A Appendix

A.1 Proof of Theorem 4.2

The key to proving Theorem 4.2 is the use of Girsanov's theorem.

Lemma 1(Girsanov's theorem) For $t \in [0, T]$, let $\mathcal{L}_t = \int_0^t b_s \mathrm{d}B_s$ where B is a Q-Brownian motion. Assume $\mathbb{E}_Q \int_0^T \|b_s\|^2 \mathrm{d}s < \infty$. Then, $\mathcal{L}$ is a Q-martingale in $L^2(Q)$. Moreover, if

$$\mathbb{E}_Q \mathcal{E}(\mathcal{L})_T = 1, where \mathcal{E}(\mathcal{L})_t := \exp\left(\int_0^t b_s \mathrm{d}B_s - \frac{1}{2}\int_0^t \|b_s\|^2 \mathrm{d}s\right),$$

then $\mathcal{E}(\mathcal{L})$ is also a Q-martingale and the process

$$t \mapsto B_t - \int_0^t b_s \mathrm{d}s$$

is a Brownian motion under $P := \mathcal{E}_T Q$, the probability distribution with density $\mathcal{E}(\mathcal{L})_T$ w.r.t. Q.

In the proof below, for any fixed $t \in \{2, \cdots, H\}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, let $p_{\mathrm{data}} = T_t(\cdot|s, a)$, we denote the path measure of the backward SDE 7 and forward SDE 6 (they share the same solution) to be $Q_T := Q_T(\cdot|t, s, a)$. Denote the path measure generated from the conditional likelihood training to be $P_T := P_T(\cdot|t, s, a, \phi, \theta)$. Denote $\widehat{\mathbf{Z}}(\cdot, \cdot, \tilde{\theta}) := \widehat{\mathbf{Z}}$ and $\mathbf{Z} := \mathbf{Z}(\cdot, \cdot, \tilde{\phi})$. By Assumption (1)∼(6), the following analysis holds for any given $t = 2, \cdots, H$ and $(s_{t-1}, a_{t-1}) \in \mathcal{S} \times \mathcal{A}$.

Theorem 4.2 For any $t = 2, \cdots, H$ and any $(s_{t-1}, a_{t-1}) \in \mathcal{S} \times \mathcal{A}$, suppose the diffusion time $T \geq \max\{1, \frac{1}{\tau^2}\}$, we have

$$\mathrm{TV}(\widehat{T}_t(\cdot|s, a), T_t(\cdot|s, a)) \lesssim (\epsilon + M^3 L^{3/2} T\sqrt{dh} + LMmh)\sqrt{T}.$$

Proof. We start by proving

$$\sum_{k=0}^{N-1} \mathbb{E}_{Q_T} \int_{kh}^{(k+1)h} \left\|\widehat{\mathbf{Z}}(kh, \mathbf{X}_{kh}) - c\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(r, \mathbf{X}_r)\right\|^2 \mathrm{d}r \lesssim (\epsilon^2 + M^6 L^3 dh + M^2 h^2 m^2)T.$$

For $r \in [kh, (k+1)h]$, we can decompose

$$\begin{aligned}
&\mathbb{E}_{Q_T}[\left\|\widehat{\mathbf{Z}}(kh, \mathbf{X}_{kh}) - c\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(r, \mathbf{X}_r)\right\|^2] \\
\lesssim &\mathbb{E}_{Q_T}[\left\|\widehat{\mathbf{Z}}(kh, \mathbf{X}_{kh}) - c\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(kh, \mathbf{X}_{kh})\right\|^2] \\
&+ \mathbb{E}_{Q_T}[\left\|g\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(kh, \mathbf{X}_{kh}) - g\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(r, \mathbf{X}_{kh})\right\|^2] \\
&+ \mathbb{E}_{Q_T}[\left\|g\nabla_{\boldsymbol{x}} \log \widehat{Psi}(r, \mathbf{X}_{kh}) - g\nabla_{\boldsymbol{x}} \log \widehat{Psi}(r, \mathbf{X}_r)\right\|^2] \\
\lesssim &\epsilon^2 + \mathbb{E}_{Q_T}\left\|g\nabla_{\boldsymbol{x}} \log\left(\frac{\widehat{\Psi}(kh, \mathbf{X}_{kh})}{\widehat{\Psi}(r, \mathbf{X}_{kh})}\right)\right\|^2 + M^2 L^2 \mathbb{E}_{Q_T} \|\mathbf{X}_{kh} - \mathbf{X}_r\|^2
\end{aligned}$$

Notice that if $S : \mathbb{R}^d \to \mathbb{R}^d$ is the mapping $S(x) = \exp(-(r - kh))x$, then $\widehat{\Psi}(T - kh, \cdot) = S(\widehat{\Psi}(T - r, \cdot) * \mathcal{N}(0, 1 - \exp(-2(r - kh))))$. We can use Lemma 2 with $\alpha = \exp(r - kh) = 1 + O(h)$ and $\sigma^2 = 1 - \exp(-2(r - kh)) = O(h)$ and obtain

$$\mathbb{E}_{Q_T}\left\|g\nabla_{\boldsymbol{x}} \log\left(\frac{\widehat{\Psi}(kh, \mathbf{X}_{kh})}{\widehat{\Psi}(r, \mathbf{X}_{kh})}\right)\right\|^2 \lesssim M^2(L^2 dh + L^2 h^2 \|\mathbf{X}_{kh}\|^2 + L^2 h^2 \left\|\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(r, \mathbf{X}_{kh})\right\|^2).$$

Also we have

$$\left\|\nabla_{\boldsymbol{x}}\widehat{Psi}(r,\mathbf{X}_{kh})\right\|^2 \leq \left\|\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(r,\mathbf{X}_r)\right\|^2 + \left\|\nabla_{\boldsymbol{x}}\log\widehat{Psi}(r,\mathbf{X}_{kh}) - \nabla_{\boldsymbol{x}}\log\widehat{\Psi}(r,\mathbf{X}_r)\right\|^2$$
$$\leq \left\|\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(r,\mathbf{X}_r)\right\|^2 + L^2\left\|\mathbf{X}_{kh}-\mathbf{X}_r\right\|^2.$$

So

$$\mathbb{E}_{Q_T}[\left\|\widehat{\mathbf{Z}}(kh,\mathbf{X}_{kh}) - c\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(r,\mathbf{X}_r)\right\|^2]$$
$$\lesssim \epsilon^2 + M^2(L^2dh + L^2h^2\mathbb{E}_{Q_T}\left\|\mathbf{X}_{kh}\right\|^2 + L^2h^2\mathbb{E}_{Q_T}\left\|\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(T-r,\mathbf{X}_r)\right\|^2 + L^2\mathbb{E}_{Q_T}\left\|\mathbf{X}_{kh}-\mathbf{X}_r\right\|^2).$$

Using L-smoothness of $\nabla_{\boldsymbol{x}}\log\widehat{\Psi}$ and $\nabla_{\boldsymbol{x}}\log\Psi$, by (Vempala & Wibisono (2019), Lemma 9) and (Chen et al. (2023a), Lemma 10) , we have

$$\mathbb{E}\left\|\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(r,X_r)\right\|^2 \leq Ld,$$

and

$$\mathbb{E}\left\|\nabla_{\boldsymbol{x}}\log\Psi(r,X_r)\right\|^2 \leq Ld.$$

On the other hand, for $0 \leq s < r$, by the forward process 6, we have

$$\mathbb{E}_{Q_T}\left\|\mathbf{X}_r-\mathbf{X}_s\right\|^2 = \mathbb{E}_{Q_T}[\left\|\int_s^r (f + c^2\nabla_{\boldsymbol{x}}\log\Psi(r,\mathbf{X}_r))\mathrm{d}r + c(B_r - B_s)\right\|^2]$$
$$\lesssim (r-s)\int_s^r \mathbb{E}\left\|f + c^2\nabla_{\boldsymbol{x}}\log\Psi(r,\mathbf{X}_r)\right\|^2 \mathrm{d}r + M(r-s)d$$
$$\lesssim (r-s)^2M^2 + (r-s)^2M^4Ld + M(r-s)d$$

As a result, we get

$$\mathbb{E}\left\|\mathbf{X}_{kh}\right\|^2 \leq \mathbb{E}\left\|\mathbf{X}_0\right\|^2 + T^2M^2 + T^2M^4Ld + MTd$$
$$\leq m^2 + T^2M^2 + T^2M^4Ld + MTd$$

and

$$\mathbb{E}\left\|\mathbf{X}_{kh}-\mathbf{X}_r\right\|^2 \leq h^2M^2 + h^2M^4Ld + Mhd.$$

Combining the results above, we get that

$$\mathbb{E}_{Q_T}[\left\|\widehat{\mathbf{Z}}(kh,\mathbf{X}_{kh}) - c\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(r,\mathbf{X}_r)\right\|^2]$$
$$\lesssim \epsilon^2 + M^2[L^2dh + L^2h^2(m^2 + T^2M^2 + T^2M^4Ld + MTd) + L^2h^2Ld + L^2(h^2M^2 + h^2M^4Ld + Mhd)]$$
$$\lesssim \epsilon^2 + M^6L^3T^2dh + M^2L^2h^2m^2.$$

(Suppose $T \geq 1$ and $h \lesssim \frac{1}{L}$) So we have

$$\sum_{k=0}^{N-1}\mathbb{E}_{Q_T}\int_{kh}^{(k+1)h}\left\|\widehat{\mathbf{Z}}(kh,\mathbf{X}_{kh}) - c\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(r,\mathbf{X}_r)\right\|^2\mathrm{d}r \lesssim (\epsilon^2 + M^6L^3T^2dh + M^2L^2h^2m^2)T.$$

Now we apply an approximation argument to use Girsanov's theorem and prove Theorem 4.2.

For $r \in [0,T]$, let $\mathcal{L}_r = \int_0^r b_s\mathrm{d}B_s$ where B is a Q_T-Brownian motion. For $r \in [kh,(k+1)h]$, define

$$b_r = -c\nabla_{\boldsymbol{x}}\log\widehat{\Psi}(r,\mathbf{X}_r) + \widehat{\mathbf{Z}}(kh,\mathbf{X}_{kh}).$$

From above,

$$\mathbb{E}_{Q_T}\int_0^T\left\|b_s\right\|^2\mathrm{d}s \lesssim (\epsilon^2 + M^6L^3T^2dh + M^2L^2h^2m^2)T < \infty,$$

using (Le Gall (2016), Proposition 5.11), $(\mathcal{E}(\mathcal{L})_r)_{r\in[0,T]}$ (see the definition in Lemma 1) is a local martingale (see Definition 1). Therefore, there exists a non-decreasing sequence of stopping time $T_n \uparrow T$ such that $(\mathcal{E}(\mathcal{L})_{r\wedge T_n})_{r\in[0,T]}$ is a martingale. Notice that $\mathcal{E}(\mathcal{L})_{r\wedge T_n} = \mathcal{E}(\mathcal{L}^n_r)$ where $\mathcal{L}^n_r = \mathcal{L}_{r\wedge T_n}$. Since $\mathcal{E}(\mathcal{L}^n_r)_{r\in[0,T]}$ is a martingale, we have

$$\mathbb{E}_{Q_T}\mathcal{E}(\mathcal{L}^n)_T = \mathbb{E}_{Q_T}\mathcal{E}(\mathcal{L}^n)_0 = 1,$$

so that $\mathbb{E}_{Q_T}\mathcal{E}(\mathcal{L})_{T_n} = 1$.

Apply Girsanov's theorem to $\mathcal{L}^n_r = \int_0^r b_s \mathbf{1}_{[0,T_n]}(s)\mathrm{d}B_s$ where B is a Q_T-Brownian motion and get that under $P^n := \mathcal{E}(\mathcal{L})_T Q_T$, there exists a Brownian motion β^n such that for $r \in [0,T]$,

$$dB_r = \left[-c\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(r, \mathbf{X}_r) + \widehat{\mathbf{Z}}(kh, \mathbf{X}_{kh})\right] \mathbf{1}_{[0,T_n]}(r)\mathrm{d}r + \mathrm{d}\beta^n_r.$$

By the backward SDE 7, under Q_T we have

$$\mathrm{d}\mathbf{X}_r = -[f - c^2\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(r, \mathbf{X}_r)]\mathrm{d}r + c\mathrm{d}B_r,\ \mathbf{X}_0 \sim p_{\text{prior}}.$$

The equation still holds P^n-a.s. since $P^n \ll Q_T$. Combining the two equations above then we obtain that P^n-a.s.,

$$\mathrm{d}\mathbf{X}_r = \left[-f + c\widehat{\mathbf{Z}}(kh, \mathbf{X}_{kh})\right] \mathbf{1}_{[0,T_n]}(r)\mathrm{d}r + \left[-f + c^2\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(T-r, \mathbf{X}_r)\right] \mathbf{1}_{[T_n,T]}(r)\mathrm{d}r + c\mathrm{d}\beta^n_r,\ \mathbf{X}_0 \sim p_{\text{prior}}.$$

i.e. path measure P^n is the solution to the above SDE. So we have

$$\begin{aligned}\mathrm{KL}(Q_T|P^n) =& \mathbb{E}_{Q_T} \log \mathcal{E}(\mathcal{L})^{-1}_{T_n} = \mathbb{E}_{Q_T}[-\mathcal{L}_{T_n} + \frac{1}{2}\int_0^{T_n} \|b_s\|^2 \mathrm{d}s] = \mathbb{E}_{Q_T}\frac{1}{2}\int_0^{T_n} \|b_s\|^2 \mathrm{d}s \\ \leq& \mathbb{E}_{Q_T}\frac{1}{2}\int_0^{T} \|b_s\|^2 \mathrm{d}s \lesssim (\epsilon^2 + M^6L^3T^2dh + M^2L^2h^2m^2)T\end{aligned}$$

where we used that $\mathbb{E}_{Q_T}\mathcal{L}_{T_n} = 0$ because $\mathcal{L}$ is a Q_T-martingale and T_n is a bounded stopping time.(Le Gall (2016), Corollary 3.23)

Consider a coupling of $(P^n)_{n\in\mathbb{N}}, P_T$: a sequence of stochastic process $(\mathbf{X}^n)_{n\in\mathbb{N}}$ over the same proability space, a stochastic process $\mathbf{X}$ and a single Brownian motion W over the same space s.t.

$$\mathrm{d}\mathbf{X}^n_r = \left[-f + c\widehat{\mathbf{Z}}(kh, \mathbf{X}^n_{kh})\right] \mathbf{1}_{[0,T_n]}(r)\mathrm{d}r + \left[-f + c^2\nabla_{\boldsymbol{x}} \log \widehat{\Psi}(T-r, \mathbf{X}^n_r)\right] \mathbf{1}_{[T_n,T]}(r)\mathrm{d}r + c\mathrm{d}W_r,$$

$$\mathrm{d}\mathbf{X}_r = \left[-f + c\widehat{\mathbf{Z}}(kh, \mathbf{X}_{kh})\right] \mathrm{d}r + c\mathrm{d}W_r,$$

$$\mathbf{X}_0 = \mathbf{X}^n_0 \sim p_{\text{prior}}.$$

By definition of P^n and P_T, the distribution of $\mathbf{X}^n$ ($\mathbf{X}$) is P^n (P_T).

Let $\delta > 0$ and consider the map $\pi_\delta : \mathcal{C}([0,T];\mathbb{R}^d) \to \mathcal{C}([0,T];\mathbb{R}^d)$ defined by

$$\pi_\delta(\omega)(r) := \omega(r \wedge (T-\delta)).$$

Notice that $\mathbf{X}^n_r = \mathbf{X}_r$ for every $r \in [0, T_n]$, using Lemma 3, we have $\pi_\delta(\mathbf{X}^n) \to \pi_\delta(\mathbf{X})$ a.s., uniformly over $[0,T]$. Therefore, $\pi_{\delta\#}P^n \to \pi_{\delta\#}P_T$ weakly. Using the lower semicontinuity of the KL divergence and the data-processing inequality (Amb (2005), Lemma 9.4.3 and Lemma 9.4.5), we get

$$\begin{aligned}\mathrm{KL}((\pi_\delta)_\# Q_T|(\pi_\delta)_\# P_T) \leq& \liminf_{n\to\infty} \mathrm{KL}((\pi_\delta)_\# Q_T|(\pi_\delta)_\# P^n) \\ \leq& \liminf_{n\to\infty} \mathrm{KL}(Q_T|P^n) \\ \lesssim& (\epsilon^2 + M^6L^3T^2dh + M^2L^2h^2m^2)T.\end{aligned}$$

Finally, using Lemma 4, $\pi_\delta(\omega) \to \omega$ as $\delta \to 0$ uniformly over $[0,T]$. Therefore, using (Amb (2005), Corollary 9.4.6), $\mathrm{KL}((\pi_\delta)_\# Q_T|(\pi_\delta)_\# P_T) \to \mathrm{KL}(Q_T|P_T)$ as $\delta \to 0$. Since the marginal distribution at $T=0$ of Q_T is $T_t(\cdot|s,a)$ and the marginal distribution at $T=0$ of P_T is $\widehat{T}_t(\cdot|s,a)$, by data processing inequality we ultimately have

$$\mathrm{KL}(T_t(\cdot|s,a)|\widehat{T}_t(\cdot|s,a)) \lesssim (\epsilon^2 + M^6L^3T^2dh + M^2L^2h^2m^2)T.$$

We conclude the proof using Pinsker's inequality ($\mathrm{TV}^2 \leq \mathrm{KL}$). □

A.2 PROOF OF THEOREM 4.1

In this section, we give the proof of Theorem 4.1, which is our main theorem.

Theorem 4.1 Under Assumptions (1)-(6), let $\widehat{V}^\pi$ be the output of CDSB estimator, and suppose that the step size $h := \frac{T}{N}$ satisfies $h \lesssim \frac{1}{L}$, where $L \geq 1$. Suppose the diffusion time $T \geq \max\{1, \frac{1}{\tau^2}\}$, then it holds that

$$|\widehat{V}^\pi - V^\pi| \lesssim R_{\max}\tau^2 H^2(\epsilon + M^3 L^{3/2} T\sqrt{dh} + LMmh)\sqrt{T}. \tag{12}$$

Proof. We have

$$V^\pi = \sum_{t=1}^{H} \int_{\mathcal{A}} \int_{\mathcal{S}^t} R_t(s_t, a_t)\pi(a_t|s_t) P_t^\pi(s_t|s_{t-1}) \cdots \widehat{P}_2^\pi(s_2|s_1) d_0(s_1) \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t,$$

and

$$\widehat{V}^\pi = \sum_{t=1}^{H} \int_{\mathcal{A}} \int_{\mathcal{S}^t} \widehat{R}_t(s_t, a_t)\pi(a_t|s_t) \widehat{P}_t^\pi(s_t|s_{t-1}) \cdots \widehat{P}_2^\pi(s_2|s_1) d_0(s_1) \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t.$$

By Theorem 4.2, assumption (6) and the definition of total-variation norm, for all $s \in \mathcal{S}$ and all $t \in \{2, \cdots, T\}$, we have

$$\begin{aligned} \int_{\mathcal{S}} |P_t^\pi(s'|s) - \widehat{P}_t^\pi(s'|s)| \mathrm{d}s' &= \int_{\mathcal{S}} | \int_{\mathcal{A}} \pi(a|s)(T_t(s'|s, a) - \widehat{T}_t(s'|s, a)) \mathrm{d}a | \mathrm{d}s \\ &\lesssim \tau(\epsilon + M^3 L^{3/2} T\sqrt{dh} + LMmh)\sqrt{T} =: \delta_0, \end{aligned}$$

$$\int_{\mathcal{A}} |\widehat{R}_t(s, a) - R_t(s, a)| \mathrm{d}a \leq \epsilon \lesssim \delta_0,$$

since $T \geq \max\{1, \frac{1}{\tau^2}\}$.

So

$$\begin{aligned} &|\widehat{V}^\pi - V^\pi| \\ \leq &\tau \sum_{t=1}^{H} \left| \int_{\mathcal{A}} \int_{\mathcal{S}^t} \widehat{R}_t(s_t, a_t) \widehat{P}_t^\pi(s_t|s_{t-1}) \cdots \widehat{P}_2^\pi(s_2|s_1) d_0(s_1) \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t - \right. \\ &\left. \int_{\mathcal{A}} \int_{\mathcal{S}^t} R_t(s_t, a_t) P_t^\pi(s_t|s_{t-1}) \cdots P_2^\pi(s_2|s_1) d_0(s_1) \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t \right| \\ \leq &\tau \sum_{t=1}^{H} \int_{\mathcal{A}} \int_{\mathcal{S}^t} \left| \widehat{R}_t(s_t, a_t) \widehat{P}_t^\pi(s_t|s_{t-1}) \cdots \widehat{P}_2^\pi(s_2|s_1) d_0(s_1) - R_t(s_t, a_t) P_t^\pi(s_t|s_{t-1}) \cdots P_2^\pi(s_2|s_1) d_0(s_1) \right| \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t \\ \leq &\tau \sum_{t=1}^{H} \int_{\mathcal{A}} \int_{\mathcal{S}^t} \left(\left| \left(\widehat{R}_t(s_t, a_t) - R_t(s_t, a_t) \right) \widehat{P}_t^\pi(s_t|s_{t-1}) \cdots \widehat{P}_2^\pi(s_2|s_1) d_0(s_1) \right| \right. \\ &\left. + \left| R_t(s_t, a_t) \left(\widehat{P}_t^\pi(s_t|s_{t-1}) \cdots \widehat{P}_2^\pi(s_2|s_1) - P_t^\pi(s_t|s_{t-1}) \cdots P_2^\pi(s_2|s_1) \right) d_0(s_1) \right| \right) \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t \\ &\cdots \\ \leq &\tau \sum_{t=1}^{H} \left(\int_{\mathcal{A}} \int_{\mathcal{S}^t} \left| \widehat{R}_t(s_t, a_t) - R_t(s_t, a_t) \right| |P_t^\pi(s_t|s_{t-1})| \left| P_{t-1}^\pi(s_{t-1}|s_{t-2}) \right| \cdots |P_2^\pi(s_2|s_1) d_0(s_1)| \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t \right. \\ &\left. + \cdots + \int_{\mathcal{A}} \int_{\mathcal{S}^t} \left| \widehat{R}_t(s_t, a_t) - R_t(s_t, a_t) \right| \left| \widehat{P}_t^\pi(s_t|s_{t-1}) - P_t^\pi(s_t|s_{t-1}) \right| \cdots \left| \widehat{P}_2^\pi(s_2|s_1) - P_2^\pi(s_2|s_1) \right| d_0(s_1) \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t \right) \end{aligned}$$

The summation above contains $2^{t-1} - 1$ items, each term $|\cdot|$ in the integration of each item is either $|\widehat{P}_j^\pi(s_j|s_{j-1}) - P_j^\pi(s_j|s_{j-1})|$ ($|\widehat{R}_t(s_t, a_t) - R_t(s_t, a_t)|$) or $|P_j^\pi(s_j|s_{j-1})|$ ($|R_t(s_t, a_t)|$) for $j = 2, \cdots, t$, but not all $|P_t^\pi(s_j|s_{j-1})|$. Relax all the $|\widehat{P}_j^\pi(s_j|s_{j-1}) - P_j^\pi(s_j|s_{j-1})|$ and $|\widehat{R}_t(s_t, a_t) -$

$R_t(s_t, a_t)|$ to their uniform upper bound (with respect to s_{j-1} and s_t) δ_0. Since P_j^π are non-negative for $t = 1, \cdots, t-1$, the terms of each item in the summation are then relaxed to

$$\delta_0^{t-1-k} \int_{\mathcal{A}} \int_{\mathcal{S}\times\cdots\times\mathcal{S}} R_t(s_t, a_t) P_{j_k}^\pi(s_{j_k}|s_{j_k-1}) \cdots P_{j_1}^\pi(s_{j_1}|s_{j_1-1}) d_0(s_1) \mathrm{d}s_t \cdots \mathrm{d}s_1 \mathrm{d}a_t,$$

or

$$\delta_0^{t-k} \int_{\mathcal{S}\times\cdots\times\mathcal{S}} R_t(s_t, a_t) P_{j_k}^\pi(s_{j_k}|s_{j_k-1}) \cdots P_{j_1}^\pi(s_{j_1}|s_{j_1-1}) d_0(s_1) \mathrm{d}s_t \cdots \mathrm{d}s_1,$$

where $1 \le k \le t-1$, $j_1 < \cdots < j_k$ and $\{j_1, \cdots, j_k\} \in \{2, \cdots, t\}$. By the definition of P_j^π, it's easy to verify that

$$\int_{\mathcal{S}^t} P_{j_k}^\pi(s_{j_k}|s_{j_k-1}) \cdots P_{j_1}^\pi(s_{j_1}|s_{j_1-1}) d_0(s_1) \mathrm{d}s_t \cdots \mathrm{d}s_1 = 1$$

and

$$\int_{\mathcal{A}} \int_{\mathcal{S}^t} R_t(s_t, a_t) P_{j_k}^\pi(s_{j_k}|s_{j_k-1}) \cdots P_{j_1}^\pi(s_{j_1}|s_{j_1-1}) d_0(s_1) \mathrm{d}s_t \cdots \mathrm{d}s_1 \mathrm{d}a_t \le R_{\max}$$

for any $1 \le k \le t-1$, $j_1 < \cdots < j_k$ and $\{j_1, \cdots, j_k\} \in \{2, \cdots, t\}$. So that the summation

$$\begin{aligned}
&\int_{\mathcal{A}} \int_{\mathcal{S}^t} \left|\widehat{R}_t(s_t, a_t) - R_t(s_t, a_t)\right| |P_t^\pi(s_t|s_{t-1})| \left|P_{t-1}^\pi(s_{t-1}|s_{t-2})\right| \cdots |P_2^\pi(s_2|s_1) d_0(s_1)| \,\mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t \\
&+ \cdots + \int_{\mathcal{A}} \int_{\mathcal{S}^t} \left|\widehat{R}_t(s_t, a_t) - R_t(s_t, a_t)\right| \left|\widehat{P}_t^\pi(s_t|s_{t-1}) - P_t^\pi(s_t|s_{t-1})\right| \cdots \left|\widehat{P}_2^\pi(s_2|s_1) - P_2^\pi(s_2|s_1)\right| d_0(s_1) \mathrm{d}s_1 \cdots \mathrm{d}s_t \mathrm{d}a_t \\
\le& R_{\max} \left(\delta_0^t + t\delta_0^{t-1} + \cdots + t\delta_0\right) \\
=& R_{\max} \left((\delta_0 + 1)^t - 1\right) \\
\le& R_{\max} \left((\delta_0 + 1)^H - 1\right).
\end{aligned}$$

Noting that $\delta_0 = \tau(\epsilon + M^3 L^{3/2} T\sqrt{dh} + LMmh)\sqrt{T}$, so for ϵ and h that is sufficiently small, there exists a universal constant η, such that

$$|\widehat{V}^\pi - V^\pi| \le H\tau H R_{\max} \eta \delta_0 \lesssim R_{\max} \tau^2 H^2 (\epsilon + M^3 L^{3/2} T\sqrt{dh} + LMmh)\sqrt{T},$$

which finishes the proof of Theorem 4.1. □

A.3 Auxiliary Lemmas

In this section, we presents the definitions and auxiliary lemmas which are used to prove Theorem 4.2.

Definition 1 A local martingale $(L_t)_{t\in[0,T]}$ is a stochastic process such that there exists a sequence of non-decreasing stopping times $T_n \to T$ such that $L^n = (L_{t\wedge T_n})_{t\in[0,T]}$ is a martingale.

Lemma 2(Chen et al. (2023a), Lemma 16) Let $0 < \zeta < 1$. Suppose that $\mathbf{M}_0, \mathbf{M}_1 \in \mathbb{R}^{2d\times 2d}$ are two matrices, where $\mathbf{M}_1$ is symmetric. Also, assume that $\|\mathbf{M}_0 - \mathbf{I}_{2d}\|_{op} \le \zeta$, so that $\mathbf{M}_0$ is invertible. Let $\mathbf{q} = \exp(-\mathbf{H})$ be a probability density on $\mathbb{R}^{2d}$ such that $\nabla\mathbf{H}$ is L-lipschitz with $L \le \frac{1}{4\|\mathbf{M}_1\|_{op}}$, it holds that

$$\left\|\nabla \log \frac{(\mathbf{M}_0)_{\#}\mathbf{q} * \mathcal{N}(0, \mathbf{M}_1)}{\mathbf{q}}(\theta)\right\| \lesssim L\sqrt{\|\mathbf{M}_1\|_{op}\, d} + L\zeta\,\|\theta\| + (\zeta + L\,\|\mathbf{M}_1\|_{op})\,\|\nabla\mathbf{H}(\theta)\|.$$

The following lemmas are very straightforward, so the proof is omitted.

Lemma 3 Consider $f_n, f : [0,T] \to \mathbb{R}^d$ s.t. there exists an increasing sequence $(T_n)_{n\in\mathbb{N}} \subseteq [0,T]$ satisfying $T_n \to T$ as $n \to \infty$ and $f_n(t) = f(t)$ for every $t \le T_n$. Then for every $\epsilon > 0$, $f_n \to f$ uniformly over $[0, T-\epsilon]$. In particular, $f_n(\cdot \wedge T - \epsilon) \to f(\cdot \wedge T - \epsilon)$ uniformly over $[0,T]$.

Lemma 4 Consider $f : [0,T] \to \mathbb{R}^d$ continuous, and $f_\epsilon : [0,T] \to \mathbb{R}^d$ s.t. $f_\epsilon(r) = f(r \wedge (T-\epsilon))$ for $\epsilon > 0$. Then $f_\epsilon \to f$ uniformly over $[0,T]$ as $\epsilon \to 0$.

A.4 EXPERIMENTS

We have made our code publicly available[1].

[1] https://anonymous.4open.science/r/bridge_OPE-302D/