \mathcal{A}\}$ with weights $\{b_\mu(s,a) : s, a \in \mathcal{S} \times \mathcal{A}\}$, as noted in Corollary 4.14 of Wu et al. (2023).

Table 10: Mean $\mathbb{W}_1$-inaccuracy of marginal distributions (standard deviation in parenthesis) over 100 simulations under realizability ($\gamma = 0.99$) for $\sigma_0^2 = 20$ (top) versus $\sigma_0^2 = 5000$ (bottom). Smallest inaccuracy values are in boldface.

Sample size	500	1000	2000	5000	10000	20000
EBRM	**2.052**	**1.406**	**2.052**	**0.843**	**0.629**	**0.508**
	(1.524)	(1.005)	(0.778)	(0.556)	(0.350)	(0.255)
FLE	22.835	15.694	12.328	8.013	5.596	4.591
	(12.810)	(9.272)	(7.859)	(5.113)	(3.786)	(2.945)
QRTD	97.856	56.694	47.738	45.764	49.851	49.877
	(13.235)	(14.905)	(25.251)	(25.656)	(27.463)	(25.945)

Sample size	2000	5000	10000	20000	50000	100000
EBRM	**18.021**	**11.613**	**7.528**	**5.441**	**3.607**	**3.062**
	(12.227)	(8.077)	(5.306)	(4.184)	(2.487)	(1.981)
FLE	94.556	71.430	46.740	44.915	31.774	23.378
	(62.630)	(51.205)	(36.986)	(31.905)	(23.190)	(18.076)
QRTD	247.308	198.257	191.908	195.787	194.908	202.076
	(16.843)	(29.911)	(48.057)	(73.533)	(91.526)	(86.031)

D.4.2 Non-realizable Scenario

Now we tried non-realizable settings with the misspecified model (201) of Section D.2.3, based on tuning parameters determined in Tables 6 and 7. We could see EBRM approached the minimum possible level of Energy Distance (13.327 for $\gamma = 0.50$ and 63.216 for $\gamma = 0.99$) as we increased the sample size N. This forms contrast with FLE that even deteriorated as sample size grows, which we can supposedly attribute to huge violation of completeness that FLE is based upon.

Table 11: Mean $\bar{\mathcal{E}}$-inaccuracy (standard deviation in parenthesis) under non-realizability for $\gamma = 0.50$ (top) VS $\gamma = 0.99$ (bottom). Smallest inaccuracy values are in boldface. Minimum possible $\bar{\mathcal{E}}$-inaccuracy values are 13.237 ($\gamma = 0.50$) and 63.216 ($\gamma = 0.99$).

Sample size	2000	3000	5000	10000
EBRM	**14.323**	**14.306**	**14.299**	**13.544**
	(0.209)	(0.152)	(0.128)	(0.065)
FLE	15.199	15.206	15.171	15.171
	(0.490)	(0.374)	(0.306)	(0.227)

Sample size	2000	3000	5000	10000
EBRM	**124.162**	**96.462**	**82.102**	**70.381**
	(42.117)	(48.178)	(37.470)	(9.503)
FLE	448.837	488.535	625.682	781.287
	(38.256)	(43.141)	(41.130)	(41.192)

Table 12: Mean $\overline{\mathbb{W}}_1$-inaccuracy (standard deviation in parenthesis) under non-realizability for $\gamma = 0.50$ (top) VS $\gamma = 0.99$ (bottom). Smallest inaccuracy values are in boldface.

Sample size	2000	3000	5000	10000
EBRM	19.245	19.232	19.202	17.589
	(0.405)	(0.308)	(0.258)	(0.240)
FLE	**15.036**	**15.049**	**15.047**	**15.037**
	(0.392)	(0.330)	(0.256)	(0.197)

Sample size	2000	3000	5000	10000
EBRM	**168.231**	**108.196**	**91.621**	**82.293**
	(24.009)	(24.756)	(17.601)	(5.490)
FLE	258.802	280.319	350.925	433.074
	(19.127)	(21.680)	(20.697)	(21.534)

Table 13: Mean $\mathbb{W}_1$-inaccuracy of marginal distributions (standard deviation in parenthesis) under non-realizability for $\gamma = 0.50$ (top) VS $\gamma = 0.99$ (bottom). Smallest inaccuracy values are in boldface.

Sample size	2000	3000	5000	10000
EBRM	14.098	14.088	14.087	**13.218**
	(0.227)	(0.167)	(0.142)	(0.093)
FLE	**13.954**	**13.977**	**13.968**	13.970
	(0.359)	(0.292)	(0.234)	(0.184)

Sample size	2000	3000	5000	10000
EBRM	**106.850**	**84.839**	**73.338**	**66.560**
	(39.076)	(29.761)	(17.974)	(5.392)
FLE	258.141	279.687	349.984	432.014
	(19.073)	(21.771)	(20.467)	(21.903)