Appendix A: Additional Theoretical Discussions and Remarks

REMARK 7 (THE POLYAK-RUPPERT AVERAGING). In Theorem 2 we show a last-iterate convergence; we can show the same sample complexity bound when we use the Polyak-Ruppert averaging (Ruppert 1988, Polyak and Juditsky 1992) $(T - T_0)^{-1} \sum_{t=T_0+1}^{T} \boldsymbol{\eta}_t^{\pi}$, where $T_0 \in \mathbb{N}$ is the burn-in period. See Appendix C.10 for more details.

REMARK 8 (REDUCTION TO THE ALL-STATE GENERATIVE MODEL). If we replace the random projection matrix $\boldsymbol{\Lambda}_t$ defined in Section 3.1 with the identity matrix $\boldsymbol{I}$, our conclusions and proofs still remain valid under the formal shorthand $\mu_{\min} = 1$. This corresponds to shifting the sampling protocol to the *all-state generative model*. Specifically, at the t-th iteration, instead of sampling a single state $s_t \sim \mu$, we independently draw transitions for *every* state $s \in \mathcal{S}$: $a_t(s) \sim \pi(\cdot \mid s), r_t(s) \sim \mathcal{P}_R(\cdot \mid s, a_t(s)), s_t'(s) \sim P(\cdot \mid s, a_t(s))$. Consequently, the empirical operator $\boldsymbol{\mathcal{T}}_t^{\pi}$ is fully defined at every coordinate $s \in \mathcal{S}$ using these actual samples:

$$\left[\boldsymbol{\mathcal{T}}_t^{\pi} \boldsymbol{\eta}\right](s) = (b_{r_t(s),\gamma})_{\#}(\eta(s_t'(s))), \quad \forall s \in \mathcal{S},\ \forall \boldsymbol{\eta} \in \mathscr{P}^{\mathcal{S}},$$

and every state is updated synchronously. We note that the shorthand $\mu_{\min} = 1$ used in our sample complexity tables simply means that the random update matrix $\boldsymbol{\Lambda}_t$ is replaced by the deterministic $\boldsymbol{I}$; it does not correspond to a valid probability distribution over multiple states.

REMARK 9 (THE METHODOLOGICAL TRADE-OFF AND NECESSITY OF VARIANCE REDUCTION).

We conjecture that the inability to simultaneously achieve the optimal dependency on both the mixing time t_{mix} and the effective horizon $(1-\gamma)^{-1}$ for vanilla (distributional) TD learning arises from the absence of a sharp vector-valued Bernstein inequality for ergodic Markov chains, rather than an inherent algorithmic deficiency. Without such a tool, analyzing the algorithms over Markovian data forces one to either discard precise variance information, yielding the loose quartic horizon dependency (Li et al. 2024), or use coupling arguments to preserve variance at the cost of a multiplicative t_{mix} penalty (as in our Theorem 3).

From a theoretical perspective, to provably mitigate this multiplicative overhead of t_{mix} while preserving the cubic horizon dependency, one can employ variance reduction techniques. Variance reduction circumvents this by decoupling the problem into a static estimation and a variance-decaying dynamic update. In the outer loop, the algorithm evaluates a reference operator at a fixed reference point. Because this target is static, its estimation error can be bounded using Bernstein inequalities

for independent random elements, where the Markovian dependence merely incurs an additive t_{mix} penalty. In the inner loop, while parameters update dynamically, the variance reduction mechanism ensures that the variance of the stochastic updates is driven to zero as the algorithm converges. By decoupling the estimation, the variance reduction mechanism achieves the optimal joint dependency on both $(1-\gamma)^{-1}$ and t_{mix}. In this way, the analysis no longer relies on vector-valued Bernstein-type inequalities for ergodic Markov chains.

REMARK 10 (ON THE SEPARABILITY AND COMPLETENESS OF SPACES). In Lemma 18, we establish the separability of both $(\mathcal{M}, \|\cdot\|_{\ell_2})$ and $(\mathcal{M}, \|\cdot\|_{W_1})$, which ensures that the random elements encountered in this paper are well-defined and measurable. However, as shown in Lemma 19, $(\mathcal{M}, \|\cdot\|_{W_1})$ is not complete. For the non-asymptotic analysis of vanilla distributional TD learning (Theorem 2), completeness is not necessary because the sequences involved remain within the original space $\mathcal{M}$, meaning they possess finite total variation. For terminological simplicity, we informally refer to $(\mathcal{M}, \|\cdot\|_{\ell_2})$ as a Hilbert space in those contexts. However, we emphasize that for the analysis of the variance-reduced algorithm (Theorem 4), rigorous completion techniques are required. Specifically, defining the fixed point of the perturbed operator $\check{\boldsymbol{\mathcal{T}}}$ in Appendix C.17 necessitates the completion of the metric space to apply the contraction mapping principle. While such theoretical fixed points might reside in the completion and possess infinite variation, they serve strictly as auxiliary analytical tools. Our error bounds are evaluated under the W_1 and ℓ_2 distance, which extend naturally to the completion and remain finite, ensuring the validity of the proof. It is worth noting that for investigating asymptotic behavior or using the operator $(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{T}}^{\pi})^{-1}$, topics that are not considered in this work, one also must rigorously employ completion techniques and extend the operators via the BLT theorem (Hunter and Nachtergaele 2001, Theorem 5.19).

Appendix B: Omitted Results and Proofs in Section 6

B.1. Proof of Theorem 5

Proof. For the sake of clarity and self-containment, we detail the construction of Pinelis's cosh-based positive supermartingale (Pinelis 1994, Theorem 3.2), tailored specifically to the Hilbert space setting.

For any $\lambda > 0$, $t \in [0,1]$ and $j \in [n]$, let $\phi(t) = \phi_{j,\lambda}(t) \coloneqq \mathbb{E}_{j-1}[\cosh(\lambda\|Y_{j-1} + tX_j\|)] = \mathbb{E}_{j-1}[\cosh(\lambda u(t))]$, where $u(t) \coloneqq \|Y_{j-1} + tX_j\|$. We aim to use the Newton-Leibniz formula to establish the relationship between $\phi(1) = \mathbb{E}_{j-1}[\cosh(\lambda\|Y_j\|)]$ and $\phi(0) = \cosh(\lambda\|Y_{j-1}\|)$. This will allow us to construct a positive supermartingale $(B_i)_{i=0}^n$.

Firstly, we calculate the derivative of ϕ. In particular,

$$u'(t) = \frac{\langle Y_{j-1} + tX_j, X_j \rangle}{u(t)},$$

$$\phi'(t) = \lambda \mathbb{E}_{j-1} \left[\sinh(\lambda u(t)) u'(t) \right] = \lambda \mathbb{E}_{j-1} \left[\sinh(\lambda u(t)) \frac{\langle Y_{j-1} + tX_j, X_j \rangle}{u(t)} \right],$$

$$\phi'(0) = \lambda \mathbb{E}_{j-1} \left[\sinh(\lambda u(0)) \frac{\langle Y_{j-1}, X_j \rangle}{u(0)} \right] = \lambda \sinh\left(\lambda \left\|Y_{j-1}\right\|\right) \frac{\langle Y_{j-1}, \mathbb{E}_{j-1}[X_j] \rangle}{\left\|Y_{j-1}\right\|} = 0.$$

Utilizing the Newton-Leibniz formula, we have

$$\phi(1) = \phi(0) + \int_0^1 \phi'(s) \mathrm{d}s = \phi(0) + \int_0^1 \int_0^s \phi''(t) t \mathrm{d}s = \phi(0) + \int_0^1 (1-t)\phi''(t) \mathrm{d}t.$$

We now calculate the second order derivative of ϕ.

$$\begin{aligned}
\phi''(t) &= \lambda \mathbb{E}_{j-1} \left\{ \frac{d}{dt} \left[\sinh(\lambda u(t)) u'(t) \right] \right\} \\
&= \lambda \mathbb{E}_{j-1} \left[\lambda (u'(t))^2 \cosh(\lambda u(t)) + u''(t) \sinh(\lambda u(t)) \right] \\
&\leq \lambda^2 \mathbb{E}_{j-1} \left[\left((u'(t))^2 + u''(t) u(t) \right) \cosh(\lambda u(t)) \right] \\
&= \frac{\lambda^2}{2} \mathbb{E}_{j-1} \left[\left(u^2\right)'' (t) \cosh(\lambda u(t)) \right] \\
&= \lambda^2 \mathbb{E}_{j-1} \left[\left\|X_j\right\|^2 \cosh\left(\lambda \left\|Y_{j-1} + tX_j\right\|\right) \right] \\
&\leq \lambda^2 \cosh\left(\lambda \left\|Y_{j-1}\right\|\right) \mathbb{E}_{j-1} \left[\left\|X_j\right\|^2 \exp\left(\lambda t \left\|X_j\right\|\right) \right],
\end{aligned}$$

where in the first inequality, we used $h(x) = x\cosh(x) - \sinh(x) \geq 0$ $h'(x) = x\sinh(x) \geq 0$ for any $x \geq 0$, and $h(0) = 0$. In the third equality, we used $(u^2)''(t) = 2[(u'(t))^2 + u''(t)u(t)]$. In the fourth equality, we used

$$\left(u^2\right)'' (t) = \frac{d^2}{dt^2} \left\|Y_{j-1} + tX_j\right\|^2 = \frac{d}{dt} \left(2 \left\langle Y_{j-1} + tX_j, X_j \right\rangle\right) = 2 \left\|X_j\right\|^2.$$

In the last inequality, we used

$$\cosh\left(\lambda \left\|Y_{j-1} + tX_j\right\|\right) \leq \cosh\left(\lambda \left\|Y_{j-1}\right\|\right) \exp\left(\lambda t \left\|X_j\right\|\right),$$

which follows from that

$$\exp\left(\lambda \left\|Y_{j-1} + tX_j\right\|\right) \leq \exp\left\{\lambda \left(\left\|Y_{j-1}\right\| + t \left\|X_j\right\|\right)\right\} = \exp\left(\lambda \left\|Y_{j-1}\right\|\right) \exp\left(\lambda t \left\|X_j\right\|\right),$$

$$\exp\left(-\lambda \left\|Y_{j-1} + tX_j\right\|\right) \leq \exp\left\{-\lambda \left(\left\|Y_{j-1}\right\| - t \left\|-X_j\right\|\right)\right\} = \exp\left(-\lambda \left\|Y_{j-1}\right\|\right) \exp\left(\lambda t \left\|X_j\right\|\right).$$

Hence, we can derive the following inequality for all $j \in [n]$

$$
\begin{aligned}
&\mathbb{E}_{j-1}\left[\cosh\left(\lambda\left\|Y_j\right\|\right)\right] = \phi(1) \\
&\quad = \phi(0) + \int_0^1 (1-t)\phi''(t)\mathrm{d}t \\
&\quad \le \cosh\left(\lambda\left\|Y_{j-1}\right\|\right) + \lambda^2 \cosh\left(\lambda\left\|Y_{j-1}\right\|\right)\mathbb{E}_{j-1}\left[\left\|X_j\right\|^2 \int_0^1 (1-t)\exp\left(\lambda t\left\|X_j\right\|\right)\mathrm{d}t\right] \\
&\quad = \cosh\left(\lambda\left\|Y_{j-1}\right\|\right) + \lambda^2 \cosh\left(\lambda\left\|Y_{j-1}\right\|\right)\mathbb{E}_{j-1}\left[\left\|X_j\right\|^2 \frac{\exp\left(\lambda\left\|X_j\right\|\right) - \lambda\left\|X_j\right\| - 1}{\lambda^2\left\|X_j\right\|^2}\right] \\
&\quad = \mathbb{E}_{j-1}\left[\exp\left(\lambda\left\|X_j\right\|\right) - \lambda\left\|X_j\right\|\right]\cosh\left(\lambda\left\|Y_{j-1}\right\|\right) \\
&\quad = \mathbb{E}_{j-1}\left[1 + \sum_{k=0}^{\infty}\frac{1}{(k+2)!}\left(\lambda\left\|X_j\right\|\right)^{k+2}\right]\cosh\left(\lambda\left\|Y_{j-1}\right\|\right) \\
&\quad \le \mathbb{E}_{j-1}\left[1 + \frac{\lambda^2\left\|X_j\right\|^2}{2}\sum_{k=0}^{\infty}\left(\frac{\lambda b}{3}\right)^k\right]\cosh\left(\lambda\left\|Y_{j-1}\right\|\right) \\
&\quad = \left(1 + \frac{\lambda^2\sigma_j^2}{2(1-\lambda b/3)}\right)\cosh\left(\lambda\left\|Y_{j-1}\right\|\right) \\
&\quad \le \exp\left\{\frac{\lambda^2\sigma_j^2}{2(1-\lambda b/3)}\right\}\cosh\left(\lambda\left\|Y_{j-1}\right\|\right),
\end{aligned}
\tag{7}
$$

which holds for any $\lambda \in (0, \frac{3}{b})$, where we used Taylor expansion $e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$ for $x \in \mathbb{R}$, and $\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$ for $x \in (-1,1)$. And in the second inequality, we used $(k+2)! \ge 2(3^k)$ and $\|X_j\| \le b$.

Let $B_0 := 1$, $B_i := \exp\{-\frac{\lambda^2 W_i}{2(1-\lambda b/3)}\}\cosh(\lambda\|Y_i\|)$. Then

$$
\begin{aligned}
\mathbb{E}_{i-1}\left[B_i\right] &= \exp\left\{-\frac{\lambda^2 W_{i-1}}{2(1-\lambda b/3)}\right\}\exp\left\{-\frac{\lambda^2\sigma_i^2}{2(1-\lambda b/3)}\right\}\mathbb{E}_{i-1}\left[\cosh\left(\lambda\left\|Y_i\right\|\right)\right] \\
&\le \exp\left\{-\frac{\lambda^2 W_{i-1}}{2(1-\lambda b/3)}\right\}\cosh\left(\lambda\left\|Y_{i-1}\right\|\right) \\
&= B_{i-1},
\end{aligned}
$$

showing that $(B_i)_{i=0}^n$ is positive supermartingale. By the optional stopping theorem (Durrett 2019, Theorem 4.8.4), for any stopping time τ, we have $\mathbb{E}[B_\tau] \le \mathbb{E}[B_0] = 1$.

Let $\tau := \inf\{k \in [n] : \|Y_k\| \ge \varepsilon\}$ be a stopping time, and $\inf\emptyset := \infty$. Define an event

$$
A := \left\{\exists k \in [n], \text{s.t. } \|Y_k\| \ge \varepsilon \text{ and } W_k \le \sigma^2\right\}.
$$

Then on A, we have $\tau < \infty$, $\|Y_\tau\| \geq \varepsilon$ and $W_\tau \leq \sigma^2$, noting that W_k is non-decreasing with k. Our goal is to provide an upper bound for $\mathbb{P}(A)$. Note that

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{E}\left[\sqrt{B_\tau}\frac{1}{\sqrt{B_\tau}}\mathbb{1}(A)\right] \\
&\leq \sqrt{\mathbb{E}\left[B_\tau\right]\mathbb{E}\left[\frac{1}{B_\tau}\mathbb{1}(A)\right]} \\
&\leq \sqrt{\mathbb{E}\left[\frac{\exp\left\{\frac{\lambda^2 W_\tau}{2(1-\lambda b/3)}\right\}}{\cosh\left(\lambda\left\|Y_\tau\right\|\right)}\mathbb{1}(A)\right]} \\
&\leq \sqrt{\mathbb{E}\left[\frac{\exp\left\{\frac{\lambda^2 \sigma^2}{2(1-\lambda b/3)}\right\}}{\cosh\left(\lambda\varepsilon\right)}\mathbb{1}(A)\right]} \\
&\leq \sqrt{2\exp\left\{-\lambda\varepsilon + \frac{\lambda^2\sigma^2}{2(1-\lambda b/3)}\right\}\mathbb{P}(A)},
\end{aligned}$$

where in the first inequality, we used Cauchy-Schwartz inequality. In the second inequality, we used $\mathbb{E}[B_\tau] \leq 1$. In the third inequality, we used $\|Y_\tau\| \geq \varepsilon$ and $W_\tau \leq \sigma^2$ on A, and $\cosh(x)$ is increasing when $x \geq 0$. In the last inequality, we used $\cosh(x) \geq \frac{1}{2}e^x$.

Hence for any $\lambda \in (0, \frac{3}{b})$,

$$\mathbb{P}(A) \leq 2\exp\left\{-\lambda\varepsilon + \frac{\lambda^2\sigma^2}{2\left(1-\lambda b/3\right)}\right\}.$$

We choose $\lambda^\star = \frac{\varepsilon}{\sigma^2+\varepsilon b/3} \in (0, \frac{3}{b})$. Hence,

$$\begin{aligned}
\mathbb{P}(A) &\leq 2\exp\left\{-\lambda^\star\varepsilon + \frac{\left(\lambda^\star\right)^2\sigma^2}{2\left(1-\lambda^\star b/3\right)}\right\} \\
&= 2\exp\left\{-\frac{\varepsilon^2}{\sigma^2+\varepsilon b/3} + \frac{\sigma^2}{2\left(1-\frac{\varepsilon b/3}{\sigma^2+\varepsilon b/3}\right)}\frac{\varepsilon^2}{\left(\sigma^2+\varepsilon b/3\right)^2}\right\} \\
&= 2\exp\left\{-\frac{\varepsilon^2/2}{\sigma^2+\varepsilon b/3}\right\},
\end{aligned}$$

which is the desired conclusion.

B.2. Proof of Theorem 6

Proof. The proof extends the approach of Theorem 6 in Li et al. (2024) to the setting of Hilbert space-valued martingales.

According to Theorem 5, for any $\varepsilon, \tilde{\sigma} > 0$, we have

$$\mathbb{P}\big(\exists k \in [n],\ \|Y_k\| \geq \varepsilon \text{ and } W_k \leq \tilde{\sigma}^2\big) \leq 2\exp\left\{-\frac{\varepsilon^2/2}{\tilde{\sigma}^2 + b\varepsilon/3}\right\}.$$

We can check that when $\varepsilon = \sqrt{4\tilde{\sigma}^2 \log\frac{2}{\delta}} + \frac{4}{3}b\log\frac{2}{\delta}$, the upper bound on RHS is less than δ. Hence,

$$\mathbb{P}\left(\exists k \in [n],\ \|Y_k\| \geq \sqrt{4\tilde{\sigma}^2 \log\frac{2}{\delta}} + \frac{4}{3}b\log\frac{2}{\delta} \text{ and } W_k \leq \tilde{\sigma}^2\right) \leq \delta. \tag{8}$$

For each $k \in [n]$, define the events

$$\begin{aligned}
\mathcal{H}_H^{(k)} &:= \left\{\|Y_k\| \geq \sqrt{8\max\left\{W_k, \frac{\sigma^2}{2^H}\right\}\log\frac{2H}{\delta}} + \frac{4}{3}b\log\frac{2H}{\delta}\right\},\\
\mathcal{B}_{H,H}^{(k)} &:= \left\{\|Y_k\| \geq \sqrt{4\frac{\sigma^2}{2^{H-1}}\log\frac{2H}{\delta}} + \frac{4}{3}b\log\frac{2H}{\delta} \text{ and } W_k \leq \frac{\sigma^2}{2^{H-1}}\right\},\\
\mathcal{B}_{h,H}^{(k)} &:= \left\{\|Y_k\| \geq \sqrt{4\frac{\sigma^2}{2^{h-1}}\log\frac{2H}{\delta}} + \frac{4}{3}b\log\frac{2H}{\delta} \text{ and } \frac{\sigma^2}{2^h} \leq W_k \leq \frac{\sigma^2}{2^{h-1}}\right\}, \qquad 1 \leq h \leq H-1.
\end{aligned}$$

By the definition, we only need to show $\mathbb{P}(\bigcup_{k\in[n]}\mathcal{H}_H^{(k)}) \leq \delta$. Since $W_k \leq W_n \leq \sigma^2$ almost surely, we can find that $\mathcal{H}_H^{(k)} \subseteq \bigcup_{h\in[H]}\mathcal{B}_{h,H}^{(k)}$ (we will justify this later). Then $\bigcup_{k\in[n]}\mathcal{H}_H^{(k)} \subseteq \bigcup_{h\in[H]}\bigcup_{k\in[n]}\mathcal{B}_{h,H}^{(k)}$. By the inequality (8) with $\tilde{\sigma}^2 = \frac{\sigma^2}{2^{h-1}}$ and setting δ as $\frac{\delta}{H}$, we have $\mathbb{P}(\bigcup_{k\in[n]}\mathcal{B}_{h,H}^{(k)}) \leq \frac{\delta}{H}$ for all $h \in [H]$. By the union bound, we can arrive at the conclusion:

$$\mathbb{P}\left(\bigcup_{k\in[n]}\mathcal{H}_H^{(k)}\right) \leq \sum_{h=1}^{H}\mathbb{P}\left(\bigcup_{k\in[n]}\mathcal{B}_{h,H}^{(k)}\right) \leq \delta.$$

To justify $\mathcal{H}_H^{(k)} \subseteq \bigcup_{h\in[H]}\mathcal{B}_{h,H}^{(k)}$, we can consider the decomposition $\mathcal{H}_H^{(k)} = \bigcup_{h\in[H]}(\mathcal{H}_H^{(k)} \cap \mathcal{C}_{h,H}^{(k)})$, where

$$\mathcal{C}_{H,H}^{(k)} := \left\{W_k \leq \frac{\sigma^2}{2^{H-1}}\right\}, \qquad \mathcal{C}_{h,H}^{(k)} := \left\{\frac{\sigma^2}{2^h} \leq W_k \leq \frac{\sigma^2}{2^{h-1}}\right\}, \quad 1 \leq h \leq H-1.$$

The decomposition holds because $W_k \leq W_n \leq \sigma^2$ almost surely. We only need to show that $\mathcal{H}_H^{(k)} \cap \mathcal{C}_{h,H}^{(k)} \subseteq \mathcal{B}_{h,H}^{(k)}$ for each $h \in [H]$. On the event $\mathcal{H}_H^{(k)} \cap \mathcal{C}_{h,H}^{(k)}$, we have

$$\|Y_k\| \geq \sqrt{8\max\left\{W_k, \frac{\sigma^2}{2^H}\right\}\log\frac{2H}{\delta}} + \frac{4}{3}b\log\frac{2H}{\delta} \geq \sqrt{4\frac{\sigma^2}{2^{h-1}}\log\frac{2H}{\delta}} + \frac{4}{3}b\log\frac{2H}{\delta},$$

hence $\mathcal{H}_H^{(k)} \cap \mathcal{C}_{h,H}^{(k)} \subseteq \mathcal{B}_{h,H}^{(k)}$.

Appendix C: Omitted Results and Proofs in Section 7

C.1. Proof of Theorem 1

Proof. For any $\boldsymbol{\eta} \in \mathscr{P}^S$, the metric of interest is the supremum p-Wasserstein metric $\bar{W}_p$. Evidently, considering Markov reward processes (MRPs) is sufficient. For $p = 1$, the W_1 lower bound could alternatively be derived from a value-function lower bound under fixed-distribution sampling, since W_1 upper-bounds the absolute difference of means. We instead work directly with return distributions to establish lower bounds under general W_p and ℓ_p metrics.

By abusing notation, let $\mathscr{M}(S)$ represent the set of all MRPs with state space size $|\mathcal{S}| = S$ and $\boldsymbol{\eta}_M \in \mathscr{P}^S$ the return distribution of any MRP $M \in \mathscr{M}(S)$. Our task then reduces to providing lower bounds for $\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M\in\mathscr{M}(S)} \mathbb{E}[\bar{W}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)]$ and $\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M\in\mathscr{M}(S)} \mathbb{E}[\bar{\ell}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)]$.

The lower bounds we aim to establish consists of two components: the dependence on the state space size S and the dependence on $\mu_{\min}$. We first consider the state-space-dependent lower bound, which is derived using Fano's method.

We consider the two-state MRP building block $M(q)$ with deterministic rewards and parameter $q \in [0, 1]$, as illustrated in Figure 2(a) in Pananjady and Wainwright (2020). The transition matrix $\boldsymbol{P}_0$ and the reward vector $\boldsymbol{r}_0$ are defined as:

$$\boldsymbol{P}_{M(q)} = \begin{bmatrix} q & 1-q \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{r}_{M(q)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

where the first state is a transient state and the second state is an absorbing state. This building block serves as the cornerstone for proving Theorem 2(b) in Pananjady and Wainwright (2020). While their theorem focuses on value function estimation and assumes the ability to simultaneously visit all states in a single iteration, our work addresses return distribution estimation under a fixed-distribution setting where samples are drawn from a specific distribution μ.

p-Wasserstein metric. Recall the following equivalent definition of W_p distance in terms of the inverse cumulative distribution function (CDF):

$$W_p(\nu_1, \nu_2) = \left(\int_0^1 |F_{\nu_1}^{-1}(u) - F_{\nu_2}^{-1}(u)|^p \mathrm{d}u \right)^{1/p}, \tag{9}$$

for any $\nu_1, \nu_2 \in \mathscr{P}$, where the inverse CDF (or quantile function) F_ν^{-1} is formally defined as the left-continuous generalized inverse of F_ν for any $\nu \in \mathscr{P}$:

$$F_\nu^{-1}(u) = \inf\{x \in \mathbb{R} : F_\nu(x) \geq u\}, \quad \text{for } u \in (0, 1]. \tag{10}$$

Therefore, it suffices to calculate the inverse CDF of the return distribution for the two-state MRP building block $M(q)$. We denote the return distribution vector of $M(q)$ by $\boldsymbol{\eta}_{M(q)}$.

Clearly, the return distribution in the absorbing state is δ_0. Thus, it suffices to focus on the return distribution in the transient state. Let G be the random return starting from the transient state. The process stays in the transient state for exactly k steps if and only if it transitions to the absorbing state at the $(k+1)$-th step. The probability of this trajectory is $q^{k-1}(1-q)$. The corresponding cumulative return is

$$v_k = \sum_{t=0}^{k-1} \gamma^t = \frac{1-\gamma^k}{1-\gamma}.$$

Thus, the return distribution in the transient state is a discrete distribution supported on the countable set $\{v_k\}_{k=1}^{\infty}$ with the probability mass function (PMF)

$$\mathbb{P}(G = v_k) = q^{k-1}(1-q).$$

We denote the corresponding cumulative distribution function (CDF) by $F_{M(q)}$. For any $x \in [v_k, v_{k+1})$

$$\begin{aligned} F_{M(q)}(x) &= \sum_{j=1}^{k} \mathbb{P}(G = v_j) \\ &= \sum_{j=1}^{k} q^{j-1}(1-q) \\ &= (1-q)\frac{1-q^k}{1-q} \\ &= 1-q^k. \end{aligned}$$

For $x < v_1$, we trivially have $F_{M(q)}(x) = 0$.

To determine the explicit form of $F^{-1}_{M(q)}(u)$, we invert the step function $F_{M(q)}(x)$. Notice that $F_{M(q)}(x)$ reaches the value $1-q^k$ at exactly $x = v_k$, and stays constant on $[v_k, v_{k+1})$. Therefore, for any probability $u \in (0,1]$, if $u \in (1-q^{k-1}, 1-q^k]$, the smallest x that satisfies $F_{M(q)}(x) \geq u$ is exactly v_k. Thus, the inverse cdf can be written as a sum of weighted indicator functions: for any $u \in (0,1]$

$$F^{-1}_{M(q)}(u) = \sum_{k=1}^{\infty} v_k \mathbb{I}_{(1-q^{k-1}, 1-q^k]}(u) = \sum_{j=1}^{\infty} (v_j - v_{j-1}) \mathbb{I}_{(1-q^{j-1}, 1]}(u) = v_1 + \sum_{k=1}^{\infty} \gamma^k \mathbb{I}_{(1-q^k, 1]}(u),$$

where $v_0 \coloneqq 0$, and $\mathbb{I}_A(u)$ is the indicator function defined as $\mathbb{I}_A(u) = 1$ if $u \in A$ and $\mathbb{I}_A(u) = 0$ otherwise. Thus, for any $0 \leq p_2 < p_1 \leq 1$, we have

$$F^{-1}_{M(p_1)}(u) - F^{-1}_{M(p_2)}(u) = \sum_{k=1}^{\infty} \gamma^k \left(\mathbb{I}_{(1-p_1^k, 1]}(u) - \mathbb{I}_{(1-p_2^k, 1]}(u) \right) = \sum_{k=1}^{\infty} \gamma^k \mathbb{I}_{(1-p_1^k, 1-p_2^k]}(u).$$

Because the function $x \mapsto x^p$ is superadditive for $p \geq 1$ and $x \geq 0$, it holds that:

$$\begin{aligned}
\bar{W}_p^p(\boldsymbol{\eta}_{M(p_1)}, \boldsymbol{\eta}_{M(p_2)}) &= \int_0^1 \left(\sum_{k=1}^{\infty} \gamma^k \mathbb{I}_{(1-p_1^k, 1-p_2^k]}(u) \right)^p \mathrm{d}u \\
&\geq \int_0^1 \sum_{k=1}^{\infty} (\gamma^k)^p \mathbb{I}_{(1-p_1^k, 1-p_2^k]}(u) \mathrm{d}u \\
&= \sum_{k=1}^{\infty} \gamma^{pk} (p_1^k - p_2^k) \\
&= \frac{p_1 \gamma^p}{1 - p_1 \gamma^p} - \frac{p_2 \gamma^p}{1 - p_2 \gamma^p} \\
&= \frac{\gamma^p (p_1 - p_2)}{(1 - \gamma^p p_1)(1 - \gamma^p p_2)}.
\end{aligned}$$

Taking the $1/p$-th root yields

$$\bar{W}_p(\boldsymbol{\eta}_{M(p_1)}, \boldsymbol{\eta}_{M(p_2)}) \geq \frac{\gamma (p_1 - p_2)^{1/p}}{[(1 - \gamma^p p_1)(1 - \gamma^p p_2)]^{1/p}}. \tag{11}$$

We construct a family of $S/2$ master MRPs $\{M_1, \ldots, M_{S/2}\}$ by replicating this two-state block. (Without loss of generality, we assume that S is even.) Specifically, these MRPs are structured as follows: given two parameters $\frac{1}{2} \leq p_2 < p_1 \leq 1$, the state space of each M_j is partitioned into $S/2$ disjoint pairs of states (two-state building blocks), with transitions occurring only within each pair. The k-th building block is parameterized by p_1 if $k \neq j$, and by p_2 otherwise. According to our sampling mechanism, for the j-th MRP M_j, we observe pairs $(s_0, s_1) \sim Q_j$ with $s_0 \sim \mu$ and $s_1 \mid s_0 \sim P_j(\cdot \mid s_0)$, where P_j represents the transition dynamics associated with M_j. And we denote Q_j as the joint distribution of (s_0, s_1) and $\boldsymbol{\eta}_j$ as the return distribution vector of M_j.

By Fano's method (Wainwright 2019a, Proposition 15.12), it holds that

$$\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M \in \mathscr{M}(S)} \mathbb{E}[\bar{W}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)] \geq \frac{1}{2} \min_{j \neq k} \bar{W}_p(\boldsymbol{\eta}_j, \boldsymbol{\eta}_k) \left(1 - \frac{\frac{T}{(S/2)^2} \sum_{j=1}^{S/2} \sum_{k=1}^{S/2} D_{\mathrm{KL}}(Q_j \| Q_k) + \log 2}{\log(S/2)} \right).$$

By the construction of MRPs and Eqn. (11), it holds that

$$\min_{j \neq k} \bar{W}_p(\boldsymbol{\eta}_j, \boldsymbol{\eta}_k) \geq \frac{\gamma (p_1 - p_2)^{1/p}}{[(1 - \gamma^p p_1)(1 - \gamma^p p_2)]^{1/p}}.$$

It now remains to bound the KL divergence term. By the chain rule for KL divergence, for $j \neq k$, we have

$$\begin{aligned}
D_{\mathrm{KL}}(Q_j \| Q_k) =& D_{\mathrm{KL}}(\mu \| \mu) + \mathbb{E}_{s_0 \sim \mu} \left[D_{\mathrm{KL}} \left(P_j(\cdot | s_0) \| P_k(\cdot | s_0) \right) \right] \\
=& \mu(2j-1) D_{\mathrm{KL}}(\mathrm{Ber}(p_2) \| \mathrm{Ber}(p_1)) + \mu(2k-1) D_{\mathrm{KL}}(\mathrm{Ber}(p_1) \| \mathrm{Ber}(p_2)) \\
\leq& (\mu(2j-1) + \mu(2k-1)) \frac{(p_1 - p_2)^2}{p_1(1-p_1)},
\end{aligned}$$

where the last inequality follows from the standard upper bound of the KL divergence between two Bernoulli distributions (Pananjady and Wainwright 2020, Lemma 4). Hence,

$$\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M\in\mathscr{M}(S)} \mathbb{E}[\bar{W}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)] \geq \frac{\gamma(p_1-p_2)^{1/p}}{2[(1-\gamma^p p_1)(1-\gamma^p p_2)]^{1/p}} \left(1 - \frac{\frac{4T}{S}\frac{(p_1-p_2)^2}{p_1(1-p_1)} + \log 2}{\log(S/2)}\right).$$

We set

$$p_1 = \frac{4\gamma-1}{3\gamma}, \quad p_2 = p_1 - \frac{1}{8}\sqrt{\frac{p_1(1-p_1)S\log(S/2)}{T}},$$

which ensures that

$$p_1 \in [\frac{2}{3}, 1), \quad \frac{1}{3}(1-\gamma) \leq p_1(1-p_1) \leq \frac{4}{9}(1-\gamma),$$

and

$$1 - \frac{\frac{4T}{S}\frac{(p_1-p_2)^2}{p_1(1-p_1)} + \log 2}{\log(S/2)} = \frac{15}{16} - \frac{\log 2}{\log(S/2)}.$$

Note that, for any $p \geq 1$, Bernoulli's inequality gives $1-\gamma^p \leq p(1-\gamma)$. Hence, assuming $\gamma \geq 1/2$,

$$1-\gamma^p p_1 = 1-\gamma^p + \gamma^p(1-p_1) \leq p(1-\gamma) + (1-p_1) = p(1-\gamma) + \frac{1-\gamma}{3\gamma} \leq (p+2/3)(1-\gamma).$$

Similarly,

$$1-\gamma^p p_2 = 1-\gamma^p p_1 + \gamma^p(p_1-p_2) \leq (p+2/3)(1-\gamma) + (p_1-p_2).$$

From the bounds below, taking $T \geq \frac{3}{256}\frac{S\log(S/2)}{1-\gamma}$ ensures that $p_1 - p_2 \leq \frac{4\sqrt{3}}{9}(1-\gamma)$ and $p_2 \geq \frac{1}{2}$. Thus,

$$1-\gamma^p p_2 \leq (p+2)(1-\gamma).$$

To summarize, we have

$$\begin{aligned}\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M\in\mathscr{M}(S)} \mathbb{E}[\bar{W}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)] &\geq \frac{\gamma}{2[(p+2)^2(1-\gamma)^2]^{1/p}} \left(\frac{1}{8}\sqrt{\frac{(1-\gamma)S\log(S/2)}{3T}}\right)^{\frac{1}{p}} \left(\frac{15}{16} - \frac{\log 2}{\log(S/2)}\right) \\ &\geq \frac{1}{4(p+2)^{2/p}8^{1/p}} \left(\frac{1}{3}\right)^{\frac{1}{2p}} \left(\frac{15}{16} - \frac{\log 2}{\log(S/2)}\right) \left(\frac{S\log(S/2)}{(1-\gamma)^3 T}\right)^{\frac{1}{2p}},\end{aligned}$$

by using $\gamma \geq 1/2$ and taking $S \geq 5$, we ensure that $\frac{15}{16} - \frac{\log 2}{\log(S/2)}$ is positive.

Next, we address the lower bound with respect to $\mu_{\min}$ by employing Le Cam's two-point method.

We modify the two-state MRP building block $M(q)$ to construct two MRPs, M_1 and M_2, of state space size S. Let $s^\star = \arg\min_{s\in\mathcal{S}} \mu(s)$ and choose an arbitrary state $s^\circ \in \mathcal{S} \setminus \{s^\star\}$. For both M_1 and M_2, all states other than $s^\star$ are absorbing states with zero reward. At state $s^\star$, the transition dynamics are defined as follows: the process stays at $s^\star$ with probability p_1 (for M_1) or p_2 (for M_2), and moves

to s° with probability $1-p_1$ or $1-p_2$, respectively. Furthermore, a reward of 1 is associated with state $s^\star$.

By Le Cam's two-point method (Wainwright 2019a, Eqn. (15.14)) and Pinsker's inequality (Wainwright 2019a, Lemma 15.2), it holds that

$$\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M\in\mathscr{M}(S)} \mathbb{E}[\bar{W}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)] \geq \frac{1}{2}\bar{W}_p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\left(1-\sqrt{\frac{T}{2}D_{\mathrm{KL}}(Q_1\|Q_2)}\right).$$

As before,

$$\bar{W}_p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \geq \frac{\gamma(p_1-p_2)^{1/p}}{[(1-\gamma^p p_1)(1-\gamma^p p_2)]^{1/p}},$$

and

$$\begin{aligned} D_{\mathrm{KL}}(Q_1\|Q_2) =& D_{\mathrm{KL}}(\mu\|\mu) + \mathbb{E}_{s_0\sim\mu}\left[D_{\mathrm{KL}}\left(P_1(\cdot|s_0)\|P_2(\cdot|s_0)\right)\right] \\ =& \mu_{\min} D_{\mathrm{KL}}(\mathrm{Ber}(p_1)\|\mathrm{Ber}(p_2)) \\ \leq& \mu_{\min}\frac{(p_1-p_2)^2}{p_1(1-p_1)}. \end{aligned}$$

Hence,

$$\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M\in\mathscr{M}(S)} \mathbb{E}[\bar{W}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)] \geq \frac{\gamma(p_1-p_2)^{1/p}}{2[(1-\gamma^p p_1)(1-\gamma^p p_2)]^{1/p}}\left(1-\sqrt{\frac{T\mu_{\min}}{2}\frac{(p_1-p_2)^2}{p_1(1-p_1)}}\right).$$

We set

$$p_1 = \frac{4\gamma-1}{3\gamma},$$

as before, and

$$p_2 = p_1 - \frac{1}{4}\sqrt{\frac{p_1(1-p_1)}{T\mu_{\min}}},$$

which ensures that

$$\begin{aligned} 1-\gamma p_2 =& (1-\gamma p_1) + \gamma(p_1-p_2) \\ =& \frac{4}{3}(1-\gamma) + \frac{\gamma}{4}\sqrt{\frac{p_1(1-p_1)}{T\mu_{\min}}} \\ \leq& \left(\frac{4}{3} + \frac{1}{4}\sqrt{\frac{1}{3T(1-\gamma)\mu_{\min}}}\right)(1-\gamma) \\ \leq& 2(1-\gamma), \end{aligned}$$

if we take

$$T \geq \frac{3}{64}\frac{1}{\mu_{\min}(1-\gamma)},$$

which also ensures that $p_1 - p_2 \le \frac{4\sqrt{3}}{9}(1-\gamma)$, $1-\gamma^p p_i \le (p+2)(1-\gamma)$ and $p_2 \ge \frac{1}{2}$ as derived before. To summarize, we have

$$\begin{aligned}\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M\in\mathscr{M}(S)} \mathbb{E}[\bar{W}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)] &\ge \frac{\gamma}{2[(p+2)^2(1-\gamma)^2]^{1/p}} \left(\frac{1}{4}\sqrt{\frac{1-\gamma}{3T\mu_{\min}}}\right)^{\frac{1}{p}} \left(1-\frac{1}{\sqrt{32}}\right) \\ &\ge \frac{1}{4(p+2)^{2/p}4^{1/p}} \left(\frac{1}{3}\right)^{\frac{1}{2p}} \left(1-\frac{1}{\sqrt{32}}\right) \left(\frac{1}{\mu_{\min}(1-\gamma)^3 T}\right)^{\frac{1}{2p}},\end{aligned}$$

which is the desired conclusion.

ℓ_p metric. We employ the same hard instances and Le Cam/Fano's method arguments established in the proof for the W_p distance. Specifically, we consider the identical two-state MRP building block $M(q)$ and its replicated master MRPs.

Recall that

$$F_{M(q)}(x) = \sum_{k=1}^{\infty} (1-q^k)\mathbb{I}_{[v_k, v_{k+1})}(x).$$

Thus, for any $0 \le p_2 < p_1 \le 1$, we have

$$F_{M(p_2)}(x) - F_{M(p_1)}(x) = \sum_{k=1}^{\infty} \left[(1-p_2^k) - (1-p_1^k)\right] \mathbb{I}_{[v_k, v_{k+1})}(x) = \sum_{k=1}^{\infty} \left(p_1^k - p_2^k\right) \mathbb{I}_{[v_k, v_{k+1})}(x).$$

Hence,

$$\bar{\ell}_p^p(\boldsymbol{\eta}_{M(p_1)}, \boldsymbol{\eta}_{M(p_2)}) = \int_{\mathbb{R}} \left|F_{M(p_1)}(x) - F_{M(p_2)}(x)\right|^p dx = \sum_{k=1}^{\infty} \gamma^k \left(p_1^k - p_2^k\right)^p.$$

By the Mean Value Theorem, $p_1^k - p_2^k \ge k p_2^{k-1}(p_1 - p_2)$. Therefore,

$$\bar{\ell}_p^p(\boldsymbol{\eta}_{M(p_1)}, \boldsymbol{\eta}_{M(p_2)}) \ge (p_1-p_2)^p \sum_{k=1}^{\infty} k^p \gamma^k p_2^{p(k-1)} = \frac{(p_1-p_2)^p}{p_2^p} \sum_{k=1}^{\infty} k^p (\gamma p_2^p)^k. \tag{12}$$

Let $z = \gamma p_2^p \in (0,1)$. To derive a lower bound for the series $\sum_{k=1}^{\infty} k^p z^k$, we use the probabilistic method. Specifically, we consider a coupling between a geometric distribution and an exponential distribution.

Let $Y \sim \mathrm{Exp}(\lambda)$ be an exponentially distributed random variable with rate $\lambda = -\log z > 0$. We define a discrete random variable $X = \lfloor Y \rfloor + 1$. The probability mass function of X exactly matches a geometric distribution with success probability $1-z$:

$$\mathbb{P}(X=k) = \mathbb{P}(k-1 \le Y < k) = \int_{k-1}^{k} \lambda e^{-\lambda y} \mathrm{d}y = e^{-\lambda(k-1)}(1-e^{-\lambda}) = z^{k-1}(1-z).$$

Thus, the series of interest can be rewritten exactly as the p-th moment of X:

$$\sum_{k=1}^{\infty} k^p z^k = \frac{z}{1-z} \sum_{k=1}^{\infty} k^p (1-z) z^{k-1} = \frac{z}{1-z} \mathbb{E}[X^p].$$

Crucially, by construction, we have $X = \lfloor Y \rfloor + 1 > Y$ almost surely. Since the function $x \mapsto x^p$ is strictly increasing for $x > 0$ and $p \geq 1$, it follows that $\mathbb{E}[X^p] > \mathbb{E}[Y^p]$. The p-th moment of the exponential random variable Y is precisely given by the Gamma function:

$$\mathbb{E}[Y^p] = \int_0^{\infty} y^p \lambda e^{-\lambda y} \mathrm{d}y = \frac{1}{\lambda^p} \int_0^{\infty} t^p e^{-t} \mathrm{d}t = \frac{\Gamma(p+1)}{(-\log z)^p}.$$

Combining these, and using the standard logarithmic inequality $-\log z \leq \frac{1-z}{z}$, we obtain:

$$\mathbb{E}[X^p] > \mathbb{E}[Y^p] \geq \Gamma(p+1) \left(\frac{z}{1-z} \right)^p.$$

Therefore, the series is strictly lower bounded by:

$$\sum_{k=1}^{\infty} k^p z^k = \frac{z}{1-z} \mathbb{E}[X^p] > \Gamma(p+1) \frac{z^{p+1}}{(1-z)^{p+1}}.$$

Substituting $z = \gamma p_2^p$ back into the lower bound (Eqn. (12)), we obtain

$$\begin{aligned} \bar{\ell}_p^p(\boldsymbol{\eta}_{M(p_1)}, \boldsymbol{\eta}_{M(p_2)}) &\geq \frac{(p_1 - p_2)^p}{p_2^p} \cdot \Gamma(p+1) \frac{(\gamma p_2^p)^{p+1}}{(1-\gamma p_2^p)^{p+1}} \\ &= \Gamma(p+1) \gamma^{p+1} p_2^{p^2} \frac{(p_1 - p_2)^p}{(1-\gamma p_2^p)^{p+1}}. \end{aligned}$$

We now bound the denominator term $1 - \gamma p_2^p$. Since $1 - x^p \leq p(1-x)$ for any $x \in [0,1]$ and $p \geq 1$, we have:

$$1 - \gamma p_2^p = 1 - \gamma + \gamma(1 - p_2^p) \leq 1 - \gamma + \gamma p (1 - p_2) \leq (1-\gamma) + p(1-p_2).$$

By setting the transition probability $p_1 = \frac{4\gamma - 1}{3\gamma}$ as before (which implies $1 - p_1 = \frac{1-\gamma}{3\gamma}$), and ensuring $p_2 \geq 1/2$ and the perturbation satisfies $p_1 - p_2 \leq \frac{4\sqrt{3}}{9}(1-\gamma)$ as explicitly verified later in the construction, we deduce that for $\gamma \geq 1/2$:

$$1 - p_2 = (1 - p_1) + (p_1 - p_2) \leq \left(\frac{2}{3} + \frac{4\sqrt{3}}{9} \right) (1-\gamma) \leq 2(1-\gamma).$$

Consequently, $1 - \gamma p_2^p \leq (2p+1)(1-\gamma)$.

Substituting this upper bound back into the denominator, we establish

$$\bar{\ell}_p^p(\boldsymbol{\eta}_{M(p_1)}, \boldsymbol{\eta}_{M(p_2)}) \geq \frac{\Gamma(p+1)}{2^{p^2+p+1}(2p+1)^{p+1}} \frac{(p_1-p_2)^p}{(1-\gamma)^{p+1}}.$$

Taking the $1/p$-th root yields

$$\bar{\ell}_p(\boldsymbol{\eta}_{M(p_1)}, \boldsymbol{\eta}_{M(p_2)}) \geq \frac{[\Gamma(p+1)]^{1/p}}{2^{p+1/p+1}(2p+1)^{1+1/p}} \cdot \frac{p_1-p_2}{(1-\gamma)^{1+1/p}}. \tag{13}$$

Finally, we explicitly specify the construction of p_2 and the required sample size T to incorporate the KL divergence constraints. Depending on the specific lower bound component being analyzed, we set the transition probability p_2 as:

$$p_2 = p_1 - \frac{1}{8}\sqrt{\frac{p_1(1-p_1)S\log(S/2)}{T}} \quad \text{or} \quad p_2 = p_1 - \frac{1}{4}\sqrt{\frac{p_1(1-p_1)}{T\mu_{\min}}}.$$

To ensure the validity of $p_2 \geq 1/2$ and $1-\gamma p_2^p \leq (2p+1)(1-\gamma)$ used before, we require that $p_1 - p_2 \leq \frac{4\sqrt{3}}{9}(1-\gamma)$ holds. By observing that $p_1(1-p_1) \leq \frac{4}{9}(1-\gamma)$, these conditions are strictly satisfied provided that the sample size is sufficiently large, specifically:

$$T \geq \frac{3}{64}\frac{S\log(S/2) \vee \mu_{\min}^{-1}}{1-\gamma}.$$

This explicitly verifies the universal constant condition $T \geq C\,(1-\gamma)^{-1}\,(\mu_{\min}^{-1} + S\log S)$ stated in the theorem.

Substituting these into Eqn. (13), we deduce that

$$\inf_{\hat{\boldsymbol{\eta}}_T} \sup_{M \in \mathscr{M}(S)} \mathbb{E}[\bar{\ell}_p(\hat{\boldsymbol{\eta}}_T, \boldsymbol{\eta}_M)] \geq c \cdot 2^{-p}\sqrt{\frac{S\log S + \mu_{\min}^{-1}}{T(1-\gamma)^{1+2/p}}}.$$

where we use the fact that $\frac{[\Gamma(p+1)]^{1/p}}{2^{p+1/p+1}(2p+1)^{1+1/p}} \gtrsim 2^{-p}$, and $c > 0$ is a universal constant independent of p. This completes the proof.

C.2. Range of Iterations

LEMMA 6 (**RANGE OF $\boldsymbol{\eta}_t^\pi$**). *Suppose that $\alpha_t \in [0,1]$ for all $t \geq 0$, and $\boldsymbol{\eta}_0^\pi \in \mathscr{P}^{\mathcal{S}}$. Then we have, for all $t \geq 0$, $\boldsymbol{\eta}_t^\pi \in \mathscr{P}^{\mathcal{S}}$ in the case of* `NTD`*. Similarly, if $\boldsymbol{\eta}_0^\pi \in \mathscr{P}_K^{\mathcal{S}}$, then we have, for all $t \geq 0$, $\boldsymbol{\eta}_t^\pi \in \mathscr{P}_K^{\mathcal{S}}$ in the case of* `CTD`*.*

Proof. We will only prove the case of `NTD`, and the proof for `CTD` is similar by utilizing the property of the projection operator $\boldsymbol{\Pi}_K \colon \mathscr{P}^{\mathcal{S}} \to \mathscr{P}_K^{\mathcal{S}}$. And we only consider the generative model setting, the same conclusion holding for the Markovian setting.

We prove the result by induction. It is trivial that $\boldsymbol{\eta}_t^\pi \in \mathscr{P}^{\mathcal{S}}$ for $t = 0$. Suppose that $\boldsymbol{\eta}_{t-1}^\pi \in \mathscr{P}^{\mathcal{S}}$, and recall the updating scheme of NTD

$$\boldsymbol{\eta}_t^\pi = (\boldsymbol{\mathcal{I}} - \alpha_t \boldsymbol{\Lambda}_t)\, \boldsymbol{\eta}_{t-1}^\pi + \alpha_t \boldsymbol{\Lambda}_t \boldsymbol{\mathcal{T}}_t^\pi \boldsymbol{\eta}_{t-1}^\pi.$$

It is evident that $\mathscr{P}^{\mathcal{S}}$ is a convex set, considering that $\mathscr{P}^{\mathcal{S}}$ is a subset of the linear space $\mathcal{M}^{\mathcal{S}}$ introduced in Section 7.1. Therefore, we only need to show that $\boldsymbol{\mathcal{T}}_t^\pi \boldsymbol{\eta}_{t-1}^\pi \in \mathscr{P}^{\mathcal{S}}$, which trivially holds because $\boldsymbol{\mathcal{T}}_t^\pi$ is a random operator mapping from $\mathscr{P}^{\mathcal{S}}$ to $\mathscr{P}^{\mathcal{S}}$, and $\boldsymbol{\eta}_{t-1}^\pi \in \mathscr{P}^{\mathcal{S}}$. Applying the induction argument, we can arrive at the conclusion.

C.3. Remove the Dependence on T for Step Sizes

The proof adapts the sequence manipulation techniques established in Appendix C.3 in Li et al. (2024). We have shown that the conclusion holds for

$$T \geq \frac{C_4 \log^3 T}{\varepsilon^2 \mu_{\min}(1-\gamma)^3} \log \frac{|\mathcal{S}|\, T}{\delta},$$

$$\frac{1}{1 + \frac{c_5 \mu_{\min}(1-\sqrt{\gamma})T}{\log^2 T}} \leq \alpha_t \leq \frac{1}{1 + \frac{c_6 \mu_{\min}(1-\sqrt{\gamma})t}{\log^2 T}},$$

where $c_5 c_6 \leq \frac{1}{8}$, $c_5 > c_6 > 0$ and $C_4 > 0$.

Then for some $c_2 > c_3 > 0$ to be determined, we now assume

$$\frac{1}{1 + \frac{c_2 \mu_{\min}(1-\sqrt{\gamma})t}{\log^2 t}} \leq \alpha_t \leq \frac{1}{1 + \frac{c_3 \mu_{\min}(1-\sqrt{\gamma})t}{\log^2 t}}.$$

Next, we will show that if we consider the result of the $\frac{T}{2}$-th iteration with this step size scheme as the initialization of a new iteration process, then the step sizes in the subsequent $\frac{T}{2}$ iterations lie in the previously established range. If this is done, the conclusion still holds if we choose $T \geq \frac{2C_4 \log^3 T}{\varepsilon^2 \mu_{\min}(1-\gamma)^3} \log \frac{|\mathcal{S}|T}{\delta}$, because the initialization $\boldsymbol{\eta}_{T/2}^\pi \in \mathscr{P}^{\mathcal{S}}$ (or $\mathscr{P}_K^{\mathcal{S}}$ in the case of CTD) is independent of the samples obtained for $\frac{T}{2} < t \leq T$.

For any $\frac{T}{2} < t \leq T$, we denote $\tau := t - \frac{T}{2}$. We can see that there exist $c_2 > c_3 > 0$, such that the last inequality in both of the following lines hold simultaneously, which is expected.

$$\tilde{\alpha}_\tau := \alpha_t \leq \frac{1}{1 + \frac{c_3 \mu_{\min}(1-\sqrt{\gamma})(\tau + T/2)}{\log^2(\tau + T/2)}} \leq \frac{1}{1 + \frac{c_3 \mu_{\min}(1-\sqrt{\gamma})\tau}{\log^2 T}} \leq \frac{1}{1 + \frac{c_6 \mu_{\min}(1-\sqrt{\gamma})\tau}{\log^2(T/2)}},$$

and

$$\tilde{\alpha}_\tau = \alpha_t \geq \frac{1}{1 + \frac{c_2 \mu_{\min}(1-\sqrt{\gamma})(\tau + T/2)}{\log^2(\tau + T/2)}} \geq \frac{1}{1 + \frac{2c_2 \mu_{\min}(1-\sqrt{\gamma})T/2}{\log^2(T/2)}} \geq \frac{1}{1 + \frac{c_5 \mu_{\min}(1-\sqrt{\gamma})T/2}{\log^2(T/2)}}.$$

C.4. Proof of Lemma 1

Proof. Note that

$$(1-\sqrt{\gamma})\alpha_t \geq \frac{1-\sqrt{\gamma}}{1+\frac{c_5\mu_{\min}(1-\sqrt{\gamma})T}{\log^2 T}} \geq \frac{1-\sqrt{\gamma}}{\frac{2c_5\mu_{\min}(1-\sqrt{\gamma})T}{\log^2 T}} = \frac{\log^2 T}{2c_5\mu_{\min}T},$$

where in the second inequality, we used the lower bound assumption on T in Theorem 2: T is sufficiently large such that $\frac{c_5\mu_{\min}(1-\sqrt{\gamma})T}{\log^2 T} \geq 1$.

Then for any $0 \leq k \leq \frac{t}{2}$, we have

$$\begin{aligned}
\beta_k^{(t)} &\leq \left[1-\alpha_{t/2}\mu_{\min}(1-\sqrt{\gamma})\right]^{t/2} \\
&\leq \left(1-\frac{\log^2 T}{2c_5 T}\right)^{t/2} \\
&\leq \left(1-\frac{\log^2 T}{2c_5 T}\right)^{\frac{T}{2c_6\log T}} \\
&= \left\{\left(1-\frac{\log^2 T}{2c_5 T}\right)^{\frac{2c_5 T}{\log^2 T}}\right\}^{\frac{\log T}{4c_5c_6}} \\
&\leq \frac{1}{T^2},
\end{aligned}$$

where in the last inequality we used $c_5c_6 \leq \frac{1}{8}$.

And for any $\frac{t}{2} < k \leq t$, we have

$$\beta_k^{(t)} \leq \alpha_k \leq \frac{1}{\frac{c_6\mu_{\min}(1-\sqrt{\gamma})k}{\log^2 T}} \leq \frac{2\log^3 T}{\mu_{\min}(1-\sqrt{\gamma})T}.$$

C.5. Proof of Lemma 2

Proof. We denote $\mathbf{\Lambda} = \mathrm{diag}\{(\lambda_s)_{s\in\mathcal{S}}\}$, then $\mu_{\min} = \min_{s\in\mathcal{S}} \lambda_s$. Accordingly,

$$\begin{aligned}
\|\boldsymbol{\mathcal{I}} - \alpha_k \mathbf{\Lambda}(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{T}})\| &= \sup_{\boldsymbol{\xi}:\, \|\boldsymbol{\xi}\|=1} \|[\boldsymbol{\mathcal{I}} - \alpha_k \mathbf{\Lambda}(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{T}})]\,\boldsymbol{\xi}\| \\
&= \sup_{\boldsymbol{\xi}:\, \|\boldsymbol{\xi}\|=1} \max_{s\in\mathcal{S}} \|\{[\boldsymbol{\mathcal{I}} - \alpha_k \mathbf{\Lambda}(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{T}})]\,\boldsymbol{\xi}\}(s)\| \\
&= \sup_{\boldsymbol{\xi}:\, \|\boldsymbol{\xi}\|=1} \max_{s\in\mathcal{S}} \left\| (1-\alpha_k\lambda_s)\xi(s) + \alpha_k\lambda_s \sum_{s'\in\mathcal{S}} \boldsymbol{\mathcal{T}}(s,s')\xi(s') \right\| \\
&\le \sup_{\boldsymbol{\xi}:\, \|\boldsymbol{\xi}\|=1} \max_{s\in\mathcal{S}} \left[(1-\alpha_k\lambda_s)\,\|\xi(s)\| + \alpha_k\lambda_s \sum_{s'\in\mathcal{S}} \|\boldsymbol{\mathcal{T}}(s,s')\xi(s')\| \right] \\
&\le \sup_{\boldsymbol{\xi}:\, \|\boldsymbol{\xi}\|=1} \max_{s\in\mathcal{S}} \left[(1-\alpha_k\lambda_s) + \alpha_k\lambda_s\sqrt{\gamma} \sum_{s'\in\mathcal{S}} P(s,s')\,\|\xi(s')\| \right] \\
&\le \max_{s\in\mathcal{S}} \left[1 - \alpha_k\lambda_s\left(1-\sqrt{\gamma}\right)\right] \\
&\le 1 - \alpha_k\mu_{\min}\left(1-\sqrt{\gamma}\right).
\end{aligned}$$

C.6. Concentration of the Martingale Term (i)

Proof of Lemma 3. We first show that the inequality holds for each $t \ge \frac{T}{c_6 \log T}$ and then apply the union bound. We denote

$$\boldsymbol{\zeta}_k^{(1)} = \alpha_k \left(\prod_{i=k+1}^{t} [\boldsymbol{\mathcal{I}} - \alpha_i \mathbf{\Lambda}(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{T}})] \right) \mathbf{\Lambda}_k (\boldsymbol{\mathcal{T}}_k - \boldsymbol{\mathcal{T}})\,\boldsymbol{\eta}_{k-1},$$

where we omit the dependency on t in the notation for brevity, then LHS in the lemma equals $\|\sum_{k=1}^t \boldsymbol{\zeta}_k^{(1)}\|$. Let $\mathcal{F}_k$ denote the σ-field that contains all information up to time step k. Then $\{\zeta_k^{(1)}(s)\}_{k=1}^t$ is a $\{\mathcal{F}_k\}_{k=1}^t$-martingale difference sequence in the Hilbert space $\mathcal{M}$:

$$\mathbb{E}_{k-1}\left[\boldsymbol{\zeta}_k^{(1)}\right] = \alpha_k \prod_{i=k+1}^{t} [\boldsymbol{\mathcal{I}} - \alpha_i \mathbf{\Lambda}(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{T}})]\, \mathbb{E}_{k-1}\left[\mathbf{\Lambda}_k (\boldsymbol{\mathcal{T}}_k - \boldsymbol{\mathcal{T}})\,\boldsymbol{\eta}_{k-1}\right] = 0,$$

the first equality holds because a Bochner integral can be exchanged with a bounded linear operator, and the second equality holds due to the definition of $\mathbf{\Lambda}_k$ and the empirical distributional Bellman operator $\boldsymbol{\mathcal{T}}_k$.

We apply Freedman's inequality (Theorem 6) to bound this martingale.

The norm of the martingale difference $\|\boldsymbol{\zeta}_k^{(1)}\|$ can be bounded as follows

$$\begin{aligned}\left\|\boldsymbol{\zeta}_k^{(1)}\right\| &\le \alpha_k \left\|\prod_{i=k+1}^{t}\left[\boldsymbol{\mathcal{I}}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\right]\right\| \|\boldsymbol{\Lambda}_k\| \left\|\left(\boldsymbol{\mathcal{T}}_k-\boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}_{k-1}\right\| \\ &\le \alpha_k \left(\prod_{i=k+1}^{t}\left[1-\alpha_i\mu_{\min}\left(1-\sqrt{\gamma}\right)\right]\right)\frac{1}{\sqrt{1-\gamma}} \\ &= \frac{\beta_k^{(t)}}{\sqrt{1-\gamma}}.\end{aligned}$$

In the second inequality, we used Lemma 2 and $\|\boldsymbol{\Lambda}_k\| \le 1$. Hence, $\max_{k\in[t]} \|\boldsymbol{\zeta}_k^{(1)}\| \le \frac{\max_{k\in[t]}\beta_k^{(t)}}{\sqrt{1-\gamma}} \le \frac{1}{\sqrt{1-\gamma}}\max\{\frac{1}{T^2}, \frac{2\log^3 T}{\mu_{\min}(1-\sqrt{\gamma})T}\} \le \frac{4\log^3 T}{\mu_{\min}(1-\gamma)^{3/2}T} =: b^{(1)}$, according to Lemma 1.

We next bound the quadratic variation.

We first introduce some notations. For any $k \in \mathbb{N}$, we denote $\mathsf{Var}(\boldsymbol{\xi}) := (\mathbb{E}[\|\xi(s)\|^2])_{s\in\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$, $\mathsf{Var}_k(\boldsymbol{\xi}) := (\mathbb{E}_k[\|\xi(s)\|^2])_{s\in\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$ for any random element $\boldsymbol{\xi}$ in $\mathcal{M}^{\mathcal{S}}$. It remains to bound $\mathsf{Var}_{k-1}(\boldsymbol{\zeta}_k^{(1)})$.

To facilitate theoretical analysis, we assume in t-th iteration, we have the following virtual samples which are not used in the algorithms: for each $s \in \mathcal{S}$, $\tilde{a}_t(s) \sim \pi(\cdot \mid s), \tilde{r}_t(s) \sim \mathcal{P}_R(\cdot \mid s, \tilde{a}_t(s)), \tilde{s}_t'(s) \sim P(\cdot \mid s, \tilde{a}_t(s))$ independently. Then, we can define $[\boldsymbol{\mathcal{T}}_t\boldsymbol{\xi}](s)$ using the tuple $(s, \tilde{a}_t(s), \tilde{r}_t(s), \tilde{s}_t'(s))$ for any $\boldsymbol{\xi} \in \mathcal{M}^{\mathcal{S}}$ and $s \ne s_t$.

For any $\boldsymbol{\xi} \in \mathcal{M}^{\mathcal{S}}$, we define its one-step update Cramér variation as $\boldsymbol{\sigma}^2(\boldsymbol{\xi}) := \mathsf{Var}((\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}) \in \mathbb{R}^{\mathcal{S}}$, where $\widehat{\boldsymbol{\mathcal{T}}}$ is a random operator and has the same distribution as $\boldsymbol{\mathcal{T}}_1$.

For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{\mathcal{S}}$, we say $\boldsymbol{x} \le \boldsymbol{y}$ if $x(s) \le y(s)$ for all $s \in \mathcal{S}$. In this part, $\|\boldsymbol{x}\| := \|\boldsymbol{x}\|_\infty = \max_{s\in\mathcal{S}}|x(s)|$, $\sqrt{\boldsymbol{x}} := (\sqrt{x(s)})_{s\in\mathcal{S}}$, and $\boldsymbol{x}^2 := (x(s)^2)_{s\in\mathcal{S}}$. And for any $\boldsymbol{U} \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$, $\|\boldsymbol{U}\| := \|\boldsymbol{U}\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{R}^{\mathcal{S}}:\ \|\boldsymbol{x}\|=1}\|\boldsymbol{U}\boldsymbol{x}\| = \max_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}|U(s,s')|$.

For any $\{\boldsymbol{x}_k\}_{k=1}^n \subset \mathbb{R}^{\mathcal{S}}$, we abuse the notations to denote $\max_{k\in[n]}\boldsymbol{x}_k$ as the element-wise maximum $(\max_{k\in[n]}x_k(s))_{s\in\mathcal{S}}$.

We denote $\boldsymbol{I} \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ as the identity matrix, $\mathbf{1} \in \mathbb{R}^{\mathcal{S}}$ as the all-ones vector, and $\boldsymbol{P} := P^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$, *i.e.*, $P(s,s') := P^\pi(s' \mid s) = \sum_{a\in\mathcal{A}}\pi(a \mid s)P(s' \mid s,a)$.

Also, we let $\boldsymbol{P}_t$ be the empirical version of $\boldsymbol{P}$ corresponding to $\boldsymbol{\mathcal{T}}_t$, for any $k \in \mathbb{N}$. Namely,

$$P_t(s_t, s') = \delta_{s',s_t'}, \tag{14}$$

and

$$P_t(s,s') = \delta_{s',\tilde{s}_t'(s)}, \tag{15}$$

for any $s \neq s_t$ and $s' \in \mathcal{S}$.

With these notations, the quadratic variation is $\boldsymbol{W}_t^{(1)} := \sum_{k=1}^t \mathsf{Var}_{k-1}(\boldsymbol{\zeta}_k^{(1)})$. To bound the quadratic variation $\boldsymbol{W}_t^{(1)}$, we need to bound $\mathsf{Var}_{k-1}(\boldsymbol{\zeta}_k^{(1)})$ using the following lemma, whose proof can be found in Appendix C.11.

LEMMA 7.

$$\mathsf{Var}_{k-1}\left(\boldsymbol{\zeta}_k^{(1)}\right) \leq \alpha_k \beta_k^{(t)} \left(\prod_{i=k+1}^t \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \right) \boldsymbol{\Lambda} \boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right).$$

Hence, the quadratic variation $\boldsymbol{W}_t^{(1)} := \sum_{k=1}^t \mathsf{Var}_{k-1}(\boldsymbol{\zeta}_k^{(1)})$ can be bounded as follows:

$$\begin{aligned}
\boldsymbol{W}_t^{(1)} &= \sum_{k=1}^t \mathsf{Var}_{t-1}\left(\boldsymbol{\zeta}_k^{(1)}\right) \\
&\leq \sum_{k=1}^t \alpha_k \beta_k^{(t)} \left(\prod_{i=k+1}^t \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \right) \boldsymbol{\Lambda} \boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right) \\
&\leq \sum_{k=1}^{t/2} \alpha_k \beta_k^{(t)} \left\| \prod_{i=k+1}^t \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \right\| \|\boldsymbol{\Lambda}\| \left\|\boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right)\right\| \mathbf{1} \\
&\quad + \sum_{k=t/2+1}^t \alpha_k \beta_k^{(t)} \left(\prod_{i=k+1}^t \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \right) \boldsymbol{\Lambda} \boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right) \\
&\leq \sum_{k=1}^{t/2} \left(\beta_k^{(t)}\right)^2 \frac{1}{1-\gamma} \mathbf{1} + \left(\max_{k: t/2 < k \leq t} \beta_k^{(t)} \right) \sum_{k=t/2+1}^t \alpha_k \left(\prod_{i=k+1}^t \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \right) \boldsymbol{\Lambda} \boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right) \\
&\leq \frac{1}{2(1-\gamma)T^3} \mathbf{1} + \frac{2\log^3 T}{\mu_{\min}(1-\sqrt{\gamma})T} \left\{ \sum_{k=t/2+1}^t \alpha_k \prod_{i=k+1}^t \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \right\} \boldsymbol{\Lambda} \max_{k: t/2 < k \leq t} \boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right) \\
&\leq \frac{1}{2(1-\gamma)T^3} \mathbf{1} + \frac{4\log^3 T}{\mu_{\min}(1-\gamma)T} (\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P})^{-1} \max_{k: t/2 < k \leq t} \boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right),
\end{aligned}$$

In the third inequality, we used Lemma 2 and

$$\left\|\boldsymbol{\sigma}^2(\boldsymbol{\eta}_{k-1})\right\| \leq \int_0^{\frac{1}{1-\gamma}} \mathrm{d}x = \frac{1}{1-\gamma}.$$

In the fourth inequality, we used Lemma 1 to bound $\beta_k^{(t)}$. And in the last inequality, note that $\max_{k: t/2 \leq k < t} \boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right) \geq \mathbf{0}$, we can use the following lemma.

LEMMA 8. *For any $t \in \mathbb{N}$, the following inequality holds entry-wise:*

$$\sum_{k=t/2+1}^t \alpha_k \prod_{i=k+1}^t \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \leq (\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P})^{-1} \boldsymbol{\Lambda}^{-1}.$$

The proof of the lemma can be found in Appendix C.12.

This yields an almost sure upper bound for $\|\boldsymbol{W}_t^{(1)}\|$:

$$\begin{aligned}\left\|\boldsymbol{W}_t^{(1)}\right\| &\le \frac{1}{2(1-\gamma)T^3} + \frac{4\log^3 T}{\mu_{\min}(1-\gamma)T}\left\|(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}\right\| \max_{k:\, t/2<k\le t}\left\|\boldsymbol{\sigma}\left(\boldsymbol{\eta}_{k-1}\right)\right\| \\ &\le \frac{1}{2(1-\gamma)T^3} + \frac{8\log^3 T}{\mu_{\min}(1-\gamma)^3 T} \\ &\le \frac{9\log^3 T}{\mu_{\min}(1-\gamma)^3 T} \\ &=: \sigma_{(1)}^2.\end{aligned}$$

Let $H = \lceil 2\log_2 \frac{1}{1-\gamma}\rceil$, we have

$$\frac{\sigma_{(1)}^2}{2^H} \le \frac{9\log^3 T}{\mu_{\min}(1-\gamma)T}.$$

By applying Freedman's inequality (Theorem 6) and the union bound over $s \in \mathcal{S}$, we obtain with probability at least $1-\delta$, for all $t \ge \frac{T}{c_6 \log T}$ and $s \in \mathcal{S}$

$$\begin{aligned}\left|\sum_{k=1}^{t}\boldsymbol{\zeta}_k^{(1)}\right| \le& \sqrt{8\left(\boldsymbol{W}_t^{(1)} + \frac{\sigma_{(1)}^2}{2^H}\mathbf{1}\right)\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta}} + \frac{4}{3}b^{(1)}\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta}\mathbf{1} \\ \le& \sqrt{16\left(\boldsymbol{W}_t^{(1)} + \frac{9\log^3 T}{\mu_{\min}(1-\gamma)T}\mathbf{1}\right)\log\frac{|\mathcal{S}|T}{\delta}} + 3b^{(1)}\log\frac{|\mathcal{S}|T}{\delta}\mathbf{1} \\ \le& 8\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)T}\left[(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}\max_{k:\, t/2<k\le t}\boldsymbol{\sigma}^2\left(\boldsymbol{\eta}_{k-1}\right) + 3\cdot\mathbf{1}\right]} \\ &+ \frac{12\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{3/2}T}\mathbf{1}.\end{aligned}$$

where we used $\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta} \le 2\log\frac{|\mathcal{S}|T}{\delta}$ in the second inequality, which holds due to the choice of T. The following lemmas are required for deriving the upper bound, which hold for both cases of `NTD` and `CTD`, whose proofs can be found in Appendix C.13 and Appendix C.14 respectively.

LEMMA 9. *For any $t \in [T]$, then*

$$\boldsymbol{\sigma}^2(\boldsymbol{\eta}_t) - \boldsymbol{\sigma}^2(\boldsymbol{\eta}) \le 4\left\|\boldsymbol{\Delta}_t\right\|_{\bar{W}_1}\mathbf{1}.$$

LEMMA 10. *It holds that*

$$(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}\boldsymbol{\sigma}^2(\boldsymbol{\eta}) \le \frac{4}{1-\gamma}\mathbf{1}.$$

Combining the upper bound with the two lemmas, we get the desired conclusion.

$$\begin{aligned}\left|\sum_{k=1}^{t}\boldsymbol{\zeta}_k^{(1)}\right| \leq & 8\left\{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)T}\left[4\max_{k:\,t/2<k\leq t}\|\boldsymbol{\Delta}_{k-1}\|_{\bar{W}_1}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)^{-1}\mathbf{1}+\frac{8}{1-\gamma}\mathbf{1}\right]\right\}^{\frac{1}{2}} \\ & +\frac{12\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{3/2}T}\mathbf{1} \\ \leq & 22\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^2 T}\left(1+\max_{k:\,t/2<k\leq t}\|\boldsymbol{\Delta}_{k-1}\|_{\bar{W}_1}\right)}\mathbf{1}+\frac{12\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{3/2}T}\mathbf{1} \\ \leq & 34\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^2 T}\left(1+\max_{k:\,t/2<k\leq t}\|\boldsymbol{\Delta}_{k-1}\|_{\bar{W}_1}\right)}\mathbf{1},\end{aligned}$$

where in the last line, we used that, excluding the constant term, the first term is larger than the second term, given the choice of $T \geq \frac{C_4 \log^3 T}{\varepsilon^2 \mu_{\min}(1-\gamma)^3}\log\frac{|\mathcal{S}|T}{\delta}$.

C.7. Concentration of the Martingale Term (ii)

Proof of Lemma 4. The proof strategy is same as that of Lemma 3, and we will omit some details which can be checked easily. We denote

$$\boldsymbol{\zeta}_k^{(2)} = \alpha_k\left(\prod_{i=k+1}^{t}\left[\boldsymbol{\mathcal{I}}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\right]\right)\left(\boldsymbol{\Lambda}-\boldsymbol{\Lambda}_k\right)\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}_{k-1},$$

then (ii) $=\sum_{k=1}^{t}\boldsymbol{\zeta}_k^{(2)}$, and $\{\boldsymbol{\zeta}_k^{(2)}\}_{k=1}^{t}$ is a $\{\mathcal{F}_k\}_{k=1}^{t}$-martingale difference sequence.

The norm of the martingale difference $\|\boldsymbol{\zeta}_k^{(2)}\|$ can be bounded as follows

$$\begin{aligned}\left\|\boldsymbol{\zeta}_k^{(2)}\right\| &\leq \alpha_k\left\|\prod_{i=k+1}^{t}\left[\boldsymbol{\mathcal{I}}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\right]\right\|\|\boldsymbol{\Lambda}-\boldsymbol{\Lambda}_k\|\left\|\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}_{k-1}\right\| \\ &\leq \alpha_k\left(\prod_{i=k+1}^{t}\left[1-\alpha_i\mu_{\min}\left(1-\sqrt{\gamma}\right)\right]\right)\frac{1}{\sqrt{1-\gamma}} \\ &\leq \frac{\beta_k^{(t)}}{\sqrt{1-\gamma}}.\end{aligned}$$

Hence, $\max_{k\in[t]}\|\boldsymbol{\zeta}_k^{(2)}\| \leq \frac{\max_{k\in[t]}\beta_k^{(t)}}{\sqrt{1-\gamma}} \leq \frac{1}{\sqrt{1-\gamma}}\max\{\frac{1}{T^2}, \frac{4\log^3 T}{\mu_{\min}(1-\sqrt{\gamma})T}\} \leq \frac{4\log^3 T}{\mu_{\min}(1-\gamma)^{3/2}T} =: b^{(2)}$.

To bound the quadratic variation, we need to bound $\mathsf{Var}_{k-1}(\boldsymbol{\zeta}_k^{(2)})$ using the following lemma, whose proof can be found in Appendix C.15.

LEMMA 11. *We have that*

$$\mathsf{Var}_{k-1}\left(\boldsymbol{\zeta}_k^{(2)}\right) \leq 2\alpha_k\beta_k^{(t)}\left(\prod_{i=k+1}^{t}\left[\boldsymbol{I}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\right)\boldsymbol{\Lambda}\left|\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}_{k-1}\right|^2,$$

where $|(\mathcal{I}-\mathcal{T})\boldsymbol{\eta}_{k-1}|^2$ *is the entry-wise square of* $(\mathcal{I}-\mathcal{T})\boldsymbol{\eta}_{k-1}$.

Hence the quadratic variation $\boldsymbol{W}_t^{(2)} \coloneqq \sum_{k=1}^t \mathsf{Var}_{k-1}(\boldsymbol{\zeta}_k^{(2)})$ can be bounded as follows

$$
\begin{aligned}
\boldsymbol{W}_t^{(2)} =& \sum_{k=1}^{t} \mathsf{Var}_{t-1}\left(\boldsymbol{\zeta}_k^{(2)}\right) \\
\leq& 2\sum_{k=1}^{t} \alpha_k \beta_k^{(t)} \left(\prod_{i=k+1}^{t} \left[\boldsymbol{I}-\alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\right) \boldsymbol{\Lambda} \left|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right|^2 \\
\leq& 2\sum_{k=1}^{t/2} \alpha_k \beta_k^{(t)} \left\|\prod_{i=k+1}^{t} \left[\boldsymbol{I}-\alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\right\| \|\boldsymbol{\Lambda}\| \left\|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right\|^2 \mathbf{1} \\
&+ 2\sum_{k=t/2+1}^{t} \alpha_k \beta_k^{(t)} \left(\prod_{i=k+1}^{t} \left[\boldsymbol{I}-\alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\right) \boldsymbol{\Lambda} \left|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right|^2 \\
\leq& 2\sum_{k=1}^{t/2} \left(\beta_k^{(t)}\right)^2 \frac{1}{1-\gamma}\mathbf{1} \\
&+ 2\left(\max_{k:\,t/2<k\leq t} \beta_k^{(t)}\right) \sum_{k=t/2+1}^{t} \alpha_k \left(\prod_{i=k+1}^{t} \left[\boldsymbol{I}-\alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\right) \boldsymbol{\Lambda} \left|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right|^2 \\
\leq& \frac{1}{(1-\gamma)T^3}\mathbf{1} \\
&+ \frac{4\log^3 T}{\mu_{\min}(1-\sqrt{\gamma})T} \left\{\sum_{k=t/2+1}^{t} \alpha_k \prod_{i=k+1}^{t} \left[\boldsymbol{I}-\alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\right\} \boldsymbol{\Lambda} \max_{k:\,t/2<k\leq t} \left|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right|^2 \\
\leq& \frac{1}{(1-\gamma)T^3}\mathbf{1} + \frac{4\log^3 T}{\mu_{\min}(1-\sqrt{\gamma})T} (\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1} \max_{k:\,t/2<k\leq t} \left|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right|^2 \\
\leq& \frac{1}{(1-\gamma)T^3}\mathbf{1} + \frac{16\log^3 T}{\mu_{\min}(1-\gamma)^2 T} \max_{k:\,t/2<k\leq t} \left|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right|^2,
\end{aligned}
$$

where in the fifth inequality, we used $\max_{k:\,t/2\leq k<t} |(\mathcal{I}-\mathcal{T})\boldsymbol{\eta}_{k-1}|^2 \geq \mathbf{0}$ entry-wise, and Lemma 8.

This yields an almost sure upper bound for $\|\boldsymbol{W}_t^{(2)}\|$:

$$
\begin{aligned}
\left\|\boldsymbol{W}_t^{(2)}\right\| &\leq \frac{1}{(1-\gamma)T^3} + \frac{16\log^3 T}{\mu_{\min}(1-\gamma)^2 T} \max_{k:\,t/2<k\leq t} \left\|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right\|^2 \\
&\leq \frac{1}{(1-\gamma)T^3} + \frac{16\log^3 T}{\mu_{\min}(1-\gamma)^3 T} \\
&\leq \frac{17\log^3 T}{\mu_{\min}(1-\gamma)^3 T} \\
&=: \sigma_{(2)}^2.
\end{aligned}
$$

Letting $H = \lceil 2\log_2 \frac{1}{1-\gamma} \rceil$ as before, we have

$$\frac{\sigma_{(2)}^2}{2^H} \le \frac{17\log^3 T}{\mu_{\min}(1-\gamma)T}.$$

Applying Freedman's inequality (Theorem 6) and utilizing the union bound over $s \in \mathcal{S}$, we obtain with probability at least $1-\delta$, for all $t \ge \frac{T}{c_6 \log T}$ and $s \in \mathcal{S}$

$$\begin{aligned}
\left|\sum_{k=1}^{t} \zeta_k^{(2)}\right| \le & \sqrt{8\left(\boldsymbol{W}_t^{(2)} + \frac{\sigma_{(2)}^2}{2^H}\mathbf{1}\right)\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta}} + \frac{4}{3}b^{(2)}\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta}\mathbf{1} \\
\le & \sqrt{16\left(\boldsymbol{W}_t^{(2)} + \frac{17\log^3 T}{\mu_{\min}(1-\gamma)T}\mathbf{1}\right)\log\frac{|\mathcal{S}|T}{\delta}} + 3b^{(2)}\log\frac{|\mathcal{S}|T}{\delta}\mathbf{1} \\
\le & 17\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^2 T}\left[\max_{k:\, t/2<k\le t}\left|(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\,\boldsymbol{\eta}_{k-1}\right|^2 + \mathbf{1}\right]} + \quad \frac{12\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{3/2}T}\mathbf{1},
\end{aligned}$$

where we used $\log\frac{8|\mathcal{S}|T\log\frac{1}{1-\gamma}}{\delta} \le 2\log\frac{|\mathcal{S}|T}{\delta}$ in the second inequality, which holds due to the choice of T.

While

$$\begin{aligned}
\left|(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\,\boldsymbol{\eta}_{k-1}\right|^2 &= \left|(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\left(\boldsymbol{\eta}_{k-1}-\boldsymbol{\eta}\right)\right|^2 \\
&\le \left|(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\,\boldsymbol{\Delta}_{k-1}\right|^2 \\
&\le 4\left\|\boldsymbol{\Delta}_{k-1}\right\|^2\mathbf{1} \\
&\le 4\left\|\boldsymbol{\Delta}_{k-1}\right\|_{\bar{W}_1}\mathbf{1},
\end{aligned} \tag{16}$$

where in the first line, we used $(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\eta} = 0$, and in the last line, we used for any $s \in \mathcal{S}$,

$$\begin{aligned}
\left\|\Delta_{k-1}(s)\right\|^2 &= \int_0^{\frac{1}{1-\gamma}} \left|F_{\Delta_{k-1}(s)}(x)\right|^2 dx \\
&\le \left(\max_{x\in\left[0,\frac{1}{1-\gamma}\right]}\left|F_{\Delta_{k-1}(s)}(x)\right|\right)\int_0^{\frac{1}{1-\gamma}}\left|F_{\Delta_{k-1}(s)}(x)\right| \mathrm{d}x \\
&\le \left\|\Delta_{k-1}(s)\right\|_{W_1}.
\end{aligned}$$

Hence

$$\begin{aligned}\left|\sum_{k=1}^{t}\boldsymbol{\zeta}_k^{(2)}\right| \leq & 17\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^2 T}\left[\max_{k:\,t/2<k\leq t}\left|(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\,\boldsymbol{\eta}_{k-1}\right|^2+\mathbf{1}\right]}+\frac{12\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{3/2}T}\mathbf{1}\\ \leq & 34\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^2 T}\left(1+\max_{k:\,t/2<k\leq t}\|\boldsymbol{\Delta}_{k-1}\|_{\bar{W}_1}\right)\mathbf{1}}+\frac{12\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{3/2}T}\mathbf{1}\\ \leq & 46\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^2 T}\left(1+\max_{k:\,t/2<k\leq t}\|\boldsymbol{\Delta}_{k-1}\|_{\bar{W}_1}\right)\mathbf{1}}.\end{aligned}$$

where in the last line, we used that, excluding the constant term, the first term is larger than the second term, given the choice of $T \geq \frac{C_4\log^3 T}{\varepsilon^2\mu_{\min}(1-\gamma)^3}\log\frac{|\mathcal{S}|T}{\delta}$.

C.8. Solve the Recurrence Relation

THEOREM 7. *Suppose for all* $t \geq \frac{T}{c_6\log T}$,

$$\|\boldsymbol{\Delta}_t\|_{\bar{W}_1} \leq 81\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}\left(1+\max_{k:\,t/2<k\leq t}\|\boldsymbol{\Delta}_{k-1}\|_{\bar{W}_1}\right)}.$$

Then there exists some large universal constant $C_7 > 0$*, such that*

$$\|\boldsymbol{\Delta}_T\|_{\bar{W}_1} \leq C_7\left(\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}}+\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}\right).$$

Proof. To resolve this recurrence relation, we employ the argument in Appendix B.4 of Li et al. (2024).

For any $k \geq 0$, we denote

$$u_k := \max\left\{\|\boldsymbol{\Delta}_t\|_{\bar{W}_1}\;\middle|\; 2^k\frac{T}{c_6\log T}\leq t\leq T\right\},$$

for $0 \leq k \leq \log_2(c_6\log T)$. We can see that $\|\boldsymbol{\Delta}_T\|_{\bar{W}_1} \leq u_k$ for any valid k. Hence, it suffices to show the upper bound holds for u_k for any valid k. It can be verified that $u_0 \leq \frac{1}{1-\gamma}$, and for $k \geq 0$

$$u_{k+1} \leq 81\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}(1+u_k)}.$$

We first show that once $u_k \leq 1$, the subsequent values of u_{k+l} will also remain upper bounded by 1. Namely, if $u_k \leq 1$ for some $k \geq 1$, then

$$u_{k+1} \leq 81\sqrt{\frac{2\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}} \leq 1,$$

if $T \geq \frac{10^4 \log^3 T \log \frac{|\mathcal{S}|T}{\delta}}{\mu_{\min}(1-\gamma)^3}$.

Let $\tau := \inf\{k : u_k \leq 1\}$, then for any $k > \tau$, we have

$$u_k \leq 81\sqrt{\frac{2\left(\log^3 T\right)\left(\log \frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}} =: a.$$

For $k \leq \tau$, we have $u_k \geq 1$ and thereby

$$u_{k+1} \leq 35\sqrt{\frac{2\left(\log^3 T\right)\left(\log \frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}} u_k = a\sqrt{u_k},$$

i.e.,

$$\log u_{k+1} - 2\log a \leq \frac{1}{2}\left(\log u_k - 2\log a\right).$$

Apply it recursively, we have

$$\log u_{k+1} \leq 2\log a + \left(\frac{1}{2}\right)^{k+1}\left(\log u_0 - 2\log a\right),$$

i.e.,

$$u_{k+1} \leq a^2\left(\frac{u_0}{a^2}\right)^{1/2^k} = a^{2\left(1-1/2^k\right)} u_0^{1/2^k} \leq \frac{a^{2\left(1-1/2^k\right)}}{(1-\gamma)^{1/2^k}}.$$

To sum up, for any $k \geq 0$, u_{k+1} is always less than the sum of the upper bounds in cases of $k > \tau$ and $k \leq \tau$,

$$u_{k+1} \leq a + a^{2\left(1-1/2^k\right)} \frac{1}{(1-\gamma)^{1/2^k}}$$

Note that, $a^{2(1-1/2^k)} \leq \max\{a, \sqrt{a}\}$, and if we take $k \geq c_8 \log\log \frac{1}{1-\gamma}$ for any constant c_8, we have $\frac{1}{(1-\gamma)^{1/2^k}} = O(1)$. We can take the constant c_8 small enough such that $c_8 \log\log \frac{1}{1-\gamma} < \log_2(c_6 \log T)$ (this can be done and c_8 is universal since $\frac{1}{1-\gamma} = o(T)$), and thereby we can find a valid $k^\star \geq c_8 \log\log \frac{1}{1-\gamma} + 1$. Then

$$\|\mathbf{\Delta}_T\|_{\bar{W}_1} \leq u_{k^\star} \leq C_7\left(\sqrt{\frac{\left(\log^3 T\right)\left(\log \frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}} + \frac{\left(\log^3 T\right)\left(\log \frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 T}\right),$$

which is the desired conclusion, and C_7 is some large universal constant related to c_8.

C.9. The Cramér Error Bound

COROLLARY 1 (**SAMPLE COMPLEXITY FOR CRAMÉR ERROR WITH A FIXED-DISTRIBUTION GENERATIVE MODEL**).

In the fixed-distribution generative model setting, consider any given $\delta \in (0,1)$, $\varepsilon \in (0,1)$, *suppose* $K > 4\varepsilon^{-2}(1-\gamma)^{-2}$ *in* `CTD`, *the initialization is* $\boldsymbol{\eta}_0^\pi \in \mathscr{P}^{\mathcal{S}}$ ($\mathscr{P}_K^{\mathcal{S}}$ *in* `CTD`), *the total update steps* T *satisfies*

$$T \geq \frac{C_1 \log^3 T}{\varepsilon^2 \mu_{\min}(1-\gamma)^{5/2}} \log \frac{|\mathcal{S}|T}{\delta},$$

for some large universal constant $C_1 > 1$, i.e., $T = \widetilde{O}(\varepsilon^{-2}\mu_{\min}^{-1}(1-\gamma)^{-5/2})$, *and the step size sequence* $(\alpha_t)_{t\geq 1}$ *is monotonically non-increasing and satisfies*

$$\frac{1}{1+\frac{c_2\mu_{\min}(1-\sqrt{\gamma})t}{\log t}} \leq \alpha_t \leq \frac{1}{1+\frac{c_3\mu_{\min}(1-\sqrt{\gamma})t}{\log t}}$$

for some small universal constants $c_2 > c_3 > 0$. *Then, for both* `NTD` *and* `CTD`, *with probability at least* $1-\delta$, *the last iterate estimator satisfies* $\bar{\ell}_2(\boldsymbol{\eta}_T^\pi, \boldsymbol{\eta}^\pi) \leq \varepsilon$.

The similar conclusion holds in the Markovian setting, we omit it for brevity.

Discussion on the ℓ_2 *Sample Complexity Gap.* As noted in Remark 6, our convergence rate under the $\bar{\ell}_2$ distance has an extra $(1-\gamma)^{-1/2}$ factor compared to the minimax lower bound $\widetilde{\Omega}(\varepsilon^{-2}\mu_{\min}^{-1}(1-\gamma)^{-2})$ (Theorem 1) achieved by model-based methods (Rowland et al. 2024b). This gap is not a calculation artifact, but rather an inherent analytical bottleneck when dealing with standard model-free TD updates in the ℓ_2 space. In model-based methods, one only need to evaluate $(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}\boldsymbol{\sigma}^2(\boldsymbol{\eta})$, which is bounded by $4(1-\gamma)^{-1}$ as shown in Lemma 10. In contrast, our model-free analysis relies on applying Freedman's inequality to the martingale difference sequence, which requires bounding the difference in variances, $\boldsymbol{\sigma}^2(\boldsymbol{\eta}_t) - \boldsymbol{\sigma}^2(\boldsymbol{\eta})$ (as in Lemma 9) in terms of $\bar{\ell}_2(\boldsymbol{\eta}_t, \boldsymbol{\eta})$. However, the Lipschitz constant of the variance functional with respect to the Cramér distance scales as $(1-\gamma)^{-1/2}$. Propagating this $(1-\gamma)^{-1/2}$ Lipschitz continuity through the iterations inherently yields an extra $(1-\gamma)^{-1/2}$ penalty on the effective horizon, which is structurally unavoidable in the standard stochastic approximation framework.

Proof of Corollary 1. The difference in the proof compared to Section 7.2 arises in Lemma 3 and Lemma 4 when we control term (II). Now we further bound the result in Lemma 9 by the Cramér norm of the error term, which is used in Lemma 3,

$$\boldsymbol{\sigma}^2(\boldsymbol{\eta}_t) - \boldsymbol{\sigma}^2(\boldsymbol{\eta}) \leq 4\,\|\boldsymbol{\Delta}_t\|_{\bar{W}_1}\,\mathbf{1} \leq \frac{4}{\sqrt{1-\gamma}}\,\|\boldsymbol{\Delta}_t\|\,\mathbf{1}.$$

It is worth noting that although we invoke Lemma 3 and Lemma 9 to bound the variance difference via the W_1 metric, this intermediate step does not introduce any artificial slackness. This variance bound is essentially tight and the extra factor on the effective horizon is unavoidable.

As for Lemma 4, we need to modify Eqn. (16) to

$$\begin{aligned}\left|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\eta}_{k-1}\right|^2 &= \left|(\mathcal{I}-\mathcal{T})\,(\boldsymbol{\eta}_{k-1}-\boldsymbol{\eta})\right|^2\\ &\le \left|(\mathcal{I}-\mathcal{T})\,\boldsymbol{\Delta}_{k-1}\right|^2\\ &\le 4\left\|\boldsymbol{\Delta}_{k-1}\right\|^2\mathbf{1}\\ &\le \frac{4}{\sqrt{1-\gamma}}\left\|\boldsymbol{\Delta}_{k-1}\right\|\mathbf{1}.\end{aligned}$$

In the same way, we can derive the following recurrence relation: with probability at least $1-\delta$, for all $t \ge \frac{T}{c_6 \log T}$

$$\|\boldsymbol{\Delta}_t\| \le 81\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{5/2}T}\left(1+\max_{k:\,t/2<k\le t}\|\boldsymbol{\Delta}_{k-1}\|\right)}.$$

By repeating the reasoning of Theorem 7, we can obtain the desired conclusion,

$$\|\boldsymbol{\Delta}_T\| \le C_7\left(\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{5/2}T}}+\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^{5/2}T}\right),$$

which is less than ε if we take $C_4 \ge 2C_7^2$ and $T \ge \frac{C_4\log^3 T}{\varepsilon^2\mu_{\min}(1-\gamma)^{5/2}}\log\frac{|\mathcal{S}|T}{\delta}$. Here, $C_7 > 1$ is a large universal constant depending on c_6.

C.10. The Polyak-Ruppert Averaging

By the derivation in Appendix C.3 and the upper bound of $\|\boldsymbol{\Delta}_t\|_{\bar{W}_1}$ derived by solving the recursive relation, if we take $T_0 \ge \frac{T}{2}$, we have with probability at least $1-\delta$, for all $T_0 < t \le T$

$$\|\boldsymbol{\Delta}_t\|_{\bar{W}_1} \le C\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3 t}},$$

for some universal constant $C > 0$. Hence

$$\begin{aligned}\left\|\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\boldsymbol{\eta}_t^{\pi}-\boldsymbol{\eta}^{\pi}\right\|_{\bar{W}_1} \leq & \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\|\boldsymbol{\Delta}_t\|_{\bar{W}_1} \\ \leq & C\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3}}\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\frac{1}{\sqrt{t}} \\ \leq & 2C\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3}}\frac{\sqrt{T}-\sqrt{T_0}}{T-T_0} \\ \leq & 2C\sqrt{\frac{\left(\log^3 T\right)\left(\log\frac{|\mathcal{S}|T}{\delta}\right)}{\mu_{\min}(1-\gamma)^3(T-T_0)}},\end{aligned}$$

which is desired.

C.11. Proof of Lemma 7

Proof. We first introduce some notations. For any matrix of operators $\boldsymbol{\mathcal{U}} \in \mathcal{L}(\mathcal{M})^{\mathcal{S}\times\mathcal{S}}$, we denote $\boldsymbol{\mathcal{U}}(s) = (\boldsymbol{\mathcal{U}}(s,s'))_{s'\in\mathcal{S}} \in \mathcal{L}(\mathcal{M})^{\mathcal{S}}$ as the s-row of $\boldsymbol{\mathcal{U}}$. And for any $\boldsymbol{\xi} \in \mathcal{M}^{\mathcal{S}}$, we define the vector inner product operation $\boldsymbol{\mathcal{U}}(s)\boldsymbol{\xi} \coloneqq \sum_{s'\in\mathcal{S}} \boldsymbol{\mathcal{U}}(s,s')\xi(s') \in \mathcal{M}$.

We need the following lemma, which holds for both cases of `NTD` and `CTD`.

LEMMA 12. *For any $n \in \mathbb{N}$, let*

$$\begin{aligned}\boldsymbol{\mathcal{U}}_n &= \prod_{i=1}^{n}\left[\boldsymbol{\mathcal{I}} - \alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\right], \\ \boldsymbol{U}_n &= \prod_{i=1}^{n}\left[\boldsymbol{I} - \alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right], \\ u_n &= \prod_{i=1}^{n}\left[1 - \alpha_i\mu_{\min}\left(1-\sqrt{\gamma}\right)\right].\end{aligned}$$

Then for any $\nu \in \mathcal{M}$, and $s, s' \in \mathcal{S}$, we have

$$\|\boldsymbol{\mathcal{U}}_n(s,s')\nu\|^2 \leq u_n U_n(s,s')\|\nu\|^2.$$

The proof can be found in Appendix C.16.

Recall that $\widehat{\boldsymbol{\mathcal{T}}}$ is a random operator and has the same distribution as $\boldsymbol{\mathcal{T}}_1$. Similarly, we define $\widehat{\boldsymbol{\Lambda}} = \mathrm{diag}\{(\hat{\lambda}_s)_{s\in\mathcal{S}}\}$ is a random matrix independent of any other random variables, and has the same distribution as $\boldsymbol{\Lambda}_1$. Then $\lambda_s = \mathbb{E}[\hat{\lambda}_s]$. By definition of $\boldsymbol{\Lambda}_1$, there exist a unique $\hat{s} \in \mathcal{S}$ such that $\hat{\lambda}_{\hat{s}} = 1$,

and $\hat{\lambda}_s = 0$ for $s \neq \hat{s}$. If we assume the all-state generative model, which samples every state at time t, then $\widehat{\boldsymbol{\Lambda}} = \boldsymbol{I}$, and the following derivation also holds.

By Lemma 12, we get the following result. For any non-random $\boldsymbol{\xi} \in \mathcal{M}^{\mathcal{S}}$,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{\mathcal{U}}_n(s)\widehat{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right\|^2\right] \\
&\quad= \mathbb{E}\left[\left\|\sum_{s'\in\mathcal{S}} \boldsymbol{\mathcal{U}}_n(s,s')\hat{\lambda}_{s'}\left[(\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s')\right\|^2\right] \\
&\quad= \sum_{s'\in\mathcal{S}} \mathbb{E}\left[\hat{\lambda}_{s'}^2\left\|\boldsymbol{\mathcal{U}}_n(s,s')\left[(\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s')\right\|^2\right] \\
&\qquad+ \sum_{s'\neq s''} \mathbb{E}\left\{\hat{\lambda}_{s'}\hat{\lambda}_{s''}\left\langle \boldsymbol{\mathcal{U}}_n(s,s')\left[(\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s'), \boldsymbol{\mathcal{U}}_n(s,s'')\left[(\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s'')\right\rangle\right\} \\
&\quad= \sum_{s'\in\mathcal{S}} \mathbb{E}\left[\hat{\lambda}_{s'}\right]\mathbb{E}\left[\left\|\boldsymbol{\mathcal{U}}_n(s,s')\left[(\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s')\right\|^2\right] \\
&\quad\leq u_n \sum_{s'\in\mathcal{S}} \lambda_{s'} U_n(s,s')\mathbb{E}\left[\left\|\left[(\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s')\right\|^2\right] \\
&\quad= u_n \sum_{s'\in\mathcal{S}} U_n(s,s')\lambda_{s'}\boldsymbol{\sigma}^2(\boldsymbol{\xi})(s') \\
&\quad= u_n \boldsymbol{U}_n(s)\boldsymbol{\Lambda}\boldsymbol{\sigma}^2(\boldsymbol{\xi}).
\end{aligned}
$$

In the third equality, we used the fact that $\hat{\lambda}_{s'}\hat{\lambda}_{s''} = 0$ for $s' \neq s''$ under the fixed-distribution generative model, which samples only one state from μ at each iteration; and we used the fact that different rows of $\widehat{\boldsymbol{\mathcal{T}}}$ are independent, $\widehat{\boldsymbol{\mathcal{T}}}(s')\boldsymbol{\xi}$ is an unbiased estimator of $\boldsymbol{\mathcal{T}}(s')\boldsymbol{\xi} \in \mathcal{M}$ under the all-state generative model. In the fourth equality, we used the fact that $\widehat{\boldsymbol{\mathcal{T}}}$ and $\widehat{\boldsymbol{\Lambda}}$ are independent, and $\hat{\lambda}_{s'}^2 = \hat{\lambda}_{s'}$. In the first inequality, we used Lemma 12. Hence, $\mathsf{Var}(\boldsymbol{\mathcal{U}}_n\widehat{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\mathcal{T}}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}) \leq u_n\boldsymbol{U}_n\boldsymbol{\Lambda}\boldsymbol{\sigma}^2(\boldsymbol{\xi})$.

We now bound $\mathsf{Var}_{k-1}(\boldsymbol{\zeta}_k^{(1)})$:

$$
\begin{aligned}
\mathsf{Var}_{k-1}\left(\boldsymbol{\zeta}_k^{(1)}\right) =& \alpha_k^2 \mathsf{Var}_{k-1}\left(\prod_{i=k+1}^{t}\left[\boldsymbol{\mathcal{I}}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\right]\boldsymbol{\Lambda}_k\left(\boldsymbol{\mathcal{T}}_k-\boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}_{k-1}\right) \\
\leq& \alpha_k^2 \prod_{i=k+1}^{t}\left[1-\alpha_i\mu_{\min}\left(1-\sqrt{\gamma}\right)\right]\prod_{i=k+1}^{t}\left[\boldsymbol{I}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\boldsymbol{\Lambda}\boldsymbol{\sigma}^2(\boldsymbol{\eta}_{k-1}) \\
=& \alpha_k\beta_k^{(t)}\prod_{i=k+1}^{t}\left[\boldsymbol{I}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\boldsymbol{\Lambda}\boldsymbol{\sigma}^2(\boldsymbol{\eta}_{k-1}).
\end{aligned}
$$

C.12. Proof of Lemma 8

Proof.

$$
\begin{aligned}
&\sum_{k=t/2+1}^{t} \alpha_k \prod_{i=k+1}^{t} \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \\
&\quad= \sum_{k=t/2+1}^{t} \prod_{i=k+1}^{t} \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \alpha_k \boldsymbol{\Lambda}(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P})(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P})^{-1}\boldsymbol{\Lambda}^{-1} \\
&\quad= \sum_{k=t/2+1}^{t} \left\{ \prod_{i=k+1}^{t} \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] - \prod_{i=k}^{t} \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \right\} (\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P})^{-1}\boldsymbol{\Lambda}^{-1} \\
&\quad= \left\{ \boldsymbol{I} - \prod_{i=t/2+1}^{t} \left[\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)\right] \right\} (\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P})^{-1}\boldsymbol{\Lambda}^{-1} \\
&\quad\le (\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P})^{-1}\boldsymbol{\Lambda}^{-1},
\end{aligned}
$$

where the last inequality holds entry-wise since we can verify that all entries of $(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P})^{-1} = \sum_{k=0}^{\infty} \left(\sqrt{\gamma}\boldsymbol{P}\right)^k$ and $\boldsymbol{I} - \alpha_i \boldsymbol{\Lambda}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)$ are non-negative.

C.13. Proof of Lemma 9

Proof. For any $s \in \mathcal{S}$,

$$
\begin{aligned}
&\boldsymbol{\sigma}^2(\boldsymbol{\eta}_t)(s) - \boldsymbol{\sigma}^2(\boldsymbol{\eta})(s) \\
&\quad= \int_0^{\frac{1}{1-\gamma}} \left\{ \left[\mathbb{E}\left[F^2_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_t\right)(s)}(x) \right] - F^2_{(\mathcal{T}\boldsymbol{\eta}_t)(s)}(x) \right] - \left[\mathbb{E}\left[F^2_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)}(x) \right] - F^2_{(\mathcal{T}\boldsymbol{\eta})(s)}(x) \right] \right\} \mathrm{d}x \\
&\quad= \int_0^{\frac{1}{1-\gamma}} \left\{ \mathbb{E}\left[F^2_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_t\right)(s)}(x) - F^2_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)}(x) \right] + \left[F^2_{(\mathcal{T}\boldsymbol{\eta})(s)}(x) - F^2_{(\mathcal{T}\boldsymbol{\eta}_t)(s)}(x) \right] \right\} \mathrm{d}x \\
&\quad= \int_0^{\frac{1}{1-\gamma}} \left\{ \mathbb{E}\Big[\left(F_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_t\right)(s)}(x) - F_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)}(x) \right) \left(F_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_t\right)(s)}(x) + F_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)}(x) \right) \Big] \right. \\
&\qquad \left. + \left(F_{(\mathcal{T}\boldsymbol{\eta})(s)}(x) - F_{(\mathcal{T}\boldsymbol{\eta}_t)(s)}(x) \right) \left(F_{(\mathcal{T}\boldsymbol{\eta})(s)}(x) + F_{(\mathcal{T}\boldsymbol{\eta}_t)(s)}(x) \right) \right\} \mathrm{d}x \\
&\quad\le 2 \int_0^{\frac{1}{1-\gamma}} \left\{ \mathbb{E}\left[\left| F_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_t\right)(s)}(x) - F_{\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)}(x) \right| \right] + \left| F_{(\mathcal{T}\boldsymbol{\eta})(s)}(x) - F_{(\mathcal{T}\boldsymbol{\eta}_t)(s)}(x) \right| \right\} \mathrm{d}x \\
&\quad= 2 \left(\mathbb{E}\left[\left\| \left[\widehat{\mathcal{T}}\left(\boldsymbol{\eta}_t - \boldsymbol{\eta}\right) \right](s) \right\|_{W_1} \right] + \left\| \left[\mathcal{T}\left(\boldsymbol{\eta}_t - \boldsymbol{\eta}\right) \right](s) \right\|_{W_1} \right).
\end{aligned}
$$

In the case of NTD, $\mathcal{T}$ and $\widehat{\mathcal{T}}$ are γ-contraction w.r.t. the supremum 1-Wasserstein metric, hence

$$\begin{aligned}\boldsymbol{\sigma}^2(\boldsymbol{\eta}_t)(s)-\boldsymbol{\sigma}^2(\boldsymbol{\eta})(s) \leq & 2\left(\mathbb{E}\left[\left\|\left[\widehat{\mathcal{T}}\left(\boldsymbol{\eta}_t-\boldsymbol{\eta}\right)\right](s)\right\|_{W_1}\right]+\left\|\left[\mathcal{T}\left(\boldsymbol{\eta}_t-\boldsymbol{\eta}\right)\right](s)\right\|_{W_1}\right) \\ \leq & 4\gamma\left\|\boldsymbol{\eta}_t-\boldsymbol{\eta}\right\|_{\bar{W}_1} \\ \leq & 4\left\|\boldsymbol{\Delta}_t\right\|_{\bar{W}_1} .\end{aligned} \tag{17}$$

In the case of CTD, if we can show $\boldsymbol{\Pi}_K$ is non-expansive w.r.t. 1-Wasserstein metric, the conclusion still holds. For any $x, y \in [0,(1-\gamma)^{-1}]$ such that $x<y$, we denote $x \in [x_k, x_{k+1})$ and $y \in [x_l, x_{l+1})$, then $k \leq l$, by the definition of $\boldsymbol{\Pi}_K$, we have

$$\boldsymbol{\Pi}_K(\delta_x)=\frac{x_{k+1}-x}{\iota_K}\delta_{x_k}+\frac{x-x_k}{\iota_K}\delta_{x_{k+1}}, \tag{18}$$

$$\boldsymbol{\Pi}_K(\delta_y)=\frac{x_{l+1}-y}{\iota_K}\delta_{x_l}+\frac{y-x_l}{\iota_K}\delta_{x_{l+1}}. \tag{19}$$

If $k=l$, we can check that $W_1(\boldsymbol{\Pi}_K\delta_x, \boldsymbol{\Pi}_K\delta_y)=\iota_K\frac{y-x}{\iota_K}=y-x$. If $k<l$, we have $W_1(\boldsymbol{\Pi}_K\delta_x, \boldsymbol{\Pi}_K\delta_y) \leq W_1(\boldsymbol{\Pi}_K\delta_x, \delta_{x_{k+1}})+W_1(\delta_{x_{k+1}}, \delta_{x_l})+W_1(\delta_{x_l}, \boldsymbol{\Pi}_K\delta_y)=(x_{k+1}-x)+(x_l-x_{k+1})+(y-x_{x_l})=y-x$. Hence, for any $\nu_1, \nu_2 \in \mathscr{P}$ and for any transport plan $\kappa \in \Gamma(\nu_1,\nu_2)$, the previous results tell us the cost of the transport plan $\boldsymbol{\Pi}_K\kappa \in \Gamma(\boldsymbol{\Pi}_K\nu_1, \boldsymbol{\Pi}_K\nu_2)$ induced by $\boldsymbol{\Pi}_K$ is no greater than the cost of κ. Consequently, $W_1(\boldsymbol{\Pi}_K\nu_1, \boldsymbol{\Pi}_K\nu_2) \leq W_1(\nu_1,\nu_2)$, *i.e.*, $\boldsymbol{\Pi}_K$ is non-expansive w.r.t. 1-Wasserstein metric, which is desired.

In addition, the projection operator $\boldsymbol{\Pi}_K$ coincides with the identity operator on distributions supported on the grid $\{x_0,\ldots,x_K\}$. Therefore, $\|\boldsymbol{\Pi}_K\|_{\bar{W}_1}$ is exactly 1, which is desired in Proposition 3.

C.14. Proof of Lemma 10

Proof. Firstly, we show that for any $\boldsymbol{v} \geq \mathbf{0}$, we have $\left\|(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}\boldsymbol{v}\right\| \leq 2\left\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{v}\right\|$.

$$\begin{aligned}\left\|(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}\boldsymbol{v}\right\| &= \left\|(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}(\boldsymbol{I}-\gamma\boldsymbol{P})(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{v}\right\| \\ &= \left\|(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}\left[(1-\sqrt{\gamma})\boldsymbol{I}+\sqrt{\gamma}(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})\right](\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{v}\right\| \\ &= \left\|\left[(1-\sqrt{\gamma})(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}+\sqrt{\gamma}\boldsymbol{I}\right](\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{v}\right\| \\ &\leq (1-\sqrt{\gamma})\left\|(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P})^{-1}(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{v}\right\|+\sqrt{\gamma}\left\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{v}\right\| \\ &\leq \left(\frac{1-\sqrt{\gamma}}{1-\sqrt{\gamma}}+\sqrt{\gamma}\right)\left\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{v}\right\| \\ &\leq 2\left\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{v}\right\| .\end{aligned}$$

In the case of NTD, by Corollary 2, we have

$$\left\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{\sigma}^2(\boldsymbol{\eta})\right\| \leq \frac{1}{1-\gamma},$$

In the case of CTD, by Corollary 5.12 in Rowland et al. (2024b), we have

$$\left\|(\boldsymbol{I}-\gamma \boldsymbol{P})^{-1} \boldsymbol{\sigma}^2(\boldsymbol{\eta})\right\| \leq \frac{2}{1-\gamma},$$

given $K > \frac{4}{1-\gamma}$.

C.15. Proof of Lemma 11

Proof. The proof is similar to that of Lemma 7 in Appendix C.11, and we will use the same notations.

Utilizing Lemma 12, we get the following result. For any non-random $\boldsymbol{\xi} \in \mathcal{M}^{\mathcal{S}}$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{\mathcal{U}}_n(s)(\boldsymbol{\Lambda}-\widehat{\boldsymbol{\Lambda}})(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right\|^2\right] \\
&\quad \leq \mathbb{E}\left[\left\|\boldsymbol{\mathcal{U}}_n(s)\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right\|^2\right] \\
&\quad = \mathbb{E}\left[\left\|\sum_{s'\in\mathcal{S}} \boldsymbol{\mathcal{U}}_n(s,s')\hat{\lambda}_{s'}\left[(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s')\right\|^2\right] \\
&\quad = \sum_{s'\in\mathcal{S}} \mathbb{E}\left[\hat{\lambda}_{s'}^2\right]\left\|\boldsymbol{\mathcal{U}}_n(s,s')\left[(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s')\right\|^2 \\
&\qquad + \sum_{s'\neq s''} \mathbb{E}\left[\hat{\lambda}_{s'}\hat{\lambda}_{s''}\right]\left\langle \boldsymbol{\mathcal{U}}_n(s,s')\left[(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s'), \boldsymbol{\mathcal{U}}_n(s,s'')\left[(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s'')\right\rangle \\
&\quad = \sum_{s'\in\mathcal{S}} \mathbb{E}\left[\hat{\lambda}_{s'}\right]\left\|\boldsymbol{\mathcal{U}}_n(s,s')\left[(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s')\right\|^2 \\
&\quad \leq u_n \sum_{s'\in\mathcal{S}} \lambda_{s'} U_n(s,s')\left\|\left[(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right](s')\right\|^2 \\
&\quad = u_n \sum_{s'\in\mathcal{S}} U_n(s,s')\lambda_{s'}\left|(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right|^2(s') \\
&\quad = u_n \boldsymbol{U}_n(s)\boldsymbol{\Lambda}\left|(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}\right|^2.
\end{aligned}$$

In the first inequality, we used the fact that, for any random element $\boldsymbol{\nu}$ in Hilbert space $\mathcal{M}$, $\mathbb{E}[\|\boldsymbol{\nu} - \mathbb{E}[\boldsymbol{\nu}]\|^2] = \mathbb{E}[\|\boldsymbol{\nu}\|^2] - \|\mathbb{E}[\boldsymbol{\nu}]\|^2 \leq \mathbb{E}[\|\boldsymbol{\nu}\|^2]$. Hence, $\mathsf{Var}(\boldsymbol{\mathcal{U}}_n(\boldsymbol{\Lambda}-\widehat{\boldsymbol{\Lambda}})(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}) \leq u_n \boldsymbol{U}_n\boldsymbol{\Lambda}|(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}})\boldsymbol{\xi}|^2$.

We now bound $\mathsf{Var}_{k-1}(\boldsymbol{\zeta}_k^{(2)})$:

$$\begin{aligned}
\mathsf{Var}_{k-1}\left(\boldsymbol{\zeta}_k^{(2)}\right) =& \alpha_k^2 \mathsf{Var}_{k-1}\left(\prod_{i=k+1}^{t}\left[\boldsymbol{\mathcal{I}}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\right]\left(\boldsymbol{\Lambda}-\boldsymbol{\Lambda}_k\right)\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}_{k-1}\right) \\
\leq& \alpha_k^2 \prod_{i=k+1}^{t}\left[1-\alpha_i\mu_{\min}\left(1-\sqrt{\gamma}\right)\right]\prod_{i=k+1}^{t}\left[\boldsymbol{I}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\boldsymbol{\Lambda}\left|\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}_{k-1}\right|^2 \\
=& \alpha_k\beta_k^{(t)}\prod_{i=k+1}^{t}\left[\boldsymbol{I}-\alpha_i\boldsymbol{\Lambda}\left(\boldsymbol{I}-\sqrt{\gamma}\boldsymbol{P}\right)\right]\boldsymbol{\Lambda}\left|\left(\boldsymbol{\mathcal{I}}-\boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}_{k-1}\right|^2.
\end{aligned}$$

C.16. Proof of Lemma 12

Proof. We prove this result by induction. For $n = 0$, we have $\boldsymbol{\mathcal{U}}_0 = \boldsymbol{\mathcal{I}}$, $\boldsymbol{U}_0 = \boldsymbol{I}$, $u_0 = 1$, thereby the inequality holds trivially. Suppose the inequality holds true for $n-1$. To prove that the inequality holds for n, it is sufficient to show that, for any $\nu \in \mathcal{M}$,

$$\begin{aligned}&\left\|\left[(1-\alpha_n\lambda_s)\delta_{s,s'} + \alpha_n\lambda_s \boldsymbol{\mathcal{T}}(s,s')\right]\nu\right\|^2 \\ &\quad\le \left[(1-\alpha_n\lambda_{\min}) + \alpha_n\lambda_{\min}\sqrt{\gamma}\right]\left[(1-\alpha_n\lambda_s)\delta_{s,s'} + \alpha_n\lambda_s\sqrt{\gamma}P(s,s')\right]\|\nu\|^2,\end{aligned}$$

where $\delta_{s,s'} = 1$ if $s = s'$, and 0 otherwise.

LHS can be bounded as follows

$$\begin{aligned}&\left\|\left[(1-\alpha_n\lambda_s)\delta_{s,s'} + \alpha_n\lambda_s \boldsymbol{\mathcal{T}}(s,s')\right]\nu\right\|^2 \\ &\quad= (1-\alpha_n\lambda_s)^2\delta_{s,s'}\|\nu\|^2 + 2(1-\alpha_n\lambda_s)\alpha_n\lambda_s\delta_{s,s'}\langle\nu, \boldsymbol{\mathcal{T}}(s,s')\nu\rangle + \alpha_n^2\lambda_s^2\|\boldsymbol{\mathcal{T}}(s,s')\nu\|^2 \\ &\quad\le (1-\alpha_n\lambda_s)^2\delta_{s,s'}\|\nu\|^2 + 2(1-\alpha_n\lambda_s)\alpha_n\lambda_s\delta_{s,s'}\|\nu\|\,\|\boldsymbol{\mathcal{T}}(s,s')\nu\| + \alpha_n^2\lambda_s^2\|\boldsymbol{\mathcal{T}}(s,s')\nu\|^2,\end{aligned}$$

where we used the Cauchy-Schwarz inequality. We need to give an upper bound for $\|\boldsymbol{\mathcal{T}}(s,s')\nu\|^2$.

Note that $(\boldsymbol{\Pi}_K\boldsymbol{\mathcal{T}}^\pi)(s,s') = \boldsymbol{\Pi}_K(\boldsymbol{\mathcal{T}}^\pi(s,s'))$ and $\|\boldsymbol{\Pi}_K\| = 1$, we only need to consider the case of `NTD`, by the definition of $\boldsymbol{\mathcal{T}}(s,s')$, we have

$$\begin{aligned}\|\boldsymbol{\mathcal{T}}(s,s')\nu\|^2 =& \int_0^{\frac{1}{1-\gamma}}\left[\sum_{a\in\mathcal{A}}\pi(a|s)P(s'|s,a)\int_0^1 F_\nu\left(\frac{x-r}{\gamma}\right)\mathcal{P}_R(\mathrm{d}r|s,a)\right]^2\mathrm{d}x \\ =&P(s,s')^2\int_0^{\frac{1}{1-\gamma}}\left[\sum_{a\in\mathcal{A}}\frac{\pi(a|s)P(s'|s,a)}{P(s,s')}\int_0^1 F_\nu\left(\frac{x-r}{\gamma}\right)\mathcal{P}_R(\mathrm{d}r|s,a)\right]^2\mathrm{d}x \\ =&P(s,s')^2\int_0^{\frac{1}{1-\gamma}}\left\{\mathbb{E}_{a\sim\pi(\cdot|s),r\sim\mathcal{P}_R(\cdot|s,a)}\left[F_\nu\left(\frac{x-r}{\gamma}\right)\Big|s'\right]\right\}^2\mathrm{d}x \\ \le&P(s,s')^2\mathbb{E}_{a\sim\pi(\cdot|s),r\sim\mathcal{P}_R(\cdot|s,a)}\left\{\int_0^{\frac{1}{1-\gamma}}\left[F_\nu\left(\frac{x-r}{\gamma}\right)\right]^2\mathrm{d}x\Big|s'\right\} \\ =&\gamma P(s,s')^2\|\nu\|^2,\end{aligned}$$

where we used Jensen's inequality and Fubini's theorem. Substitute it back into the upper bound,

$$\begin{aligned}&\left\|\left[(1-\alpha_n\lambda_s)\delta_{s,s'} + \alpha_n\lambda_s \boldsymbol{\mathcal{T}}(s,s')\right]\nu\right\|^2 \\ &\quad\le (1-\alpha_n\lambda_s)^2\delta_{s,s'}\|\nu\|^2 + 2(1-\alpha_n\lambda_s)\alpha_n\lambda_s\delta_{s,s'}\|\nu\|\,\|\boldsymbol{\mathcal{T}}(s,s')\nu\| + \alpha_n^2\lambda_s^2\|\boldsymbol{\mathcal{T}}(s,s')\nu\|^2 \\ &\quad\le \left[(1-\alpha_n\lambda_s)^2\delta_{s,s'} + 2(1-\alpha_n\lambda_s)\alpha_n\lambda_s\delta_{s,s'}\sqrt{\gamma}P(s,s') + \alpha_n^2\lambda_s^2\gamma P(s,s')^2\right]\|\nu\|^2 \\ &\quad= \left[(1-\alpha_n\lambda_s)^2\delta_{s,s'} + \alpha_n\lambda_s\sqrt{\gamma}P(s,s')\right]^2\|\nu\|^2 \\ &\quad\le \left[(1-\alpha_n\lambda_s) + \alpha_n\lambda_s\sqrt{\gamma}\right]\left[(1-\alpha_n\lambda_s)\delta_{s,s'} + \alpha_n\lambda_s\sqrt{\gamma}P(s,s')\right]\|\nu\|^2 \\ &\quad\le \left[(1-\alpha_n\lambda_{\min}) + \alpha_n\lambda_{\min}\sqrt{\gamma}\right]\left[(1-\alpha_n\lambda_s)\delta_{s,s'} + \alpha_n\lambda_s\sqrt{\gamma}P(s,s')\right]\|\nu\|^2,\end{aligned}$$

which is desired.

C.17. Proof of Lemma 5

Proof. For simplicity, we use the same abbreviations as in Section 7.2, and we abuse the notation to abbreviate $\mu_{\pi,\min}$ as $\mu_{\min}$. Since we are conducting a per-epoch analysis, we will also omit e in the notations. As in Theorem 4 in Li et al. (2021), we define

$$\begin{aligned} t_{\text{frame}} &:= \frac{443 t_{\text{mix}}}{\mu_{\min}} \log \frac{24\,|\mathcal{S}|\, t_{\text{epoch}}}{\delta}, \\ \mu_{\text{frame}} &:= \frac{1}{2} \mu_{\min} t_{\text{frame}}, \\ \beta &:= \left(1-\sqrt{\gamma}\right)\left(1-(1-\alpha)^{\mu_{\text{frame}}}\right). \end{aligned}$$

Phase 1: when $\|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\| > 1$

$$\begin{aligned} \boldsymbol{\Delta}_t &= \boldsymbol{\eta}_t - \boldsymbol{\eta} \\ &= \left(\boldsymbol{\mathcal{I}} - \alpha \boldsymbol{\Lambda}_t\right) \boldsymbol{\eta}_{t-1} + \alpha \boldsymbol{\Lambda}_t \left(\boldsymbol{\mathcal{T}}_t \boldsymbol{\eta}_{t-1} - \boldsymbol{\mathcal{T}}_t \bar{\boldsymbol{\eta}} + \widetilde{\boldsymbol{\mathcal{T}}} \bar{\boldsymbol{\eta}}\right) - \boldsymbol{\eta} \\ &= \left(\boldsymbol{\mathcal{I}} - \alpha \boldsymbol{\Lambda}_t\right) \boldsymbol{\Delta}_{t-1} + \alpha \boldsymbol{\Lambda}_t \left(\boldsymbol{\mathcal{T}}_t \left(\boldsymbol{\eta}_{t-1} - \bar{\boldsymbol{\eta}}\right) + \widetilde{\boldsymbol{\mathcal{T}}} \bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) \\ &= \left(\boldsymbol{\mathcal{I}} - \alpha \boldsymbol{\Lambda}_t\right) \boldsymbol{\Delta}_{t-1} + \alpha \boldsymbol{\Lambda}_t \left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\right) \bar{\boldsymbol{\eta}} + \alpha \boldsymbol{\Lambda}_t \left(\boldsymbol{\mathcal{T}}_t - \boldsymbol{\mathcal{T}}\right) \left(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\right) + \alpha \boldsymbol{\Lambda}_t \boldsymbol{\mathcal{T}}_t \boldsymbol{\Delta}_{t-1}. \end{aligned}$$

Applying it recursively, we can further decompose the error into four terms:

$$\begin{aligned} \boldsymbol{\Delta}_t =& \left\{\sum_{k=1}^{t} \alpha \left(\prod_{i=k+1}^{t} \left(\boldsymbol{\mathcal{I}} - \alpha \boldsymbol{\Lambda}_i\right)\right) \boldsymbol{\Lambda}_k \left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\right) \bar{\boldsymbol{\eta}}\right\} + \left\{\sum_{k=1}^{t} \alpha \left(\prod_{i=k+1}^{t} \left(\boldsymbol{\mathcal{I}} - \alpha \boldsymbol{\Lambda}_i\right)\right) \boldsymbol{\Lambda}_k \left(\boldsymbol{\mathcal{T}}_k - \boldsymbol{\mathcal{T}}\right) \left(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\right)\right\} \\ &+ \left\{\sum_{k=1}^{t} \alpha \left(\prod_{i=k+1}^{t} \left(\boldsymbol{\mathcal{I}} - \alpha \boldsymbol{\Lambda}_i\right)\right) \boldsymbol{\Lambda}_k \boldsymbol{\mathcal{T}}_k \boldsymbol{\Delta}_{k-1}\right\} + \left\{\prod_{k=1}^{t} \left(\boldsymbol{\mathcal{I}} - \alpha \boldsymbol{\Lambda}_k\right) \boldsymbol{\Delta}_0\right\} \\ =& (\mathrm{I})_t + (\mathrm{II})_t + (\mathrm{III})_t + (\mathrm{IV})_t. \end{aligned}$$

For Term (I), we have the following upper bound whose proof can be found in Appendix C.18.

LEMMA 13. *For any* $\delta \in (0,1)$*, and* $N \geq t_{\text{frame}}$*, with probability at least* $1 - \frac{\delta}{3}$*, we have for all* $t \in [t_{\text{epoch}}]$*, in both cases of* `VR-NTD` *and* `VR-CTD`*,*

$$\|(I)_t\| \leq 2\sqrt{\frac{\log \frac{12|\mathcal{S}|N}{\delta}}{N\mu_{\min}}} \left(2\left\|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\right\| + \frac{1}{\sqrt{1-\gamma}}\right).$$

For Term (II), we have the following upper bound whose proof can be found in Appendix C.19.

LEMMA 14. *For any $\delta \in (0,1)$, with probability at least $1-\frac{\delta}{3}$, we have for all $t \in [t_{\text{epoch}}]$, in both cases of* `VR-NTD` *and* `VR-CTD`*,*

$$\|(II)_t\| \le 3\sqrt{\alpha \log \frac{6\,|\mathcal{S}|\,t_{\text{epoch}}}{\delta}}\,\|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\|\,.$$

For Term (III), we can check that

$$|(\text{III})_t| \le \sqrt{\gamma}\sum_{k=1}^{t}\|\boldsymbol{\Delta}_{k-1}\|\,\alpha\left(\prod_{i=k+1}^{t}(\boldsymbol{I}-\alpha\boldsymbol{\Lambda}_i)\right)\boldsymbol{\Lambda}_k\mathbf{1}.$$

For Term (IV), when $t < t_{\text{frame}}$, we use a trivial upper bound $\|(\text{IV})_t\| \le \|\boldsymbol{\Delta}_0\|$. And using a concentration inequality of empirical distributions of Markov chains (Li et al. 2021, Lemma 8), one can show that, for any $\delta \in (0,1)$, we have with probability at least $1-\frac{\delta}{3}$, for all $t_{\text{frame}} \le t \le t_{\text{epoch}}$, $\|(\text{IV})_t\| \le (1-\alpha)^{t\mu_{\min}/2}\,\|\boldsymbol{\Delta}_0\|$.

To summarize, we arrive at, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following results hold: For all $t < t_{\text{frame}}$,

$$\begin{aligned}|\boldsymbol{\Delta}_t| \le& \sqrt{\gamma}\sum_{k=1}^{t}\|\boldsymbol{\Delta}_{k-1}\|\,\alpha\left(\prod_{i=k+1}^{t}(\boldsymbol{I}-\alpha\boldsymbol{\Lambda}_i)\right)\boldsymbol{\Lambda}_k\mathbf{1} + 2\sqrt{\frac{\log\frac{12|\mathcal{S}|N}{\delta}}{N\mu_{\min}}}\left(2\,\|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\| + \frac{1}{\sqrt{1-\gamma}}\right)\mathbf{1}\\ &+ 3\sqrt{\alpha\log\frac{6\,|\mathcal{S}|\,t_{\text{epoch}}}{\delta}}\,\|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\|\,\mathbf{1} + \|\boldsymbol{\Delta}_0\|\,\mathbf{1},\end{aligned}$$

and for all $t_{\text{frame}} \le t \le t_{\text{epoch}}$, the above inequality still holds if we replace the last term $\|\boldsymbol{\Delta}_0\|\mathbf{1}$ with $(1-\alpha)^{t\mu_{\min}/2}\|\boldsymbol{\Delta}_0\|\mathbf{1}$.

To solve the recursive relation, we first have the crude upper bound whose proof is identical to Eqn. (49) in Li et al. (2021)

$$\begin{aligned}\|\boldsymbol{\Delta}_t\| \le& \frac{1}{1-\sqrt{\gamma}}\left\{2\sqrt{\frac{\log\frac{12|\mathcal{S}|N}{\delta}}{N\mu_{\min}}}\left(2\,\|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\| + \frac{1}{\sqrt{1-\gamma}}\right) + 3\sqrt{\alpha\log\frac{6\,|\mathcal{S}|\,t_{\text{epoch}}}{\delta}}\,\|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\| + \|\boldsymbol{\Delta}_0\|\right\}\\ =& \frac{1}{1-\sqrt{\gamma}}\left\{2\sqrt{\frac{\log\frac{12|\mathcal{S}|N}{\delta}}{(1-\gamma)N\mu_{\min}}} + \|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\|\left(2\sqrt{\frac{\log\frac{12|\mathcal{S}|N}{\delta}}{N\mu_{\min}}} + 3\sqrt{\alpha\log\frac{6\,|\mathcal{S}|\,t_{\text{epoch}}}{\delta}}\right) + \|\boldsymbol{\Delta}_0\|\right\}.\end{aligned}$$

To refine the upper bound, we use the argument in Theorem 5 in Li et al. (2021). Let $\xi := \frac{1}{8}$, $u_0 := \frac{\|\boldsymbol{\Delta}_0\|}{1-\sqrt{\gamma}}$, $u_t := \|\boldsymbol{v}_t\|$, where

$$\boldsymbol{v}_t = \begin{cases}\sqrt{\gamma}\sum_{k=1}^{t}u_{k-1}\alpha\left(\prod_{i=k+1}^{t}(\boldsymbol{I}-\alpha\boldsymbol{\Lambda}_i)\right)\boldsymbol{\Lambda}_k\mathbf{1} + \|\Delta_0\|\,\mathbf{1} & \text{for } 1 \le t \le t_{\text{th},\xi},\\ \sqrt{\gamma}\sum_{k=1}^{t}u_{k-1}\alpha\left(\prod_{i=k+1}^{t}(\boldsymbol{I}-\alpha\boldsymbol{\Lambda}_i)\right)\boldsymbol{\Lambda}_k\mathbf{1} & \text{for } t > t_{\text{th},\xi},\end{cases}$$

and $t_{\text{th},\xi}$ is defined as

$$t_{\text{th},\xi} := \max\left\{\frac{2\log\frac{1}{\xi(1-\sqrt{\gamma})\sqrt{1-\gamma}}}{\alpha\mu_{\min}}, t_{\text{frame}}\right\}.$$

We can find that $(1-\alpha)^{t\mu_{\min}/2}\|\boldsymbol{\Delta}_0\| \le (1-\sqrt{\gamma})\xi$ for any $t > t_{\text{th},\xi}$. Following the proof of Lemma 3 and Lemma 4 in Li et al. (2021), one can show that with probability at least $1-\delta$, for all $t \in [t_{\text{epoch}}]$

$$\begin{aligned}\|\boldsymbol{\Delta}_t\| \le& u_t + L + \xi,\\ \le& (1-\beta)^k \frac{\|\boldsymbol{\Delta}_0\|}{1-\sqrt{\gamma}} + L + \xi,\end{aligned}$$

where $k = (\lfloor \frac{t-t_{\text{th},\xi}}{t_{\text{frame}}} \rfloor)_+$, and

$$L = \frac{1}{1-\sqrt{\gamma}}\left\{2\sqrt{\frac{\log\frac{24|\mathcal{S}|N}{\delta}}{(1-\gamma)N\mu_{\min}}} + \|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\|\left(2\sqrt{\frac{\log\frac{24|\mathcal{S}|N}{\delta}}{N\mu_{\min}}} + 3\sqrt{\alpha\log\frac{12\,|\mathcal{S}|\,t_{\text{epoch}}}{\delta}}\right)\right\}.$$

If we take $\alpha = \min\{\frac{(1-\sqrt{\gamma})^2}{C\log\frac{|\mathcal{S}|t_{\text{epoch}}}{\delta}}, \frac{1}{\mu_{\text{frame}}}\}$, and $N \ge \max\{t_{\text{frame}}, \frac{C\log\frac{|\mathcal{S}|N}{\delta}}{(1-\sqrt{\gamma})^2(1-\gamma)\mu_{\min}}\}$ for some sufficiently large universal constant $C > 0$, we have

$$L \le \frac{1}{8} + \frac{1}{8}\|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\|.$$

If we take $t_{\text{epoch}} \ge t_{\text{frame}} + t_{\text{th},\xi} + \frac{8\log\frac{2}{1-\sqrt{\gamma}}}{(1-\sqrt{\gamma})\alpha\mu_{\min}}$, we have

$$(1-\beta)^k \le \frac{1}{8}\left(1-\sqrt{\gamma}\right).$$

Consequently,

$$\begin{aligned}\left\|\boldsymbol{\Delta}_{t_{\text{epoch}}}\right\| \le& \frac{1}{8}\|\boldsymbol{\Delta}_0\| + \left(\frac{1}{8} + \frac{1}{8}\|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\|\right) + \frac{1}{8}\\ \le& \frac{1}{2}\max\left\{1, \|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\|\right\},\end{aligned}$$

where we used $\boldsymbol{\Delta}_0 = \bar{\boldsymbol{\eta}} - \boldsymbol{\eta}$.

Phase 2: when $\|\bar{\boldsymbol{\eta}}-\boldsymbol{\eta}\| \le 1$ Following the idea of Section B.2 in Wainwright (2019b), we introduce a new operator $\check{\boldsymbol{\mathcal{T}}}$, which is a perturbation of distributional Bellman operator $\boldsymbol{\mathcal{T}}$ and defined as

$$\check{\boldsymbol{\mathcal{T}}}(\boldsymbol{\xi}) := \boldsymbol{\mathcal{T}}\boldsymbol{\xi} + \left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\right)\bar{\boldsymbol{\eta}},$$

for any $\boldsymbol{\xi} \in \mathcal{M}_1^{\mathcal{S}}$. And $\mathcal{M}_1$ is defined as

$$\mathcal{M}_1 := \left\{\nu:\ |\nu|(\mathbb{R}) < \infty, \nu(\mathbb{R}) = 1, \text{supp}(\nu) \subseteq \left[0, (1-\gamma)^{-1}\right]\right\}.$$

For the operator construction, we work on the event $\mathcal{E}_N$ defined in Appendix C.18, on which every state appears among the reference samples.

One can find that the Cramér metric ℓ_2 can be naturally extended to $\mathcal{M}_1$, making $(\mathcal{M}_1, \ell_2)$ a separable but not complete metric space, and $\check{\mathcal{T}}$ is a $\sqrt{\gamma}$-contraction in the product space $(\mathcal{M}_1^{\mathcal{S}}, \bar{\ell}_2)$. To resolve the completeness problem, we will use the completion of $\mathcal{M}_1$ instead. Using some basic analysis, one can show that $\check{\mathcal{T}}$ can be uniquely extended to be a $\sqrt{\gamma}$-contraction in the completion of $\mathcal{M}_1^{\mathcal{S}}$. For simplicity, we abuse the notations to use $\mathcal{M}_1$ to denote the completion space, and $\check{\mathcal{T}}$ to denote the unique extension. By the contraction mapping theorem (Bellemare et al. 2023, Proposition 4.7), $\check{\mathcal{T}}$ admits a unique fixed point $\check{\boldsymbol{\eta}} = (\check{\eta}(s))_{s\in\mathcal{S}} \in \mathcal{M}_1^{\mathcal{S}}$. We remark that $\check{\mathcal{T}}$ and $\check{\boldsymbol{\eta}}$ are only used for proof.

By the triangle inequality,

$$\|\boldsymbol{\Delta}_t\| = \left\|\boldsymbol{\eta}_t - \boldsymbol{\eta}\right\| \le \left\|\check{\boldsymbol{\Delta}}_t\right\| + \|\check{\boldsymbol{\eta}} - \boldsymbol{\eta}\|,$$

where $\check{\boldsymbol{\Delta}}_t := \boldsymbol{\eta}_t - \check{\boldsymbol{\eta}}$. First, we give an upper bound for the norm of $\check{\boldsymbol{\eta}} - \boldsymbol{\eta} \in \mathcal{M}^{\mathcal{S}}$, whose proof can be found in Appendix C.20.

LEMMA 15. *For any $\delta \in (0,1)$, when $N \ge \frac{C \log \frac{|\mathcal{S}|N}{\delta}}{\varepsilon^2(1-\sqrt{\gamma})^2(1-\gamma)\mu_{\min}}$ with probability at least $1 - \frac{\delta}{2}$, in both cases of* `VR-NTD` *and* `VR-CTD`*,*

$$\|\check{\boldsymbol{\eta}} - \boldsymbol{\eta}\| \le 11\sqrt{\frac{\log \frac{4|\mathcal{S}|N}{\delta}}{(1-\sqrt{\gamma})^2 N \mu_{\min}}} \le \frac{1}{16}\sqrt{1-\gamma}\varepsilon.$$

Next, we have the following upper bound for $\left\|\check{\boldsymbol{\Delta}}_t\right\|$, whose proof can be found in Appendix C.21.

LEMMA 16. *For any $\delta \in (0,1)$, with probability at least $1 - \frac{\delta}{2}$, in both cases of* `VR-NTD` *and* `VR-CTD`*,*

$$\left\|\check{\boldsymbol{\Delta}}_{t_{\text{epoch}}}\right\| \le \frac{1}{4}\left\|\check{\boldsymbol{\Delta}}_0\right\| + \frac{1}{8}\sqrt{1-\gamma}\varepsilon.$$

To summarize, we have

$$\begin{aligned}
\left\|\boldsymbol{\Delta}_{t_{\text{epoch}}}\right\| &\le \left\|\check{\boldsymbol{\Delta}}_{t_{\text{epoch}}}\right\| + \|\check{\boldsymbol{\eta}} - \boldsymbol{\eta}\| \\
&\le \frac{1}{4}\left\|\check{\boldsymbol{\Delta}}_0\right\| + \frac{3}{16}\sqrt{1-\gamma}\varepsilon \\
&\le \frac{1}{4}\|\boldsymbol{\Delta}_0\| + \frac{1}{4}\|\check{\boldsymbol{\eta}} - \boldsymbol{\eta}\| + \frac{3}{16}\sqrt{1-\gamma}\varepsilon \\
&\le \frac{1}{2}\max\left\{\|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\|, \sqrt{1-\gamma}\varepsilon\right\},
\end{aligned}$$

which is the desired conclusion.

C.18. Proof of Lemma 13

Proof. We can check that

$$\begin{aligned}\|(\mathrm{I})_t\| &\le \left\| \sum_{k=1}^{t} \alpha \left(\prod_{i=k+1}^{t} (\boldsymbol{\mathcal{I}} - \alpha \boldsymbol{\Lambda}_i) \right) \boldsymbol{\Lambda}_k \right\| \left\| \left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}} \right\| \\ &\le \left\| \sum_{k=1}^{t} \alpha \left(\prod_{i=k+1}^{t} (\boldsymbol{I} - \alpha \boldsymbol{\Lambda}_i) \right) \boldsymbol{\Lambda}_k \right\| \left\| \left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}} \right\| \\ &\le \left\| \left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}} \right\|,\end{aligned}$$

where the last inequality holds due to Eqn. (51b) in Li et al. (2021). We only need to bound $\|(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}})\bar{\boldsymbol{\eta}}\|$. Let $N_s = \sum_{i=0}^{N-1} \mathbb{1}\{\tilde{s}_i = s\}$ and define the event $\mathcal{E}_N = \{N_s > 0,\ \forall s \in \mathcal{S}\}$. On $\mathcal{E}_N$,

$$\left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}} = \left(\sum_{i=0}^{N-1} \widetilde{\boldsymbol{\Lambda}}_i \right)^{-1} \sum_{i=0}^{N-1} \widetilde{\boldsymbol{\Lambda}}_i \left(\widetilde{\boldsymbol{\mathcal{T}}}_i - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}}.$$

Note that, for each $s \in \mathcal{S}$, $[(\widetilde{\boldsymbol{\mathcal{T}}}_i - \boldsymbol{\mathcal{T}})\bar{\boldsymbol{\eta}}](s)$ is a mean-zero $\mathcal{M}$-valued random element bounded by

$$\begin{aligned}\left\| \left[\left(\widetilde{\boldsymbol{\mathcal{T}}}_i - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}} \right](s) \right\| &\le \left\| \left(\widetilde{\boldsymbol{\mathcal{T}}}_i - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}} \right\| \\ &\le \left\| \left(\widetilde{\boldsymbol{\mathcal{T}}}_i - \boldsymbol{\mathcal{T}} \right) (\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}) \right\| + \left\| \left(\widetilde{\boldsymbol{\mathcal{T}}}_i - \boldsymbol{\mathcal{T}} \right) \boldsymbol{\eta} \right\| \\ &\le 2 \|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\| + \frac{1}{\sqrt{1-\gamma}}.\end{aligned}$$

As in Lemma 7 of Li et al. (2021), we apply the Hilbert-space Hoeffding inequality to the centered samples at successive visits to each state. A union bound over all states and all deterministic visit counts $1 \le n \le N$ yields an event C_N with probability at least $1 - \frac{\delta}{6}$. On $C_N \cap \mathcal{E}_N$,

$$\left\| \left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}} \right\| \le \max_{s \in \mathcal{S}} \sqrt{\frac{2 \log \frac{12|\mathcal{S}|N}{\delta}}{N_s}} \left(2 \|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\| + \frac{1}{\sqrt{1-\gamma}} \right).$$

By a concentration inequality of empirical distributions of Markov chains (Li et al. 2021, Lemma 8), when $N \ge t_{\text{frame}}$, the event $\mathcal{G}_N = \left\{ N_s \ge \frac{1}{2} N \mu_{\min},\ \forall s \in \mathcal{S} \right\}$ has probability at least $1 - \frac{\delta}{6}$. Since $\mathcal{G}_N \subseteq \mathcal{E}_N$, a union bound shows that, with probability at least $1 - \frac{\delta}{3}$,

$$\left\| \left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}} \right) \bar{\boldsymbol{\eta}} \right\| \le 2 \sqrt{\frac{\log \frac{12|\mathcal{S}|N}{\delta}}{N \mu_{\min}}} \left(2 \|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\| + \frac{1}{\sqrt{1-\gamma}} \right).$$

C.19. Proof of Lemma 14

Proof. For each $s \in \mathcal{S}$, let $t_k(s)$ be the time when s is visited for the k-th time, and $K_t(s) \coloneqq \max\{k : t_k(s) \le t\}$ be the total number of times that s is visited, then

$$\begin{aligned}&\left[\sum_{k=1}^{t} \alpha \left(\prod_{i=k+1}^{t} (\mathcal{I} - \alpha \mathbf{\Lambda}_i)\right) \mathbf{\Lambda}_k \left(\mathcal{T}_k - \mathcal{T}\right)(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})\right](s) \\ &\quad = \sum_{k=1}^{K_t(s)} \alpha (1-\alpha)^{K_t(s)-k} \left(\mathcal{T}_{t_k(s)}(s) - \mathcal{T}(s)\right)(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}),\end{aligned}$$

in which

$$\left\|\alpha (1-\alpha)^{K_t(s)-k} \left(\mathcal{T}_{t_k(s)}(s) - \mathcal{T}(s)\right)(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})\right\| \le 2\alpha (1-\alpha)^{K_t(s)-k} \|\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\|.$$

Now, we can use the argument in Lemma 1 in Li et al. (2021) and replace the scalar version Hoeffding's inequality with the Hilbert space version to show that, with probability at least $1 - \frac{\delta}{3}$, for all $t \in [t_{\text{epoch}}]$

$$\left\|\sum_{k=1}^{t} \alpha \left(\prod_{i=k+1}^{t} (\mathcal{I} - \alpha \mathbf{\Lambda}_i)\right) \mathbf{\Lambda}_k \left(\mathcal{T}_k - \mathcal{T}\right)(\boldsymbol{\eta} - \bar{\boldsymbol{\eta}})\right\| \le 3\sqrt{\alpha \log \frac{6\,|\mathcal{S}|\, t_{\text{epoch}}}{\delta}}\, \|\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\|.$$

C.20. Proof of Lemma 15

Proof. Using the fixed point property, we have

$$\begin{aligned}|\check{\boldsymbol{\eta}} - \boldsymbol{\eta}| &= \left|\mathcal{T}\check{\boldsymbol{\eta}} + \left(\widetilde{\mathcal{T}} - \mathcal{T}\right)\bar{\boldsymbol{\eta}} - \mathcal{T}\boldsymbol{\eta}\right| \\ &\le |\mathcal{T}(\check{\boldsymbol{\eta}} - \boldsymbol{\eta})| + \left|\left(\widetilde{\mathcal{T}} - \mathcal{T}\right)\bar{\boldsymbol{\eta}}\right| \\ &\le \sqrt{\gamma}\boldsymbol{P}\,|\check{\boldsymbol{\eta}} - \boldsymbol{\eta}| + \left|\left(\widetilde{\mathcal{T}} - \mathcal{T}\right)\bar{\boldsymbol{\eta}}\right| \\ &\le \sum_{k=0}^{\infty} \left(\sqrt{\gamma}\boldsymbol{P}\right)^k \left|\left(\widetilde{\mathcal{T}} - \mathcal{T}\right)\bar{\boldsymbol{\eta}}\right| \\ &= \left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)^{-1} \left|\left(\widetilde{\mathcal{T}} - \mathcal{T}\right)\bar{\boldsymbol{\eta}}\right|.\end{aligned}$$

Hence, we only need to deal with $|(\widetilde{\mathcal{T}} - \mathcal{T})\bar{\boldsymbol{\eta}}|$. Using the triangle inequality,

$$\left|\left(\widetilde{\mathcal{T}} - \mathcal{T}\right)\bar{\boldsymbol{\eta}}\right| \le \left|\left(\widetilde{\mathcal{T}} - \mathcal{T}\right)(\bar{\boldsymbol{\eta}} - \boldsymbol{\eta})\right| + \left|\left(\widetilde{\mathcal{T}} - \mathcal{T}\right)\boldsymbol{\eta}\right|,$$

we need to bound the above two terms. Following the proof of Lemma 13 in Appendix C.18, we have the following upper bound for the first term, with probability at least $1 - \frac{\delta}{2}$

$$\left\|\left(\widetilde{\mathcal{T}} - \mathcal{T}\right)(\bar{\boldsymbol{\eta}} - \boldsymbol{\eta})\right\| \le 2\sqrt{\frac{\log \frac{4|\mathcal{S}|N}{\delta}}{N\mu_{\min}}}\, \|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\| \le 2\sqrt{\frac{\log \frac{4|\mathcal{S}|N}{\delta}}{N\mu_{\min}}},$$

where we used the assumption that $\|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}\| \le 1$ holds in the Phase 2. As for the second term, one can use the argument in Eqn. (94-96) in Li et al. (2021) and replace the scalar version Bernstein's inequality with the Hilbert space version to show that, with probability at least $1 - \frac{\delta}{2}$

$$\left|\left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}\right| \le 2\sqrt{\frac{2\log\frac{4|\mathcal{S}|N}{\delta}}{N\mu_{\min}}\boldsymbol{\sigma}^2(\boldsymbol{\eta})} + \frac{8\log\frac{4|\mathcal{S}|N}{\delta}}{3\sqrt{1-\gamma}N\mu_{\min}}.$$

To summarize, we have

$$\begin{aligned}
|\check{\boldsymbol{\eta}} - \boldsymbol{\eta}| &\le \left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)^{-1}\left|\left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\right)\bar{\boldsymbol{\eta}}\right| \\
&\le \left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)^{-1}\left(\left\|\left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\right)(\bar{\boldsymbol{\eta}} - \boldsymbol{\eta})\right\|\mathbf{1} + \left|\left(\widetilde{\boldsymbol{\mathcal{T}}} - \boldsymbol{\mathcal{T}}\right)\boldsymbol{\eta}\right|\right) \\
&\le 2\sqrt{\frac{\log\frac{4|\mathcal{S}|N}{\delta}}{(1-\sqrt{\gamma})^2 N\mu_{\min}}} + \frac{8\log\frac{4|\mathcal{S}|N}{\delta}}{3(1-\sqrt{\gamma})\sqrt{1-\gamma}N\mu_{\min}} + 2\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)^{-1}\sqrt{\frac{2\log\frac{4|\mathcal{S}|N}{\delta}}{N\mu_{\min}}\boldsymbol{\sigma}(\boldsymbol{\eta})} \\
&\le 5\sqrt{\frac{\log\frac{4|\mathcal{S}|N}{\delta}}{(1-\sqrt{\gamma})^2 N\mu_{\min}}} + 2\sqrt{\frac{2\log\frac{4|\mathcal{S}|N}{\delta}}{(1-\sqrt{\gamma})N\mu_{\min}}\left(\boldsymbol{I} - \sqrt{\gamma}\boldsymbol{P}\right)^{-1}\boldsymbol{\sigma}(\boldsymbol{\eta})} \\
&\le 11\sqrt{\frac{\log\frac{4|\mathcal{S}|N}{\delta}}{(1-\sqrt{\gamma})^2 N\mu_{\min}}},
\end{aligned}$$

where in the fourth inequality, we used $N \ge \frac{C\log\frac{|\mathcal{S}|N}{\delta}}{(1-\sqrt{\gamma})^2(1-\gamma)\mu_{\min}}$ and Jensen's inequality; and in the last inequality, we used Lemma 10. And the conclusion follows if we substitute N into the upper bound.

C.21. Proof of Lemma 16

Proof. For any $t \in [t_{\text{epoch}}]$,

$$\begin{aligned}
\check{\boldsymbol{\Delta}}_t &= \boldsymbol{\eta}_t - \check{\boldsymbol{\eta}} \\
&= \left(\boldsymbol{\mathcal{I}} - \alpha\boldsymbol{\Lambda}_t\right)\check{\boldsymbol{\Delta}}_{t-1} + \alpha\boldsymbol{\Lambda}_t\left(\boldsymbol{\mathcal{T}}_t\boldsymbol{\eta}_{t-1} - \boldsymbol{\mathcal{T}}_t\bar{\boldsymbol{\eta}} + \widetilde{\boldsymbol{\mathcal{T}}}\bar{\boldsymbol{\eta}} - \check{\boldsymbol{\eta}}\right) \\
&= \left(\boldsymbol{\mathcal{I}} - \alpha\boldsymbol{\Lambda}_t\right)\check{\boldsymbol{\Delta}}_{t-1} + \alpha\boldsymbol{\Lambda}_t\left(\boldsymbol{\mathcal{T}}_t\boldsymbol{\eta}_{t-1} - \boldsymbol{\mathcal{T}}_t\bar{\boldsymbol{\eta}} + \widetilde{\boldsymbol{\mathcal{T}}}\bar{\boldsymbol{\eta}} - \boldsymbol{\mathcal{T}}\check{\boldsymbol{\eta}} - \widetilde{\boldsymbol{\mathcal{T}}}\bar{\boldsymbol{\eta}} + \boldsymbol{\mathcal{T}}\bar{\boldsymbol{\eta}}\right) \\
&= \left(\boldsymbol{\mathcal{I}} - \alpha\boldsymbol{\Lambda}_t\right)\check{\boldsymbol{\Delta}}_{t-1} + \alpha\boldsymbol{\Lambda}_t\left(\boldsymbol{\mathcal{T}}_t\left(\boldsymbol{\eta}_{t-1} - \bar{\boldsymbol{\eta}}\right) - \boldsymbol{\mathcal{T}}(\check{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}})\right) \\
&= \left(\boldsymbol{\mathcal{I}} - \alpha\boldsymbol{\Lambda}_t\right)\check{\boldsymbol{\Delta}}_{t-1} + \alpha\boldsymbol{\Lambda}_t\left(\left(\boldsymbol{\mathcal{T}}_t - \boldsymbol{\mathcal{T}}\right)(\check{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}) + \boldsymbol{\mathcal{T}}_t\check{\boldsymbol{\Delta}}_{t-1}\right).
\end{aligned}$$

Applying it recursively, we can further decompose the error into three terms:

$$\begin{aligned}
\check{\boldsymbol{\Delta}}_t = &\left\{\sum_{k=1}^{t}\alpha\left(\prod_{i=k+1}^{t}\left(\boldsymbol{\mathcal{I}} - \alpha\boldsymbol{\Lambda}_i\right)\right)\boldsymbol{\Lambda}_k\left(\boldsymbol{\mathcal{T}}_k - \boldsymbol{\mathcal{T}}\right)(\check{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}})\right\} \\
&+ \left\{\sum_{k=1}^{t}\alpha\left(\prod_{i=k+1}^{t}\left(\boldsymbol{\mathcal{I}} - \alpha\boldsymbol{\Lambda}_i\right)\right)\boldsymbol{\Lambda}_k\boldsymbol{\mathcal{T}}_k\check{\boldsymbol{\Delta}}_{k-1}\right\} + \left\{\prod_{k=1}^{t}\left(\boldsymbol{\mathcal{I}} - \alpha\boldsymbol{\Lambda}_k\right)\check{\boldsymbol{\Delta}}_0\right\}
\end{aligned}$$

$$=(\mathrm{I})_t + (\mathrm{II})_t + (\mathrm{III})_t.$$

For Term (I), following the proof of Lemma 14 in Appendix C.19, we have with probability at least $1-\frac{\delta}{2}$

$$\|(\mathrm{I})_t\| \le 3\sqrt{\alpha \log \frac{4\,|\mathcal{S}|\,t_{\text{epoch}}}{\delta}}\,\|\check{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}\|\,.$$

For Term (II), we can check that

$$|(\mathrm{II})_t| \le \sqrt{\gamma}\sum_{k=1}^{t}\left\|\check{\boldsymbol{\Delta}}_{k-1}\right\|\alpha\left(\prod_{i=k+1}^{t}(\boldsymbol{I}-\alpha\boldsymbol{\Lambda}_i)\right)\boldsymbol{\Lambda}_k\mathbf{1}.$$

For Term (III), when $t < t_{\text{frame}}$, we use a trivial upper bound $\|(\mathrm{III})_t\| \le \|\check{\boldsymbol{\Delta}}_0\|$. And using a concentration inequality of empirical distributions of Markov chains (Li et al. 2021, Lemma 8), one can show that, for any $\delta \in (0,1)$, we have with probability at least $1-\frac{\delta}{2}$, for all $t_{\text{frame}} \le t \le t_{\text{epoch}}$, $\|(\mathrm{III})_t\| \le (1-\alpha)^{t\mu_{\min}/2}\,\|\check{\boldsymbol{\Delta}}_0\|$.

Following the proof in Phase 1 with ξ replaced with $\frac{1}{8}\sqrt{1-\gamma}\varepsilon$, we have with probability at least $1-\delta$, for all $t \in [t_{\text{epoch}}]$

$$\left\|\check{\boldsymbol{\Delta}}_t\right\| \le (1-\beta)^k\frac{\left\|\check{\boldsymbol{\Delta}}_0\right\|}{1-\sqrt{\gamma}} + \frac{3}{1-\sqrt{\gamma}}\sqrt{\alpha\log\frac{8\,|\mathcal{S}|\,t_{\text{epoch}}}{\delta}}\,\|\check{\boldsymbol{\eta}}-\bar{\boldsymbol{\eta}}\| + \frac{1}{8}\sqrt{1-\gamma}\varepsilon,$$

where $k = (\lfloor\frac{t-t_{\text{th},\xi}}{t_{\text{frame}}}\rfloor)_+$, and $t_{\text{th},\xi}$ is defined as

$$t_{\text{th},\xi} := \max\left\{\frac{2\log\frac{4}{\xi(1-\sqrt{\gamma})\sqrt{1-\gamma}}}{\alpha\mu_{\min}}, t_{\text{frame}}\right\}.$$

If we take $\alpha = c\min\{\frac{(1-\sqrt{\gamma})^2}{\log\frac{|\mathcal{S}|t_{\text{epoch}}}{\delta}}, \frac{1}{\mu_{\text{frame}}}\}$ for some sufficiently small universal constant $c>0$, we have

$$\frac{3}{1-\sqrt{\gamma}}\sqrt{\alpha\log\frac{8\,|\mathcal{S}|\,t_{\text{epoch}}}{\delta}}\,\|\check{\boldsymbol{\eta}}-\bar{\boldsymbol{\eta}}\| \le \frac{1}{8}\,\|\check{\boldsymbol{\eta}}-\bar{\boldsymbol{\eta}}\|\,.$$

And if we take $t_{\text{epoch}} \ge t_{\text{frame}} + t_{\text{th},\xi} + \frac{C'\log\frac{1}{1-\sqrt{\gamma}}}{(1-\sqrt{\gamma})\alpha\mu_{\min}}$ for some sufficiently large universal constant $C'>0$, we have

$$(1-\beta)^k \le \frac{1}{8}\left(1-\sqrt{\gamma}\right).$$

Consequently,

$$\begin{aligned}\left\|\check{\boldsymbol{\Delta}}_{t_{\text{epoch}}}\right\| &\le \frac{1}{8}\left\|\check{\boldsymbol{\Delta}}_0\right\| + \frac{1}{8}\,\|\check{\boldsymbol{\eta}}-\bar{\boldsymbol{\eta}}\| + \frac{1}{8}\sqrt{1-\gamma}\varepsilon\\ &\le \frac{1}{4}\left\|\check{\boldsymbol{\Delta}}_0\right\| + \frac{1}{8}\sqrt{1-\gamma}\varepsilon,\end{aligned}$$

where we used $\check{\boldsymbol{\Delta}}_0 = \bar{\boldsymbol{\eta}} - \check{\boldsymbol{\eta}}$.

Appendix D: Stochastic Distributional Bellman Equation and Operator

In this section, we use the same notations as in Appendix C and only consider the NTD setting. Inspired by stochastic categorical CDF Bellman operator introduced in Rowland et al. (2024b) used in the categorical setting, we introduce stochastic distributional Bellman operator $\mathscr{T} : \Delta(\mathscr{P}^{\mathcal{S}}) \to \Delta(\mathscr{P}^{\mathcal{S}})$ to derive an upper bound for $\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{\sigma}^2(\boldsymbol{\eta})\|$ in the case of NTD. For any $\boldsymbol{\phi} \in \Delta(\mathscr{P}^{\mathcal{S}})$, we denote $\boldsymbol{\eta}_{\boldsymbol{\phi}}$ be the random element in $\mathscr{P}^{\mathcal{S}}$ with law $\boldsymbol{\phi}$.

$$\mathscr{T}\boldsymbol{\phi} \coloneqq \mathrm{Law}\left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}_{\boldsymbol{\phi}}\right),$$

where $(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}_{\boldsymbol{\phi}})(\omega) \coloneqq (\widehat{\boldsymbol{\mathcal{T}}})(\omega)(\boldsymbol{\eta}_{\boldsymbol{\phi}})(\omega) \in \mathscr{P}^{\mathcal{S}}$ for any $\omega \in \Omega$, Ω is the corresponding probability space, and $\widehat{\boldsymbol{\mathcal{T}}}$ is independent of $\boldsymbol{\eta}_{\boldsymbol{\phi}}$. In this part, $\widehat{\boldsymbol{\mathcal{T}}}$ does not consist of Π_K since we only consider the NTD setting.

We consider the 1-Wasserstein metric W_1 on $\Delta(\mathscr{P}^{\mathcal{S}})$, the space of all probability measures on the space $(\mathscr{P}^{\mathcal{S}}, \bar{\ell}_2)$. Since $(\mathscr{P}^{\mathcal{S}}, \bar{\ell}_2)$ is Polish (complete and separable), the space $(\Delta(\mathscr{P}^{\mathcal{S}}), W_1)$ is also Polish (Villani et al. 2009, Theorem 6.18).

PROPOSITION 4. *The stochastic distributional Bellman operator $\mathscr{T}$ is a $\sqrt{\gamma}$-contraction on $\Delta(\mathscr{P}^{\mathcal{S}})$,* i.e.*, for any $\boldsymbol{\phi}, \boldsymbol{\phi}' \in \Delta(\mathscr{P}^{\mathcal{S}})$, we have*

$$W_1\left(\mathscr{T}\boldsymbol{\phi}, \mathscr{T}\boldsymbol{\phi}'\right) \le \sqrt{\gamma} W_1\left(\boldsymbol{\phi}, \boldsymbol{\phi}'\right).$$

Proof. Let $\boldsymbol{\kappa}^\star \in \Gamma(\boldsymbol{\phi}, \boldsymbol{\phi}')$ be the optimal coupling between $\boldsymbol{\phi}$ and $\boldsymbol{\phi}'$. The existence of $\boldsymbol{\kappa}^\star$ is guaranteed by Theorem 4.1 in Villani et al. (2009). And let the random element $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ in $(\mathscr{P}^{\mathcal{S}})^2$ have the law $\boldsymbol{\kappa}^\star$, where $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are both random elements in $\mathscr{P}^{\mathcal{S}}$. We denote $\mathscr{T}\boldsymbol{\kappa}^\star \coloneqq \mathrm{Law}[(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\xi}_1, \widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\xi}_2)] \in \Gamma(\mathscr{T}\boldsymbol{\phi}, \mathscr{T}\boldsymbol{\phi}')$.

$$\begin{aligned}
W_1\left(\mathscr{T}\boldsymbol{\phi}, \mathscr{T}\boldsymbol{\phi}'\right) &= \inf_{\boldsymbol{\kappa} \in \Gamma(\mathscr{T}\boldsymbol{\phi}, \mathscr{T}\boldsymbol{\phi}')} \int_{\left(\mathscr{P}^{\mathcal{S}}\right)^2} \bar{\ell}_2\left(\xi, \xi'\right) \boldsymbol{\kappa}\left(\mathrm{d}\xi, \mathrm{d}\xi'\right) \\
&\le \int_{\left(\mathscr{P}^{\mathcal{S}}\right)^2} \bar{\ell}_2\left(\xi, \xi'\right) \mathscr{T}\boldsymbol{\kappa}^\star\left(\mathrm{d}\xi, \mathrm{d}\xi'\right) \\
&= \mathbb{E}\left[\bar{\ell}_2\left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\xi}_1, \widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\xi}_2\right)\right] \\
&\le \sqrt{\gamma}\mathbb{E}\left[\bar{\ell}_2\left(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right)\right] \\
&= \sqrt{\gamma} \int_{\left(\mathscr{P}^{\mathcal{S}}\right)^2} \bar{\ell}_2\left(\xi, \xi'\right) \boldsymbol{\kappa}^\star\left(\mathrm{d}\xi, \mathrm{d}\xi'\right) \\
&= \sqrt{\gamma} \inf_{\boldsymbol{\kappa} \in \Gamma(\boldsymbol{\phi}, \boldsymbol{\phi}')} \int_{\left(\mathscr{P}^{\mathcal{S}}\right)^2} \bar{\ell}_2\left(\xi, \xi'\right) \boldsymbol{\kappa}\left(\mathrm{d}\xi, \mathrm{d}\xi'\right) \\
&= \sqrt{\gamma} W_1\left(\boldsymbol{\phi}, \boldsymbol{\phi}'\right).
\end{aligned}$$

By the proposition and contraction mapping theorem, there exists a unique fixed point of $\mathscr{T}$, we denote $\boldsymbol{\psi} \in \Delta(\mathscr{P}^{\mathcal{S}})$ as the fixed point. Hence, the stochastic distributional Bellman equation reads

$$\boldsymbol{\psi} = \mathscr{T}\boldsymbol{\psi}.$$

We denote by $\boldsymbol{\eta}_{\boldsymbol{\psi}}$ the random element in $\mathscr{P}$ with law $\boldsymbol{\psi}$, then $\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}_{\boldsymbol{\psi}}$ and $\boldsymbol{\eta}_{\boldsymbol{\psi}}$ have the same law. As shown in the following proposition, $\boldsymbol{\eta}_{\boldsymbol{\psi}}$ can be regarded as a noisy version of $\boldsymbol{\eta}$.

PROPOSITION 5. *It holds that*

$$\mathbb{E}\left[\boldsymbol{\eta}_{\boldsymbol{\psi}}\right] = \boldsymbol{\eta},$$

where the expectation is regarded as the Bochner integral in the space of all finite measures on $\mathscr{P}^{\mathcal{S}}$, which is a normed linear space equipped with the Cramér metric as its norm.

Proof.

$$\mathbb{E}\left[\boldsymbol{\eta}_{\boldsymbol{\psi}}\right] = \mathbb{E}\left[\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right] = \mathbb{E}\left\{\mathbb{E}\left[\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\middle|\boldsymbol{\eta}_{\boldsymbol{\psi}}\right]\right\} = \mathbb{E}\left[\boldsymbol{\mathcal{T}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right] = \boldsymbol{\mathcal{T}}\mathbb{E}\left[\boldsymbol{\eta}_{\boldsymbol{\psi}}\right],$$

where we used $\widehat{\boldsymbol{\mathcal{T}}}$ is independent of $\boldsymbol{\eta}_{\boldsymbol{\psi}}$. Since $\mathbb{E}[\boldsymbol{\eta}_{\boldsymbol{\psi}}]$ is the fixed point of $\boldsymbol{\mathcal{T}}$, we have $\mathbb{E}[\boldsymbol{\eta}_{\boldsymbol{\psi}}] = \boldsymbol{\eta}$.

Based on the concepts of $\mathscr{T}$ and $\boldsymbol{\psi}$, we can obtain the following second order distributional Bellman equation, which is similar to the classic second-order Bellman equation (Gheshlaghi Azar et al. 2013, Lemma 7).

Recall the one-step Cramér variation $\boldsymbol{\sigma}^2(\boldsymbol{\eta}) = (\mathbb{E}[\|(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta})(s) - \eta(s)\|^2])_{s\in\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$ used in the NTD setting. We denote $\boldsymbol{\sigma}^2 \coloneqq \boldsymbol{\sigma}^2(\boldsymbol{\eta})$ for simplicity, and $\boldsymbol{\sigma}^2_\infty \coloneqq (\mathbb{E}[\|\boldsymbol{\eta}_{\boldsymbol{\psi}}(s) - \eta(s)\|^2])_{s\in\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$.

PROPOSITION 6 (**SECOND ORDER DISTRIBUTIONAL BELLMAN EQUATION**). *The following relationship is satisfied:*

$$\boldsymbol{\sigma}^2_\infty = \boldsymbol{\sigma}^2 + \gamma \boldsymbol{P}\boldsymbol{\sigma}^2_\infty.$$

Proof. For any $s \in \mathcal{S}$,

$$\begin{aligned}
\sigma^2_\infty(s) =& \mathbb{E}\left[\left\|\boldsymbol{\eta}_{\boldsymbol{\psi}}(s) - \eta(s)\right\|^2\right] \\
=& \mathbb{E}\left[\left\|\left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s) - \eta(s)\right\|^2\right] \\
=& \mathbb{E}\left[\left\|\left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s) - \left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}\right)(s) + \left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}\right)(s) - \eta(s)\right\|^2\right] \\
=& \mathbb{E}\left[\left\|\left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}\right)(s) - \eta(s)\right\|^2\right] + \mathbb{E}\left[\left\|\left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s) - \left(\widehat{\boldsymbol{\mathcal{T}}}\boldsymbol{\eta}\right)(s)\right\|^2\right],
\end{aligned} \tag{20}$$

where the last equality holds since the cross term is zero as below

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)-\eta(s),\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s)-\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)\right\rangle\right]\\
&\quad=\mathbb{E}\left\{\mathbb{E}\left[\left\langle\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)-\eta(s),\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s)-\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)\right\rangle\middle|\widehat{\mathcal{T}}\right]\right\}\\
&\quad=\mathbb{E}\left\{\left\langle\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)-\eta(s),\mathbb{E}\left[\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s)\middle|\widehat{\mathcal{T}}\right]-\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)\right\rangle\right\}\\
&\quad=\mathbb{E}\left[\left\langle\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)-\eta(s),\mathbf{0}\right\rangle\right]\\
&\quad=0.
\end{aligned}$$

The first term in the last line of Eqn. (20) is $\sigma^2(s)$, we need to deal with the second term.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s)-\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)\right\|^2\right]\\
&\quad=\mathbb{E}\left\{\mathbb{E}\left[\left\|\left(\widehat{\mathcal{T}}\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s)-\left(\widehat{\mathcal{T}}\boldsymbol{\eta}\right)(s)\right\|^2\middle|\boldsymbol{\eta}_{\boldsymbol{\psi}}\right]\right\}\\
&\quad=\mathbb{E}\left\{\mathbb{E}_{\substack{a(s)\sim\pi(\cdot|s),\\ s'(s)\sim P(\cdot|s,a(s)),\\ r(s)\sim\mathcal{P}_R(\cdot|s,a(s))}}\left[\int_0^{\frac{1}{1-\gamma}}\left(F_{\left(\boldsymbol{\eta}_{\boldsymbol{\psi}}\right)(s'(s))}\left(\frac{x-r}{\gamma}\right)-F_{\eta(s'(s))}\left(\frac{x-r}{\gamma}\right)\right)^2\mathrm{d}x\middle|\boldsymbol{\eta}_{\boldsymbol{\psi}}\right]\right\}\\
&\quad=\gamma\sum_{s'\in\mathcal{S}}\mathbb{E}\left[\left\|\boldsymbol{\eta}_{\boldsymbol{\psi}}(s')-\eta(s')\right\|^2\right]\sum_{a\in\mathcal{A}}\pi(a\mid s)P(s'\mid s,a)\\
&\quad=\gamma\sum_{s'\in\mathcal{S}}P(s,s')\sigma_\infty^2(s').
\end{aligned}$$

Put these together, and we can arrive at the conclusion.

Now, we can derive a tighter upper bound for $\left\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{\sigma}^2(\boldsymbol{\eta})\right\|$.

COROLLARY 2. *We have that*

$$\left\|(\boldsymbol{I}-\gamma\boldsymbol{P})^{-1}\boldsymbol{\sigma}^2(\boldsymbol{\eta})\right\|=\left\|\boldsymbol{\sigma}_\infty^2\right\|\le\frac{1}{1-\gamma}.$$

Proof. Note that $\sigma_\infty^2(s)=\mathbb{E}[\|\boldsymbol{\eta}_{\boldsymbol{\psi}}(s)-\eta(s)\|^2]\le\int_0^{(1-\gamma)^{-1}}dx=\frac{1}{1-\gamma}$ for any $s\in\mathcal{S}$.

Appendix E: Other Technical Lemmas

LEMMA 17 (**BASIC INEQUALITIES FOR METRICS ON THE SPACE OF PROBABILITY MEASURES**).

For any $\nu_1,\nu_2\in\mathscr{P}$, we have $(\ell_2(\nu_1,\nu_2))^2\le W_1(\nu_1,\nu_2)\le\frac{1}{\sqrt{1-\gamma}}\ell_2(\nu_1,\nu_2)$ and $W_p(\nu_1,\nu_2)\le\frac{1}{(1-\gamma)^{1-\frac{1}{p}}}W_1^{\frac{1}{p}}(\nu_1,\nu_2)$.

Proof. By Cauchy-Schwarz inequality,

$$\begin{aligned} W_1(\nu_1,\nu_2) =& \int_0^{\frac{1}{1-\gamma}} |F_{\nu_1}(x) - F_{\nu_2}(x)| \mathrm{d}x \\ \leq& \sqrt{\int_0^{\frac{1}{1-\gamma}} 1^2 \mathrm{d}x} \sqrt{\int_0^{\frac{1}{1-\gamma}} |F_{\nu_1}(x) - F_{\nu_2}(x)|^2 \mathrm{d}x} \\ =& \frac{1}{\sqrt{1-\gamma}} \ell_2(\nu_1,\nu_2). \end{aligned}$$

For ℓ_2 distance, we have

$$\begin{aligned} \ell_2(\nu_1,\nu_2) =& \sqrt{\int_0^{\frac{1}{1-\gamma}} |F_{\nu_1}(x) - F_{\nu_2}(x)|^2 \mathrm{d}x} \\ \leq& \sqrt{\int_0^{\frac{1}{1-\gamma}} |F_{\nu_1}(x) - F_{\nu_2}(x)| \mathrm{d}x} \\ =& \sqrt{W_1(\nu_1,\nu_2)}. \end{aligned}$$

And for W_p distance, we have

$$\begin{aligned} W_p(\nu_1,\nu_2) =& \left(\inf_{\kappa \in \Gamma(\nu_1,\nu_2)} \int_{\left[0,\frac{1}{1-\gamma}\right]^2} |x-y|^p \, \kappa(\mathrm{d}x, \mathrm{d}y) \right)^{1/p} \\ \leq& \frac{1}{(1-\gamma)^{1-\frac{1}{p}}} \left(\inf_{\kappa \in \Gamma(\nu_1,\nu_2)} \int_{\left[0,\frac{1}{1-\gamma}\right]^2} |x-y| \, \kappa(\mathrm{d}x, \mathrm{d}y) \right)^{1/p} \\ =& \frac{1}{(1-\gamma)^{1-\frac{1}{p}}} W_1^{\frac{1}{p}}(\nu_1,\nu_2). \end{aligned}$$

LEMMA 18. *$(\mathcal{M}, \|\cdot\|_{\ell_2})$ and $(\mathcal{M}, \|\cdot\|_{W_1})$ are separable.*

Proof. Recall that $(\mathscr{P}, W_1)$ is separable (Villani et al. 2009, Theorem 6.18), and by Lemma 17, the Cramér distance ℓ_2 can be bounded by 1-Wasserstein distance W_1, hence $(\mathscr{P}, \ell_2)$ is also separable. Let d be either W_1 or ℓ_2, and A be the countable dense subset of $(\mathscr{P}, d)$. For any $\epsilon > 0, \nu \in \mathcal{M}$, we denote $\tilde{\nu} \coloneqq \frac{2\nu}{|\nu|(\mathbb{R})}$. Then we can find a $q \in \mathbb{Q}, \bar{\nu}_+, \bar{\nu}_- \in A, \mathsf{s.t.} |q - \frac{1}{2}|\nu|(\mathbb{R})| \leq \frac{\epsilon |\nu|(\mathbb{R})}{6\|\nu\|_d}$, $d(\bar{\nu}_+, \tilde{\nu}_+) \leq \frac{\epsilon}{3q}, d(\bar{\nu}_-, \tilde{\nu}_-) \leq \frac{\epsilon}{3q}$. Note that the set of all possible $\bar{\nu} = \bar{\nu}_+ - \bar{\nu}_-$ is countable. Let $\hat{\nu} \coloneqq q\bar{\nu}$,

then we have

$$\begin{aligned}\|\nu-\hat{\nu}\|_d &\le \|\nu-q\tilde{\nu}\|_d + \|q\tilde{\nu}-\hat{\nu}\|_d \\ &= \left\|\left(\frac{1}{2}|\nu|(\mathbb{R})-q\right)\tilde{\nu}\right\|_d + q\,\|(\tilde{\nu}_+-\bar{\nu}_+)-(\tilde{\nu}_--\bar{\nu}_-)\|_d \\ &\le \left|q-\frac{1}{2}|\nu|(\mathbb{R})\right|\frac{2\|\nu\|_d}{|\nu|(\mathbb{R})} + q\,\|\tilde{\nu}_+-\bar{\nu}_+\|_d + q\,\|\tilde{\nu}_--\bar{\nu}_-\|_d \\ &\le \frac{\epsilon|\nu|(\mathbb{R})}{6\|\nu\|_d}\frac{2\|\nu\|_d}{|\nu|(\mathbb{R})} + q\frac{\epsilon}{3q} + q\frac{\epsilon}{3q} \\ &= \epsilon.\end{aligned}$$

Hence we have found a countable dense subset for $(\mathcal{M}, \|\cdot\|_d)$ for $d = W_1$ or ℓ_2. Therefore, $(\mathcal{M}, \|\cdot\|_{\ell_2})$ and $(\mathcal{M}, \|\cdot\|_{W_1})$ are separable.

LEMMA 19. $(\mathcal{M}, \|\cdot\|_{W_1})$ *is not complete.*

Proof. Consider the Cauchy sequence in $(\mathcal{M}, \|\cdot\|_{W_1})$ $\nu_n = \sum_{i=1}^{n}(\delta_{\frac{1}{i}+\frac{1}{2^i}} - \delta_{\frac{1}{i}-\frac{1}{2^i}})$, which satisfies $\|\nu_n - \nu_{n+k}\|_{W_1} \le \sum_{i=1}^{k}\frac{1}{2^{n+i-1}} \le \frac{1}{2^{n-1}} \to 0$, but its limit $\sum_{i=1}^{\infty}(\delta_{\frac{1}{i}+\frac{1}{2^i}} - \delta_{\frac{1}{i}-\frac{1}{2^i}}) \notin \mathcal{M}$ because its total variation is infinity. Hence, this is a Cauchy sequence without a limit in $(\mathcal{M}, \|\cdot\|_{W_1})$, implying that $(\mathcal{M}, \|\cdot\|_{W_1})$ is not complete.

Appendix F: Additional Numerical Experiments

F.1. Convergence Behavior

We now verify the theoretical convergence rates from Section 4. We measure the error by $\bar{W}_1(\hat{\boldsymbol{\eta}}, \boldsymbol{\eta}^\pi) = \max_{s\le S} W_1(\hat{\eta}(s), \eta^\pi(s))$, the worst-case 1-Wasserstein distance over **recurrent states**, consistent with the sup-norm guarantee in our theorems. We evaluate this error using return distributions computed by distributional DP, as described in Section 5.2. Throughout this subsection the demand parameter is $\lambda = 10$ and each configuration is repeated over 20 independent seeds. In these convergence experiments, `CTD` and `VR-CTD` use 51 atoms. We approximate `NTD` and `VR-NTD` numerically using 1001 equally spaced atoms in $[0, (1-\gamma)^{-1}]$ and categorical projection at each update. This keeps the support size fixed but introduces projection error. We consider three data-collection mechanisms, each paired with the algorithms and step-size schedules described by the corresponding theorem:

1. *Fixed-distribution generative model* (Theorem 2). At each step a state s_t is drawn uniformly from $\mathcal{S}$ and one transition (r_t, s_t') is observed. The step size is $\alpha_t = 1/(1+\mu_{\min}(1-\sqrt{\gamma})\,t/\log t)$ with $\mu_{\min} = 1/|\mathcal{S}|$. Total samples: $T = 2\times 10^5$.

2. *Markovian with data-drop* (Theorem 3). The agent follows a single trajectory with burn-in $T_0 = \lceil t_{\text{mix}} \log(12/\delta) \rceil$ and performs one TD update every $q = \lceil t_{\text{mix}} \log(3T^\star/\delta) \rceil$ steps, where $\delta = 0.05$ and $T^\star = 5 \times 10^4$. The step size uses $\mu_{\pi,\min}$ in place of $\mu_{\min}$. Total raw samples: $T_0 + qT^\star \approx 3.75 \times 10^6$.
3. *Variance-reduced Markovian* (Theorem 4). The procedure runs $E = 50$ epochs. Each epoch collects $N = 10^4$ Markov samples for the reference operator and then performs $t_{\text{epoch}} = 5 \times 10^3$ TD updates with constant step size $\alpha = 0.3$. Total raw samples: $E(N + t_{\text{epoch}}) = 7.5 \times 10^5$.

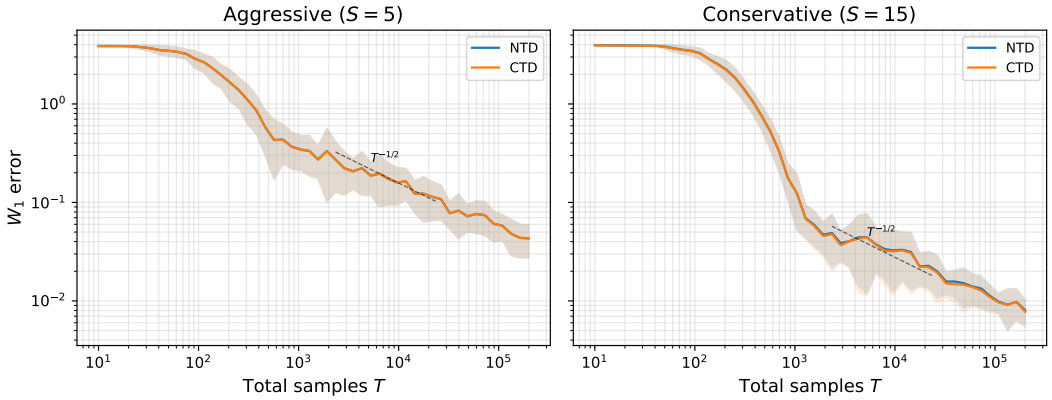


Figure 2 Convergence in the fixed-distribution generative model setting (Theorem 2). Log-log $\bar{W}_1$ error versus number of samples T for both policies and both algorithms (20 seeds, mean ± one std). The dashed reference line has slope $-1/2$.

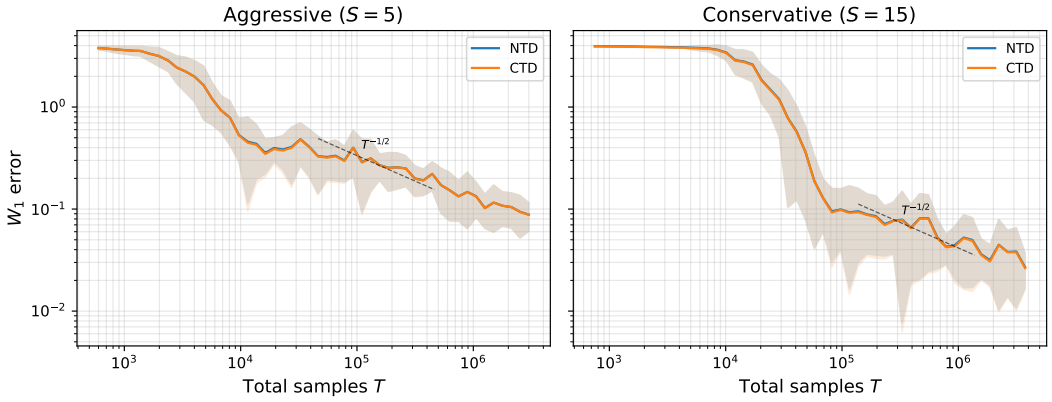


Figure 3 Convergence in the Markovian data-drop setting (Theorem 3). Log-log $\bar{W}_1$ versus total raw samples $T_0 + qT^\star$ for both policies and both algorithms (20 seeds, mean ± one std). The dashed reference line has slope $-1/2$.

The results in Figures 2–4 are consistent with our theoretical predictions. In the generative setting (Figure 2), the log-log $\bar{W}_1$ curves track the $T^{-1/2}$ reference line closely. The Markovian data-drop

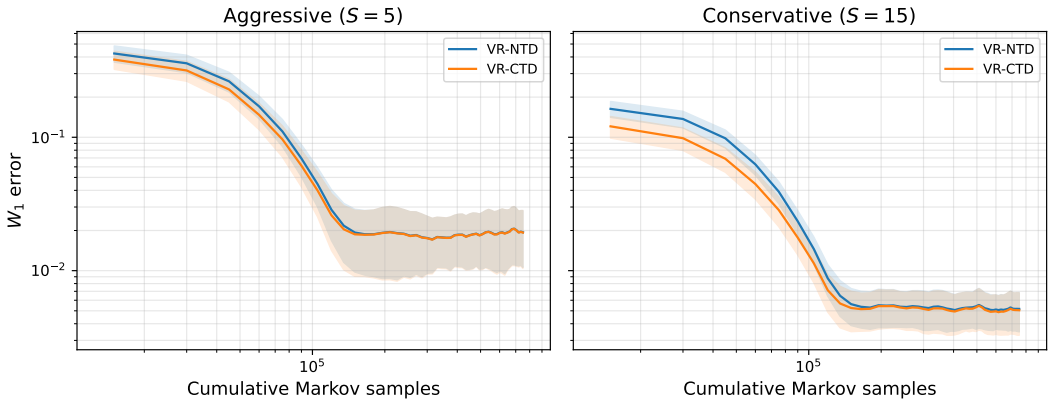


Figure 4 Convergence in the variance-reduced Markovian setting (Theorem 4). $\bar{W}_1$ versus cumulative Markov samples $E(N + t_{\text{epoch}})$ for both policies and both algorithms (20 seeds, mean ± one std). The final accuracy scales as $N^{-1/2}$ with the reference sample size N; for fixed N, the error contracts geometrically across epochs.

setting (Figure 3) also exhibits a $T^{-1/2}$ rate, but to reach the comparable accuracy as in the generative setting, roughly t_{mix} times more raw samples are needed, consistent with the $\widetilde{O}(\varepsilon^{-2} t_{\text{mix}})$ bound in Theorem 3.

The variance-reduced algorithms (Figure 4) exhibit a qualitatively different pattern. The error decreases geometrically as a function of epoch count, where each epoch uses $N + t_{\text{epoch}}$ Markovian samples. The accuracy at convergence is governed by the reference sample size N, scaling as $N^{-1/2}$. At comparable final accuracy, the variance-reduced method uses fewer samples than the data-drop method, because the sample complexity is $\widetilde{O}(\varepsilon^{-2} + t_{\text{mix}})$ (Theorem 4) when only considering the dependence on ε and t_{mix}, where t_{mix} enters additively rather than multiplicatively.

Across all three settings, `NTD` and `CTD` (and their variance-reduced counterparts) are nearly indistinguishable: here the categorical approximation error of `CTD` at $K = 50$ is dominated by the statistical error.

F.2. Sensitivity Analysis

We examine the sensitivity of the algorithms to their tuning parameters. Unless otherwise stated, experiments in this subsection use the conservative policy ($S = 15$), demand $\lambda = 10$, and 5 independent seeds.

Number of atoms K. The categorical approximation in `CTD` introduces a K-dependent error floor. To quantify this, we run `CTD` with $K \in \{11, 21, 51, 101, 201, 501\}$ in all three settings. All $\bar{W}_1$ errors are measured against a high-resolution return distribution computed by DP so that the approximation error is properly captured. Figure 5 shows that the `CTD` error decreases as $K^{-1/2}$ across all three

settings, consistent with the $O(K^{-1/2}(1-\gamma)^{-1})$ approximation bound established in Proposition 3 in Rowland et al. (2018).

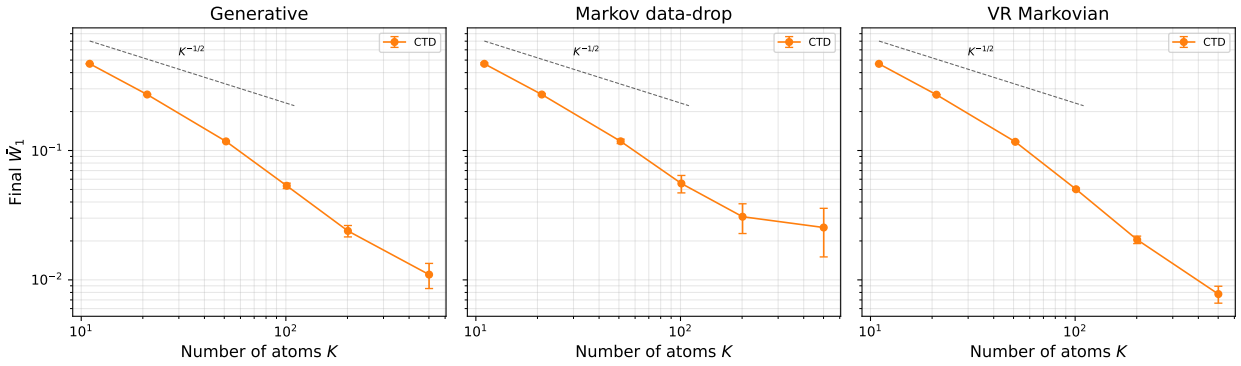


Figure 5 Categorical approximation error of CTD versus K, one panel per setting (5 seeds, mean ± one std). The gray reference line has slope $-1/2$. All $\bar{W}_1$ errors are computed against a high-resolution return distribution computed by DP.

Step size. The step-size schedules in our theorems involve unspecified universal constants. In the generative and data-drop settings, the decaying step size takes the form $\alpha_t = 1/(1 + c\,\mu_{\min}(1-\sqrt{\gamma})\,t/\log t)$, where $\mu_{\min}$ is defined as in Section F.1; in the variance-reduced setting, we use a constant step size α. Figure 6 examines the sensitivity to the constant c and to α using NTD ($K = 1000$). In the generative and data-drop settings (Figure 6(a)–(b)), a small c (e.g., $c = 0.1$) gives a large step size that reduces the error quickly at first but plateaus at a higher level. Conversely, a large c (e.g., $c = 10$) decays the step size faster, resulting in slower initial progress but a lower error floor. The range $c \in [0.5, 2.0]$ provides a good compromise. For the variance-reduced setting (Figure 6(c)), the algorithm diverges at $\alpha = 1.0$, converges slowly at $\alpha = 0.01$, and converges quickly to similar final accuracy for $\alpha \in [0.1, 0.5]$. At $\gamma = 0.9$, the theoretical step size from Theorem 4 is $\alpha \approx 0.001$, well below the effective range. This gap is expected, as the theoretical constant must ensure convergence for all MDP instances.

Reference sample size N. As Theorem 4 shows, the sample complexity of the variance-reduced algorithms depends on N through the term $\Omega(\varepsilon^{-2}(1-\gamma)^{-3} + t_{\text{mix}})$. To examine this dependence empirically, we vary $N \in \{100, 500, 10^3, 5\times10^3, 10^4, 5\times10^4\}$. Figure 7 (left) shows that all configurations exhibit geometric contraction across epochs, and the error floor decreases substantially as N increases. The log-log plot in Figure 7 (right) confirms that the final error scales as $N^{-1/2}$, consistent with Theorem 4.

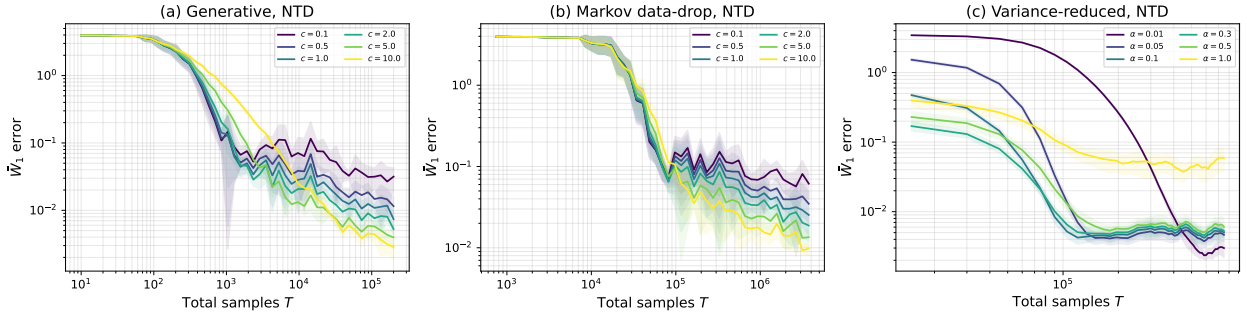


Figure 6 Sensitivity analysis of step size for NTD ($K = 1000$, 5 seeds, mean ± one std). (a) Fixed-distribution generative model setting: varying the decaying-step-size constant c. (b) Markovian data-drop setting: same sweep over c. (c) Variance-reduced Markovian setting: varying the constant step size α.

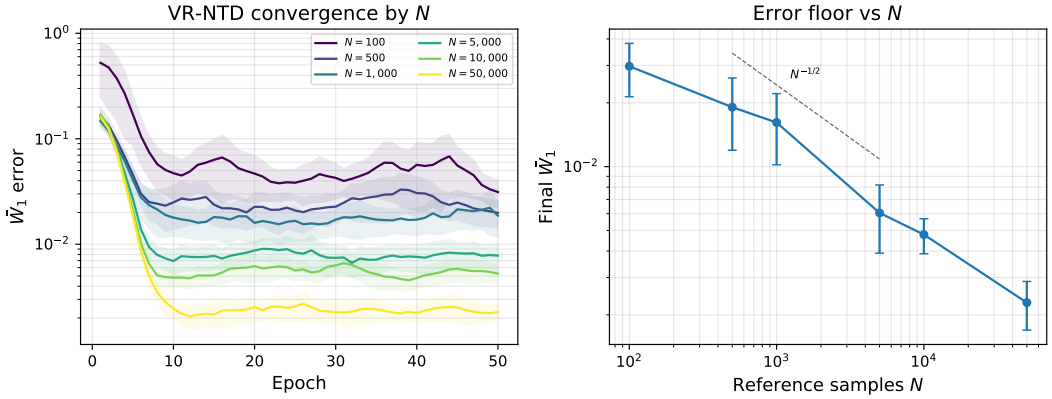


Figure 7 Sensitivity analysis of reference sample size N, illustrated with VR−NTD ($K = 1000$, 5 seeds, mean ± one std). Left: $\bar{W}_1$ versus epoch for different N. Right: final $\bar{W}_1$ versus N on a log-log scale with $N^{-1/2}$ reference line.

F.3. Effect of Variance Reduction

We compare vanilla CTD and VR−CTD directly under fixed-distribution and Markovian sampling, using the inventory model of Section 5 with $\lambda = 10$ and policies $S = 5$ and $S = 15$. Both methods use 51 equally spaced atoms and are initialized with uniform atom probabilities. For each of 20 independent seeds, the two methods receive the same sequence of samples. In the fixed-distribution setting, states are sampled uniformly from $\mathcal{S}$. In the Markovian setting, the trajectory starts at state S, and vanilla CTD updates at every transition, without burn-in or data drop.

We retain the step-size choices from Appendix F.1. For CTD, $\alpha_t = 1/[1 + \mu_{\min}(1 - \sqrt{\gamma})t/\max\{\log t, 1\}]$, with $\mu_{\min} = 1/21$ under fixed-distribution sampling and $\mu_{\pi,\min}$ in its place under Markovian sampling; here t counts all transitions, and $\mu_{\pi,\min}$ is taken over the recurrent states. For VR−CTD, $\alpha = 0.3$, and each of 50 epochs uses 10^4 samples to construct the reference operator

followed by 5×10^3 updates. Thus, both methods receive 7.5×10^5 samples in total, including all reference samples used by `VR-CTD`. The error is evaluated over recurrent states against return distributions computed by DP on the same 51-atom grid.

Figure 8 shows that `VR-CTD` reaches a lower-error region earlier under Markovian sampling. Under fixed-distribution sampling, its advantage is less pronounced, and vanilla `CTD` attains a smaller final error.

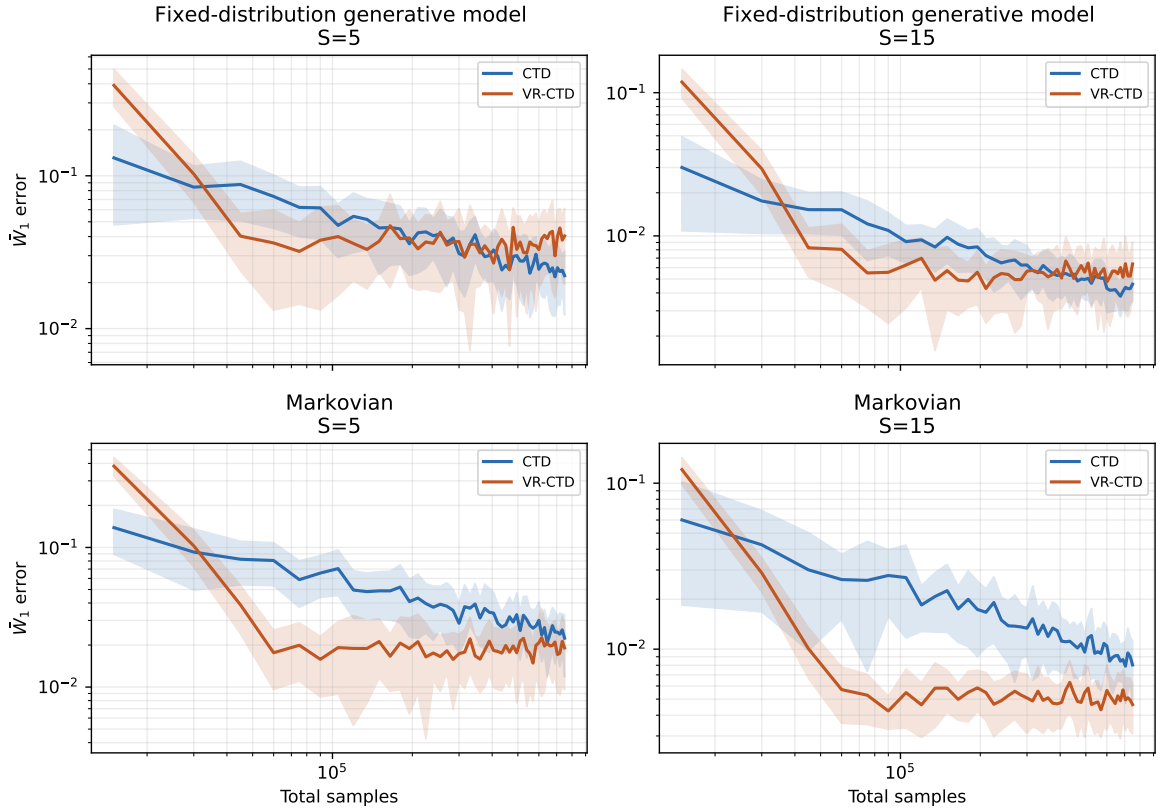


Figure 8 Direct comparison of vanilla `CTD` and `VR-CTD` under fixed-distribution sampling (top) and Markovian sampling without data drop (bottom). Curves show the mean error over 20 seeds, with shaded bands indicating one standard deviation. The horizontal axis counts all sampled transitions, including those used to construct the reference operator.