Supplementary Materials

A Appendix

A.1 Total Variations, Bellman Contraction and Symmetry Bridge

Fact 1. *Let m^1 and m^2 be probability measures on $\mathbb{R}^n$ whose singular continuous parts are zero. Decompose m^1 and m^2 into their absolutely continuous and discrete parts: $m^1 = m^1_a + m^1_d$, $m^2 = m^2_a + m^2_d$. Then*

$$D_{TV}(m^1||m^2) = \frac{1}{2}\big(\left\|m^1_a - m^2_a\right\|_1 + \left\|m^1_d - m^2_d\right\|_1 \big) \triangleq \frac{1}{2}\left\|m^1 - m^2\right\|_1 .$$

Proof. (Hewitt and Ross, 1963, Theorem 19.20) implies $D_{TV}(m^1||m^2) = D_{TV}(m^1_a||m^2_a) + D_{TV}(m^1_d||m^2_d)$, so Fact 1 is proved by combining (Hewitt and Ross, 1963, Theorem 19.20) and that TV distance = half of ℓ_1 norm for absolutely continuous or discrete measures.

To avoid using a big hammer, we provide an alternative proof by revising the usual proof of "TV distance = half of ℓ_1 norm" with the Lebesgue decomposition: $m = m_a + m_d$. Since m^1_a and m^2_a are absolutely continuous w.r.t. Lebesgue measure, let d^1, d^2 be the corresponding probability density functions.

Let $B = B_a \cup B_d$ where $B_a = \{x \in \text{Supp}(m^1_a) \cup \text{Supp}(m^2_a) : \ d^1(x) \geq d^2(x)\}$, $B_d = \{x \in \text{Supp}(m^1_d) \cup \text{Supp}(m^2_d) : \ m^1_d(x) \geq m^2_d(x)\}$. Since m_a and m_d are mutually singular, we know

$$m^1_a(B_d) = m^2_a(B_d) = 0 = m^1_d(B_a) = m^2_d(B_a) \tag{1}$$

Also, the complement operation implies

$$m^2(A^c) - m^1(A^c) = 1 - m^2(A) - 1 + m^1(A) = m^1(A) - m^2(A), \quad \text{for any measurable set } A \tag{2}$$

Hence we have an important result

$$\begin{aligned}
m^1(B) - m^2(B) =& m^1_a(B) - m^2_a(B) + m^1_d(B) - m^2_d(B) \\
\overset{(1)}{=}& m^1_a(B_a) - m^2_a(B_a) + m^1_d(B_d) - m^2_d(B_d) \\
\overset{(2)}{=}& \frac{1}{2}\big[m^1_a(B_a) - m^2_a(B_a) + m^2_a(B^c_a) - m^1_a(B^c_a) + m^1_d(B_d) - m^2_d(B_d) + m^2_d(B^c_d) - m^1_d(B^c_d)\big] \\
=& \frac{1}{2}\Big[\int_{B_a} d^1(x) - d^2(x)dx + \int_{\text{Supp}(m^1_a)\cup\text{Supp}(m^2_a)\setminus B_a} d^2(x) - d^1(x)dx \\
& + \sum_{x\in B_d} m^1_d(x) - m^2_d(x) + \sum_{x\in\text{Supp}(m^1_d)\cup\text{Supp}(m^2_d)\setminus B_d} m^2_d(x) - m^1_d(x)\Big] \\
=& \frac{1}{2}\Big[\int_{\text{Supp}(m^1_a)\cup\text{Supp}(m^2_a)} \Big|d^1(x) - d^2(x)\Big|dx + \sum_{x\in\text{Supp}(m^1_d)\cup\text{Supp}(m^2_d)} \Big|m^1_d(x) - m^2_d(x)\Big|\Big] \\
=& \frac{1}{2}\big(\left\|m^1_a - m^2_a\right\|_1 + \left\|m^1_d - m^2_d\right\|_1 \big) \triangleq \frac{1}{2}\left\|m^1 - m^2\right\|_1 .
\end{aligned} \tag{3}$$

(i) By definition of TV distance, we get

$$D_{TV}(m^1||m^2) \geq |m^1(B) - m^2(B)| = m^1(B) - m^2(B) \overset{(3)}{=} \frac{1}{2}\left\|m^1 - m^2\right\|_1$$

(ii) For any measurable set A in $\mathbb{R}^n$, we know

$$m^1(A) - m^2(A) = \left[m^1(A \cap B) - m^2(A \cap B)\right] + \left[m^1(A \cap B^c) - m^2(A \cap B^c)\right]$$

By definition of B, the first term is nonnegative while the second term is nonpositive; therefore

$$\begin{aligned}
|m^1(A) - m^2(A)| &\leq \max\left\{m^1(A \cap B) - m^2(A \cap B),\ m^2(A \cap B^c) - m^1(A \cap B^c)\right\} \\
&\leq \max\left\{m^1(B) - m^2(B),\ m^2(B^c) - m^1(B^c)\right\} \\
&\overset{(2)}{=} m^1(B) - m^2(B) \overset{(3)}{=} \frac{1}{2}\left\|m^1 - m^2\right\|_1
\end{aligned}$$

Taking a supremium over A, we arrive at

$$D_{TV}(m^1||m^2) \leq \frac{1}{2}\left\|m^1 - m^2\right\|_1.$$

Combining (i) and (ii), the result follows. □

Due to Fact 1, in the following we will treat TV distance as the half of ℓ_1 norm. Also, to unify the operations in discrete and continuous parts, we will consider "generalized" probability density functions where Dirac delta function is included. Thus, Fact 1 is rephrased as

$$D_{TV}(m^1||m^2) = \frac{1}{2}\int |d^1(x) - d^2(x)|dx,$$

where d^1, d^2 are the generalized density functions of m^1 and m^2. This allows us to prove Fact 2:

Fact 2. *$B_{\pi,T}$ is a γ-contraction w.r.t. total variation distance.*

Proof. Let $p_1(s),\ p_2(s)$ be the density functions of some state distributions.

$$\begin{aligned}
D_{TV}(B_{\pi,T}(p_1)||B_{\pi,T}(p_2)) &= \frac{1}{2}\int \left|B_{\pi,T}(p_1(s)) - B_{\pi,T}(p_2(s))\right|ds \\
&= \frac{1}{2}\int \gamma\Big|\int T(s|s',a')\pi(a'|s')\big(p_1(s') - p_2(s')\big)ds'da'\Big|ds \\
&\leq \frac{\gamma}{2}\int T(s|s',a')\pi(a'|s')\left|p_1(s') - p_2(s')\right|ds'da'ds \\
&= \frac{\gamma}{2}\int \left|p_1(s') - p_2(s')\right|ds' = \gamma D_{TV}(p_1||p_2).
\end{aligned}$$

□

The advantages of working on contractions are their convergence and unique fixed-point properties [Theorem 1.1.]Conrad (2014).

Fact 3. *Let (X, d) be a complete metric space and $f:\ X \to X$ be a map such that*

$$d(f(x),\ f(x')) \leq cd(x, x')$$

for some $0 \leq c < 1$ and all $x,\ x' \in X$. Then f has a unique fixed point in X. Moreover, for any $x_0 \in X$ the sequence of the iterates x_0, $f(x_0)$, $f(f(x_0))$,... converges to the fixed point of f.

Fact 4. *The normalized state occupancy measure $\rho_{T,\gamma}^{\rho_0,\pi}(s)$ is a fixed point of the Bellman flow operator $B_{\pi,T}(\cdot)$.*

Proof.

$$\begin{aligned}
\rho_{T,\gamma}^{\rho_0,\pi}(s) =&(1-\gamma)\sum_{i=0}^{\infty}\gamma^i f_i(s|\rho_0,\pi,T)\\
=&(1-\gamma)f_0(s|\rho_0,\pi,T)+\gamma(1-\gamma)\sum_{i=0}^{\infty}\gamma^i f_{i+1}(s|\rho_0,\pi,T)\\
=&(1-\gamma)\rho_0(s)+\gamma(1-\gamma)\sum_{i=0}^{\infty}\gamma^i\int T(s|s',a')\pi(a'|s')f_i(s'|\rho_0,\pi,T)ds'da'\\
=&(1-\gamma)\rho_0(s)+\gamma\int T(s|s',a')\pi(a'|s')(1-\gamma)\sum_{i=0}^{\infty}\gamma^i f_i(s'|\rho_0,\pi,T)ds'da'\\
=&(1-\gamma)\rho_0(s)+\gamma\int T(s|s',a')\pi(a'|s')\rho_{T,\gamma}^{\rho_0,\pi}(s')ds'da'=B_{\pi,T}(\rho_{T,\gamma}^{\rho_0,\pi}(s)).
\end{aligned}$$

□

Together, Fact 2 and 3 imply the Bellman flow operator has a unique fixed point, and according to Fact 4, the unique fixed point is the state occupancy measure. The contraction and the fixed point properties are particularly useful for proving the symmetry bridge Lemma.

Lemma 1 (symmetry bridge)**.** *Let B_2 be a Bellman flow operator with fixed-point ρ_2. Let ρ_1 be another state distribution. If B_2 is a η-contraction w.r.t. some metric $\|\cdot\|$, then $\|\rho_1-\rho_2\|\leq\|\rho_1-B_2(\rho_1)\|/(1-\eta)$.*

Proof.

$$\begin{aligned}
\|\rho_1-\rho_2\|=\|\rho_1-B_2^{\infty}(\rho_1)\|&\leq\|\rho_1-B_2(\rho_1)\|+\sum_{i=1}^{\infty}\left\|B_2^i(\rho_1)-B_2^{i+1}(\rho_1)\right\|\\
&\leq\|\rho_1-B_2(\rho_1)\|+\sum_{i=1}^{\infty}\|\rho_1-B_2(\rho_1)\|\,\eta^i=\|\rho_1-B_2(\rho_1)\|/(1-\eta).
\end{aligned}$$

The first line uses the fixed-point property and the triangle inequlaity for the distance metric $\|\cdot\|$. The second line uses the contraction property. □

A.2 Error of Policies

Lemma 2 (Error w.r.t. TV Distance between Occupancy Measures)**.** *Let $\rho_1(s,a)$, $\rho_2(s,a)$ be two normalized occupancy measures of rollouts with discount factor γ. If $0\leq r(s,a)\leq r^{\max}$, then $|R(\rho_1)-R(\rho_2)|\leq D_{TV}(\rho_1||\rho_2)r^{\max}/(1-\gamma)$. where D_{TV} is the total variation distance.*

Proof.

$$\begin{aligned}
R(\rho_1)&=\frac{1}{1-\gamma}\int r(s,a)\rho_1(s,a)dsda\leq\frac{1}{1-\gamma}\int r(s,a)\max\big(\rho_1(s,a),\rho_2(s,a)\big)dsda\\
&=R(\rho_2)+\frac{1}{1-\gamma}\int r(s,a)\Big(\max\big(\rho_1(s,a),\rho_2(s,a)\big)-\rho_2(s,a)\Big)dsda\\
&\leq R(\rho_2)+\frac{r^{\max}}{1-\gamma}\int\max\big(\rho_1(s,a),\rho_2(s,a)\big)-\rho_2(s,a)dsda\\
&=R(\rho_2)+\frac{r^{\max}}{1-\gamma}\frac{1}{2}\|\rho_1-\rho_2\|_1=R(\rho_2)+\frac{r^{\max}}{1-\gamma}D_{TV}(\rho_1||\rho_2).
\end{aligned}$$

Because the TV distance is symmetric, we may interchange the roles of ρ_1 and ρ_2; thus we conclude that

$$|R(\rho_1)-R(\rho_2)|\leq D_{TV}(\rho_1||\rho_2)r^{\max}/(1-\gamma).$$

□

Theorem 1 (Error of Policies). *If $0 \le r(s,a) \le r^{\max}$ and the discrepancy in policies is $\epsilon^T_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\pi_D}_T}[D_{TV}(\pi_D(\cdot|s)||\pi(\cdot|s))]$, then $|R(\pi_D,T) - R(\pi,T)| \le \epsilon^T_{\pi_D,\pi} r^{\max}\Big(\frac{1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\Big)$.*

Proof. Let $B_{\pi_D,T}$, $B_{\pi,T}$ be Bellman flow operators whose fixed points are $\rho^{\pi_D}_T(s)$, $\rho^{\pi}_T(s)$, respectively. According to Lemma 2, we need to upper bound $D_{TV}(\rho^{\pi_D}_T(s,a)||\rho^{\pi}_T(s,a))$. Observe that

$$\begin{aligned} D_{TV}(\rho^{\pi_D}_T(s,a)||\rho^{\pi}_T(s,a)) &= \frac{1}{2}\int \Big|\rho^{\pi_D}_T(s,a) - \rho^{\pi}_T(s,a)\Big| dsda = \frac{1}{2}\int \Big|\rho^{\pi_D}_T(s)\pi_D(a|s) - \rho^{\pi}_T(s)\pi(a|s)\Big| dsda \\ &\le \frac{1}{2}\int \rho^{\pi_D}_T(s)\Big|\pi_D(a|s) - \pi(a|s)\Big| + \pi(a|s)\Big|\rho^{\pi_D}_T(s) - \rho^{\pi}_T(s)\Big| dsda \\ &= \epsilon^T_{\pi_D,\pi} + D_{TV}(\rho^{\pi_D}_T(s)||\rho^{\pi}_T(s)) \end{aligned} \tag{4}$$

As for the rest, by the properties of the Bellman flow operators, we have

$$\begin{aligned} D_{TV}(\rho^{\pi_D}_T(s)||\rho^{\pi}_T(s)) &\le \frac{1}{1-\gamma} D_{TV}(\rho^{\pi_D}_T(s)||B_{\pi,T}(\rho^{\pi_D}_T(s))) \\ &= \frac{1}{1-\gamma} D_{TV}(B_{\pi_D,T}(\rho^{\pi_D}_T(s))||B_{\pi,T}(\rho^{\pi_D}_T(s))) \\ &\le \frac{\gamma}{2(1-\gamma)}\int T(s|s',a')\Big|\pi_D(a'|s') - \pi(a'|s')\Big|\rho^{\pi_D}_T(s')ds'da'ds \\ &= \frac{\gamma}{1-\gamma}\epsilon^T_{\pi_D,\pi}, \end{aligned} \tag{5}$$

where the top two lines follows from the symmetry bridge property (Lemma 1) and the fixed-point property. Combining Eq. (4) and (5), we know $D_{TV}(\rho^{\pi_D}_T(s,a)||\rho^{\pi}_T(s,a)) \le \epsilon^T_{\pi_D,\pi}(1 + \frac{\gamma}{1-\gamma})$; therefore by Lemma 2,

$$|R(\pi_D,T) - R(\pi,T)| \le \epsilon^T_{\pi_D,\pi} r^{\max}\Big(\frac{1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\Big)$$

□

Corollary 1 (Error of Behavior Cloning). *Let π_D and π be the expert policy and the agent policy. If $0 \le r(s,a) \le r^{\max}$ and $\mathbb{E}_{s\sim\rho^{\pi_D}_T} D_{KL}(\pi_D(\cdot|s)||\pi(\cdot|s)) \le \epsilon_{BC}$, then $|R(\pi_D,T) - R(\pi,T)| \le \sqrt{\epsilon_{BC}/2}\, r^{\max}\Big(\frac{1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\Big)$.*

Proof. The result is immediate from Theorem 1 and the Pinsker's Inequality. □

Corollary 2 (Error of GAIL). *Let π_D and π be the expert policy and the agent policy. If $0 \le r(s,a) \le r^{\max}$ and $D_{JS}(\rho^{\pi_D}_T||\rho^{\pi}_T) \le \epsilon_{GAIL}$. Then $|R(\pi_D,T) - R(\pi,T)| \le \sqrt{2\epsilon_{GAIL}}\, r^{\max}/(1-\gamma)$*

Proof. By definition of the JSD, for any distributions P, Q and their average $M = (P+Q)/2$ we know

$$D_{JS}(P||Q) = \frac{1}{2}\big[D_{KL}(P||M) + D_{KL}(Q||M)\big] \ge D_{TV}(P||M)^2 + D_{TV}(Q||M)^2 \ge \frac{1}{2} D_{TV}(P||Q)^2,$$

where the first inequality follows from Pinsker's Inequality, and the second inequality holds because that $D_{TV}(P||M) + D_{TV}(Q||M) \ge D_{TV}(P||Q)$ by triangle inequality and that $2a^2 + 2b^2 \ge c^2$ if $a + b \ge c \ge 0$. Thus, we know $D_{TV}(\rho^{\pi_D}_T||\rho^{\pi}_T) \le \sqrt{2\epsilon_{GAIL}}$. Applying Lemma 2 completes the proof. □

A.3 MBRL with Absolutely Continuous Stochastic Transitions

Theorem 2 (Error of Absolutely Continuous Stochastic Transitions). *Let π_D, T and $\hat{T}$ be the sampling policy, the real and the learned transitions. If $0 \le r(s,a) \le r^{\max}$ and the error in one-step total variation distance is $\epsilon^{\pi_D}_{T,\hat{T}} = \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_T}[D_{TV}(T(\cdot|s,a)||\hat{T}(\cdot|s,a))]$, then $|R(\pi_D,T) - R(\pi_D,\hat{T})| \le \epsilon^{\pi_D}_{T,\hat{T}} r^{\max}\gamma(1-\gamma)^{-2}$.*

Proof. If there is a upper bound for $D_{TV}(\rho^{\pi_D}_T(s,a)||\rho^{\pi_D}_{\hat{T}}(s,a))$, by Lemma 2, we are done. Also, observe that

$$D_{TV}(\rho^{\pi_D}_T(s,a)||\rho^{\pi_D}_{\hat{T}}(s,a)) = \frac{1}{2}\int \pi_D(a|s)\big|\rho^{\pi_D}_T(s) - \rho^{\pi_D}_{\hat{T}}(s)\big|dsda = D_{TV}(\rho^{\pi_D}_T(s)||\rho^{\pi_D}_{\hat{T}}(s)),$$

so $D_{TV}(\rho^{\pi_D}_T(s)||\rho^{\pi_D}_{\hat{T}}(s))$ is of interest. Employing the properties of Bellman flow operator, we have

$$\begin{aligned}
D_{TV}(\rho^{\pi_D}_T(s)||\rho^{\pi_D}_{\hat{T}}(s)) \leq& \frac{1}{1-\gamma} D_{TV}(\rho^{\pi_D}_T(s)||B_{\pi_D,\hat{T}}(\rho^{\pi_D}_T(s))) \\
=& \frac{1}{1-\gamma} D_{TV}(B_{\pi_D,T}(\rho^{\pi_D}_T(s))||B_{\pi_D,\hat{T}}(\rho^{\pi_D}_T(s))) \\
=& \frac{1}{2(1-\gamma)}\int \Big|\gamma\int \big(T(s|s',a') - \hat{T}(s|s',a')\big)\pi_D(a'|s')\rho^{\pi_D}_T(s')ds'da'\Big|ds \\
\leq& \frac{\gamma}{2(1-\gamma)}\int \Big|\big(T(s|s',a') - \hat{T}(s|s',a')\big)\Big|\rho^{\pi_D}_T(s',a')dsds'da' \\
=& \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho^{\pi_D}_T} D_{TV}(T(\cdot|s,a)||\hat{T}(\cdot|s,a)) = \frac{\gamma}{1-\gamma}\epsilon^{\pi_D}_{T,\hat{T}},
\end{aligned}$$

where the top two lines follows from the symmetry bridge property (Lemma 1) and the fixed-point property. Finally, from Lemma 2, we conclude that

$$|R(\pi_D,T) - R(\pi_D,\hat{T})| \leq \epsilon^{\pi_D}_{T,\hat{T}} r^{\max}\frac{\gamma}{(1-\gamma)^2}$$

□

Corollary 3 (Error of MBRL with Absolutely Continuous Stochastic Transition)**.** *Let π_D, π, T and $\hat{T}$ be the sampling policy, the agent policy, the real transition and the learned transition. If $0 \leq r(s,a) \leq r^{\max}$ and the discrepancies are $\epsilon^{\pi_D}_{T,\hat{T}} = \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_T} D_{TV}(T(\cdot|s,a)||\hat{T}(\cdot|s,a))$ and $\epsilon^{T,\gamma}_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\pi_D}_T} D_{TV}(\pi_D(\cdot|s)||\pi(\cdot|s))$, then $|R(\pi,T) - R(\pi,\hat{T})| \leq (\epsilon^{\pi_D}_{T,\hat{T}} + \epsilon^{T,\gamma}_{\pi_D,\pi} + \epsilon^{\hat{T},\gamma}_{\pi_D,\pi})r^{\max}\gamma/(1-\gamma)^2 + (\epsilon^{T,\gamma}_{\pi_D,\pi} + \epsilon^{\hat{T},\gamma}_{\pi_D,\pi})r^{\max}/(1-\gamma)$.*

Proof. Observe that $|R(\pi,T) - R(\pi,\hat{T})| \leq |R(\pi,T) - R(\pi_D,T)| + |R(\pi_D,T) - R(\pi_D,\hat{T})| + |R(\pi_D,\hat{T}) - R(\pi,\hat{T})|$. Combining Theorem 2 and 1, the result follows. □

Lemma 3. *Let $\gamma > \beta$ be discount factors of long and short rollouts. Let π_D and T be the sampling policy and the real transition, then $D_{TV}(\rho^{\pi_D}_{T,\gamma}||\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}) \leq (1-\gamma)\beta/(\gamma-\beta)$.*

Proof. Since $\rho^{\pi_D}_{T,\gamma}$ is generated by the triple (ρ_0,π_D,T) with discount factor γ while $\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}$ is generated by $(\rho^{\pi_D}_{T,\gamma},\pi_D,T)$ with discount factor β. By definition of the occupancy measure we have

$$\begin{aligned}
\rho^{\pi_D}_{T,\gamma}(s,a) &= \sum_{i=0}^{\infty}(1-\gamma)\gamma^i f_i(s,a). \\
\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}(s,a) &= \sum_{i=0}^{\infty}\sum_{j=0}^{i}(1-\gamma)\gamma^{i-j}(1-\beta)\beta^j f_i(s,a),
\end{aligned}$$

where $f_i(s,a)$ is the density of (s,a) at time i if generated by the triple (ρ_0,π_D,T). Then,

$$\begin{aligned}
D_{TV}(\rho^{\pi_D}_{T,\gamma}||\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}) &\leq \frac{1}{2}\sum_{i=0}^{\infty}\Big|(1-\gamma)\gamma^i - \sum_{j=0}^{i}(1-\gamma)\gamma^{i-j}(1-\beta)\beta^j\Big| = \frac{1}{2}\sum_{i=0}^{\infty}(1-\gamma)\gamma^i\Big|1 - \sum_{j=0}^{i}(1-\beta)\Big(\frac{\beta}{\gamma}\Big)^j\Big| \\
&= \frac{1}{2}\sum_{i=0}^{\infty}(1-\gamma)\gamma^i\frac{1}{\gamma-\beta}\Big| - \beta(1-\gamma) + \Big(\frac{\beta}{\gamma}\Big)^{i+1}(1-\beta)\gamma\Big| \\
&\overset{(*)}{=} \frac{(1-\gamma)\beta}{\gamma-\beta}\sum_{i=0}^{M-1} -(1-\gamma)\gamma^i + (1-\beta)\beta^i = \frac{(1-\gamma)\beta}{\gamma-\beta}(\gamma^M - \beta^M) \\
&\leq \frac{(1-\gamma)\beta}{\gamma-\beta}.
\end{aligned}$$

where $(*)$ comes from that $-\beta(1-\gamma)+(\frac{\beta}{\gamma})^i(1-\beta)\gamma$ is a strictly decreasing function in i. Since $\gamma>\beta$, its sign flips from $+$ to $-$ at some index; say M. Finally, the sum of the absolute value are the same between $\sum_{i=0}^{M-1}$ and $\sum_{i=M}^{\infty}$ because the total probability is conservative, and the difference on one side is the same as that on the other. □

Corollary 4 (Error of MBRL with A. C. Stochastic Transition and Branched Rollouts)**.** *Let $\gamma>\beta$ be discount factors of long and short rollouts. Let π_D, π, T and $\hat{T}$ be sampling policy, agent policy, real transition and learned transition. If $0\le r(s,a)\le r^{\max}$ and the discrepancies are $\epsilon^{T,\gamma}_{\pi_D,\pi}=\mathbb{E}_{s\sim\rho^{\pi_D}_{T,\gamma}}D_{TV}(\pi_D(\cdot|s)||\pi(\cdot|s))$, $\epsilon^{\hat{T},\beta}_{\pi_D,\pi}=\mathbb{E}_{s\sim\rho^{\pi_D}_{\hat{T},\beta}}D_{TV}(\pi_D(\cdot|s)||\pi(\cdot|s))$, and ,$\epsilon^{\pi_D,\beta}_{T,\hat{T}}=\mathbb{E}_{(s,a)\sim\rho^{\pi_D}_{T,\beta}}D_{TV}(T(\cdot|s,a)||\hat{T}(\cdot|s,a))$, then*

$$\Big|R_\gamma(\rho_0,\pi,T)-\frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi,\hat{T})\Big|\le r^{\max}\Big(\frac{\epsilon^{T,\gamma}_{\pi_D,\pi}\gamma}{(1-\gamma)^2}+\frac{(\epsilon^{\pi_D,\beta}_{T,\hat{T}}+\epsilon^{\hat{T},\beta}_{\pi_D,\pi})\beta}{(1-\beta)(1-\gamma)}+\frac{\epsilon^{T,\gamma}_{\pi_D,\pi}+\epsilon^{\hat{T},\beta}_{\pi_D,\pi}}{1-\gamma}+\frac{\beta}{\gamma-\beta}\Big)$$

Proof. Expand with the triangle inequality:

$$\begin{aligned}
&\Big|R_\gamma(\rho_0,\pi,T)-\frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi,\hat{T})\Big|\\
\le&\Big|R_\gamma(\rho_0,\pi,T)-R_\gamma(\rho_0,\pi_D,T)\Big|+\Big|R_\gamma(\rho_0,\pi_D,T)-\frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi_D,T)\Big|+\\
&\frac{1-\beta}{1-\gamma}\Big|R_\beta(\rho^{\pi_D}_{T,\gamma},\pi_D,T)-R_\beta(\rho^{\pi_D}_{T,\gamma},\pi_D,\hat{T})\Big|+\frac{1-\beta}{1-\gamma}\Big|R_\beta(\rho^{\pi_D}_{T,\gamma},\pi_D,\hat{T})-R_\beta(\rho^{\pi_D}_{T,\gamma},\pi,\hat{T})\Big|
\end{aligned}$$

By Theorem 1, the first term $\le\epsilon^{T,\gamma}_{\pi_D,\pi}r^{\max}\Big(\frac{1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\Big)$.
The second term is a short extension of Lemma 2 and Lemma 3:

$$\begin{aligned}
R_\gamma(\rho_0,\pi_D,T)&=\frac{1}{1-\gamma}\int r(s,a)\rho^{\pi_D}_{T,\gamma}(s,a)dsda\le\frac{1}{1-\gamma}\int r(s,a)\max\big(\rho^{\pi_D}_{T,\gamma}(s,a),\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}(s,a)\big)dsda\\
&=\frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi_D,T)+\frac{1}{1-\gamma}\int r(s,a)\Big(\max\big(\rho^{\pi}_{T}(s,a),\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}(s,a)\big)-\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}(s,a)\Big)dsda\\
&\le\frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi_D,T)+\frac{r^{\max}}{1-\gamma}\int\Big(\max\big(\rho^{\pi}_{T}(s,a),\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}(s,a)\big)-\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}(s,a)\Big)dsda\\
&\le\frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi_D,T)+\frac{r^{\max}}{1-\gamma}D_{TV}(\rho^{\pi_D}_{T,\gamma}||\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta})
\end{aligned}$$

By the symmetry of the total variation distance and Lemma 3, we obtain

$$\Big|R_\gamma(\rho_0,\pi_D,T)-\frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi_D,T)\Big|\le\frac{r^{\max}}{1-\gamma}D_{TV}(\rho^{\pi_D}_{T,\gamma}||\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta})\le r^{\max}\frac{\beta}{\gamma-\beta}.$$

By Theorem 2, the third term $\le\epsilon^{\pi_D,\beta}_{T,\hat{T}}r^{\max}\frac{\beta}{(1-\beta)(1-\gamma)}$.
By Theorem 1, the fourth term $\le\epsilon^{\hat{T},\beta}_{\pi_D,\pi}r^{\max}\Big(\frac{1}{1-\gamma}+\frac{\beta}{(1-\beta)(1-\gamma)}\Big)$. □

A.4 MBRL with Deterministic Transition and Strong Lipschitz Continuity

Assumption 1.

(1.1) $\overline{T}$, $\hat{\overline{T}}$ are $(L_{\overline{T},s},L_{\overline{T},a})$, $(L_{\hat{\overline{T}},s},L_{\hat{\overline{T}},a})$ Lipschitz w.r.t. states and actions.

(1.2) $\mathcal{A}$ is a convex, closed, bounded (diameter $\mathrm{diam}_{\mathcal{A}}$) set in a $\dim_{\mathcal{A}}$-dimensional space.

(1.3) $\pi(a|s)\sim\mathcal{P}_{\mathcal{A}}[\mathcal{N}(\mu_\pi(s),\Sigma_\pi(s))]$ and $\pi_D(a|s)\sim\mathcal{P}_{\mathcal{A}}[\mathcal{N}(\mu_{\pi_D}(s),\Sigma_{\pi_D}(s))]$

(1.4) μ_π, μ_{π_D}, $\Sigma^{1/2}_{\pi_D}$, and $\Sigma^{1/2}_{\pi}$ are $L_{\pi,\mu}$, $L_{\pi_D,\mu}$, $L_{\pi,\Sigma}$, $L_{\pi_D,\Sigma}$ Lipschitz w.r.t. states.

The validation of Assumption 1 is below.

1.1 **The real and learned transitions are Lipschitz w.r.t. states and actions.** For the real transition especially in continuous control, the Lipschitzness follows from the laws of motion, as computed in Eq. (10) in the paper. For the learned transition, the Lipschitzness can be made by spectral normalization (Miyato et al., 2018) or gradient penalty (Gulrajani et al., 2017), which are some notable approaches to ensure the Lipschitzness of the discriminator in Wasserstein GAN (Arjovsky et al., 2017).

1.2 **The action space is convex, closed and bounded in a finite dimensional linear space.** This is a standard assumption in continuous-control and is usually satisfied (or made satisfied) in practice (Fujita and Maeda, 2018). The boundness assumption, if not naturally satisfied, is addressed in 1.3.

1.3 **The policy follows truncated Gaussian, by projecting the Gaussian r.v. onto the action space.** According to (Fujita and Maeda, 2018), this is a common practice in RL experiment. The Gaussain assumption is made by training some NNs for the mean and variance of the policies. As for the projection of action to a bounded convex set, it is perfectly fine in RL experiment and is largely used in most MuJoCo experiments as MuJoCo also provides the bounds for the action space. It is also a good practice since it helps stabilize the training.

1.4 **The mean and covariance of the policy are Lipschitz w.r.t. state.** As again noted in 1.1, the Lipschitzness can be realized by spectral normalization or gradient penalty. Since the mean and covariance of the policy are represented by some NN, this assumption can be easily made in practice.

Lemma 4 (Conditional Contraction)**.** *Under assumption 1, if* $\eta_{\pi,\overline{T}} = L_{\overline{T},s} + L_{\overline{T},a}(L_{\pi,\mu} + L_{\pi,\Sigma}\sqrt{\dim_{\mathcal{A}}}) < 1/\gamma$*, where* γ *is the discount factor of* $B_{\pi,\overline{T}}$*, then* $B_{\pi,\overline{T}}$ *is a* $\gamma\eta_{\pi,\overline{T}}$*-contraction w.r.t. 1-Wasserstein distance.*

Proof. Recall $B_{\pi,\overline{T}}(\rho(s)) = (1-\gamma)\rho_0(s) + \gamma\int \delta(s - \overline{T}(s',a'))\pi(a'|s')\rho(s')ds'da'$. Let $\rho_1(s)$, $\rho_2(s)$ be some distributions over states. We have

$$
\begin{aligned}
&W_1(B_{\pi,\overline{T}}(\rho_1) \,\|\, B_{\pi,\overline{T}}(\rho_2)) \\
\overset{(a)}{\leq}&\gamma \inf_{J(s_1,a_1,s_2,a_2)\sim\Pi(\rho_1(s)\pi(a|s),\rho_2(s)\pi(a|s))} \mathbb{E}_J \left\|\overline{T}(s_1,a_1) - \overline{T}(s_2,a_2)\right\|_2 \\
=&\gamma \inf_{J(s_1,a_1,s_2,a_2)\sim\Pi(\rho_1(s)\pi(a|s),\rho_2(s)\pi(a|s))} \mathbb{E}_J \left\|\overline{T}(s_1,a_1) - \overline{T}(s_1,a_2) + \overline{T}(s_1,a_2) - \overline{T}(s_2,a_2)\right\|_2 \\
\leq&\gamma \inf_{J(s_1,a_1,s_2,a_2)\sim\Pi(\rho_1(s)\pi(a|s),\rho_2(s)\pi(a|s))} \mathbb{E}_J L_{\overline{T},a}\left\|a_1 - a_2\right\|_2 + L_{\overline{T},s}\left\|s_1 - s_2\right\|_2 \\
\overset{(b)}{=}&\gamma \inf_{J(s_1,\xi_1,s_2,\xi_2)\sim\Pi(\rho_1,\mathcal{N},\rho_2,\mathcal{N})} \mathbb{E}_J L_{\overline{T},a}\left\|\mathcal{P}_{\mathcal{A}}[\mu_\pi(s_1) + \Sigma_\pi^{1/2}(s_1)\xi_1] - \mathcal{P}_{\mathcal{A}}[\mu_\pi(s_2) - \Sigma_\pi^{1/2}(s_2)\xi_2]\right\|_2 + L_{\overline{T},s}\left\|s_1 - s_2\right\|_2 \\
\overset{(c)}{\leq}&\gamma \inf_{J(s_1,\xi_1,s_2,\xi_2)\sim\Pi(\rho_1,\mathcal{N},\rho_2,\mathcal{N})} \mathbb{E}_J L_{\overline{T},a}\left\|\mu_\pi(s_1) + \Sigma_\pi^{1/2}(s_1)\xi_1 - \mu_\pi(s_2) - \Sigma_\pi^{1/2}(s_2)\xi_2\right\|_2 + L_{\overline{T},s}\left\|s_1 - s_2\right\|_2 \\
\leq&\gamma \inf_{J(s_1,s_2)\sim\Pi(\rho_1,\rho_2)} \Big(\mathbb{E}_J(L_{\overline{T},s} + L_{\overline{T},a}L_{\pi,\mu})\left\|s_1 - s_2\right\|_2 + \inf_{K(\xi_1,\xi_2)\sim\Pi(\mathcal{N},\mathcal{N})} \mathbb{E}_K L_{\overline{T},a}\left\|\Sigma_\pi^{1/2}(s_1)\xi_1 - \Sigma_\pi^{1/2}(s_2)\xi_2\right\|_2\Big) \\
\overset{(d)}{\leq}&\gamma \inf_{J(s_1,s_2)\sim\Pi(\rho_1,\rho_2)} \Big(\mathbb{E}_J(L_{\overline{T},s} + L_{\overline{T},a}L_{\pi,\mu})\left\|s_1 - s_2\right\|_2 + \mathbb{E}_{\xi_1} L_{\overline{T},a}\left\|\Sigma_\pi^{1/2}(s_1) - \Sigma_\pi^{1/2}(s_2)\right\|_{op}\left\|\xi_1\right\|_2\Big) \\
\overset{(e)}{\leq}&\gamma \inf_{J(s_1,s_2)\sim\Pi(\rho_1,\rho_2)} \mathbb{E}_J(L_{\overline{T},s} + L_{\overline{T},a}(L_{\pi,\mu} + L_{\pi,\Sigma}\sqrt{\dim_{\mathcal{A}}}))\left\|s_1 - s_2\right\|_2 \\
=&\gamma(L_{\overline{T},s} + L_{\overline{T},a}(L_{\pi,\mu} + L_{\pi,\Sigma}\sqrt{\dim_{\mathcal{A}}}))W_1(\rho_1 \,\|\, \rho_2) = \gamma\eta_{\pi,\overline{T}}W_1(\rho_1 \,\|\, \rho_2),
\end{aligned}
$$

where $\inf_{J(s_1,s_2)\sim\Pi(\rho_1,\rho_2)}$ takes a infimum over all joint distributions $J(s_1,s_2)$ whose marginals are ρ_1 and ρ_2. (a) selects a joint distribution over $B_{\pi,\overline{T}}(\rho_1)$ and $B_{\pi,\overline{T}}(\rho_2)$ that share the same randomness of $(1-\gamma)\rho_0$, which establishes a upper bound and allows us to cancel $(1-\gamma)\rho_0$. (b) uses the Gaussian assumption of the policy, with ξ_1, ξ_2 being standard normal vectors. (c) uses the non-expansiveness property of projection onto a closed convex set.(d) selects $\xi_1 = \xi_2$ and uses the property of operator norm. (e) uses the Lipschitz assumption of $\Sigma_\pi^{1/2}(s)$ and that $\|\xi_1\| \leq \sqrt{\dim_{\mathcal{A}}}$ by Jensen inequality. □

Lemma 5 (Error w.r.t. W1 Distance between Occupancy Measures)**.** *Let $\rho_1(s,a)$, $\rho_2(s,a)$ be two normalized occupancy measures of rollouts with discount factor γ. If the reward is L_r-Lipschitz, then $|R(\rho_1) - R(\rho_2)| \leq W_1(\rho_1 \,\|\, \rho_2) L_r/(1-\gamma)$.*

Proof. The cumulative reward is bounded by

$$\begin{aligned}
R(\rho_1) &= \frac{1}{1-\gamma}\int r(s,a)\rho_1(s,a)dsda = R(\rho_2) + \frac{1}{1-\gamma}\int r(s,a)\big(\rho_1(s,a) - \rho_2(s,a)\big)dsda \\
&= R(\rho_2) + \frac{L_r}{1-\gamma}\int \frac{r(s,a)}{L_r}\big(\rho_1(s,a) - \rho_2(s,a)\big)dsda \\
&\leq R(\rho_2) + \frac{L_r}{1-\gamma}\sup_{\|f\|_{\mathrm{Lip}}\leq 1}\int f(s,a)\big(\rho_1(s,a) - \rho_2(s,a)\big)dsda \\
&= R(\rho_2) + \frac{L_r}{1-\gamma}\sup_{\|f\|_{\mathrm{Lip}}\leq 1}\mathbb{E}_{(s,a)\sim\rho_1}[f(s,a)] - \mathbb{E}_{(s,a)\sim\rho_2}[f(s,a)] \\
&= R(\rho_2) + \frac{L_r}{1-\gamma}W_1(\rho_1 \,\|\, \rho_2).
\end{aligned}$$

The third line holds because $r(s,a)/L_r$ is 1-Lipschitz and the last line follows from Kantorovich-Rubinstein duality Villani (2008). Since W_1 distance is symmetric, the same conclusion holds if interchanging ρ_1 and ρ_1; thus

$$|R(\rho_1) - R(\rho_2)| \leq W_1(\rho_1 \,\|\, \rho_2) L_r/(1-\gamma).$$

□

Theorem 3 (Error of Deterministic Transitions with Strong Lipschitzness)**.** *Under Lemma 4, let $\overline{T}$, $\hat{\overline{T}}$, r, π_D be deterministic real transition, deterministic learned transtion, reward and sampling policy. If $r(s,a)$ is L_r-Lipschitz and the ℓ_2 error is ϵ_{ℓ_2}, then $|R(\pi_D, \overline{T}) - R(\pi_D, \hat{\overline{T}})| \leq (1 + L_{\pi_D,\mu} + L_{\pi_D,\Sigma}\sqrt{\dim_{\mathcal{A}}}) L_r \frac{\gamma\epsilon_{\ell_2}}{(1-\gamma)(1-\gamma\eta_{\pi_D,\hat{\overline{T}}})}$.*

Proof. Observe that the Wasserstein distance over the joint can be upper bounded by that over the marginal.

$$\begin{aligned}
W_1(\rho^{\pi_D}_{\overline{T}}(s,a) \,\|\, \rho^{\pi_D}_{\hat{\overline{T}}}(s,a)) =& \inf_{J(s_1,a_1,s_2,a_2)\in\Pi(\rho^{\pi_D}_{\overline{T}}(s,a),\rho^{\pi_D}_{\hat{\overline{T}}}(s,a))} \mathbb{E}_J \left\|(s_1 - s_2, a_1 - a_2)\right\|_2 \\
\leq& \inf_{J(s_1,a_1,s_2,a_2)\in\Pi(\rho^{\pi_D}_{\overline{T}}(s,a),\rho^{\pi_D}_{\hat{\overline{T}}}(s,a))} \mathbb{E}_J \left\|s_1 - s_2\right\|_2 + \left\|a_1 - a_2\right\|_2 \\
\overset{(*)}{\leq}& (1 + L_{\pi_D,\mu} + L_{\pi_D,\Sigma}\sqrt{\dim_{\mathcal{A}}}) \inf_{J(s_1,s_2)\in\Pi(\rho^{\pi_D}_{\overline{T}}(s),\rho^{\pi_D}_{\hat{\overline{T}}}(s))} \mathbb{E}_J \left\|s_1 - s_2\right\|_2 \\
=& (1 + L_{\pi_D,\mu} + L_{\pi_D,\Sigma}\sqrt{\dim_{\mathcal{A}}}) W_1(\rho^{\pi_D}_{\overline{T}}(s) \,\|\, \rho^{\pi_D}_{\hat{\overline{T}}}(s)),
\end{aligned} \tag{6}$$

where (*) follows from the same analysis in Lemma 4. Also, the Wasserstein distance over the marginal is upper bounded by the ℓ_2 error:

$$\begin{aligned}
W_1(\rho^{\pi_D}_{\overline{T}}(s) \,\|\, \rho^{\pi_D}_{\hat{\overline{T}}}(s)) \leq& \frac{1}{1-\gamma\eta_{\pi_D,\hat{\overline{T}}}} W_1(\rho^{\pi_D}_{\overline{T}}(s) \,\|\, B^{\pi_D}_{\hat{\overline{T}}}(\rho^{\pi_D}_{\overline{T}}(s))) = \frac{1}{1-\gamma\eta_{\pi_D,\hat{\overline{T}}}} W_1(B^{\pi_D}_{\overline{T}}(\rho^{\pi_D}_{\overline{T}}(s)) \,\|\, B^{\pi_D}_{\hat{\overline{T}}}(\rho^{\pi_D}_{\overline{T}}(s))) \\
\leq& \frac{\gamma}{1-\gamma\eta_{\pi_D,\hat{\overline{T}}}} \inf_{J(s_1,a_1,s_2,a_2)\sim\Pi(\rho^{\pi_D}_{\overline{T}}(s)\pi_D(a|s),\rho^{\pi_D}_{\overline{T}}(s)\pi_D(a|s))} \mathbb{E}_J \left\|\overline{T}(s_1,a_1) - \hat{\overline{T}}(s_2,a_2)\right\|_2 \\
\leq& \frac{\gamma}{1-\gamma\eta_{\pi_D,\hat{\overline{T}}}} \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_{\overline{T}}(s)\pi_D(a|s)} \left\|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\right\|_2 = \frac{\gamma}{1-\gamma\eta_{\pi_D,\hat{\overline{T}}}}\epsilon_{\ell_2}.
\end{aligned} \tag{7}$$

The first line follows from conditional contraction (Lemma 4), symmetry bridge (Lemma 1) and fixed-point property. The second line uses the fact that $B^{\pi_D}_{\overline{T}}$ and $B^{\pi_D}_{\hat{\overline{T}}}$ have $1-\gamma$ fraction in common, so we can create a joint

distribution to cancel it. The third line builds a upper bound by choosing $(s_1, a_1) = (s_2, a_2) \sim \rho^{\pi_D}_{\overline{T}}(s)\pi_D(a|s)$. Combining Eq. (6), (7) and Lemma 5, we conclude that

$$|R(\pi_D, \overline{T}) - R(\pi_D, \hat{\overline{T}})| \leq (1 + L_{\pi_D,\mu} + L_{\pi_D,\Sigma}\sqrt{\dim_{\mathcal{A}}})L_r \frac{\gamma\epsilon_{\ell_2}}{(1-\gamma)(1-\gamma\eta_{\pi_D,\hat{\overline{T}}})}.$$

□

Corollary 5 (Error of MBRL with Deterministic Transition, Strong Lipschitzness and Branched Rollouts)**.** *Let $\gamma > \beta$ be discount factors of long and short rollouts. Let π_D, π, $\overline{T}$ and $\hat{\overline{T}}$ be sampling policy, agent policy, real deterministic transition and deterministic learned transition. Under assumption 1, suppose the reward is both bounded $0 \leq r(s,a) \leq r^{\max}$ and L_r-Lipschitz. Let $\epsilon^{\overline{T},\gamma}_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\pi_D}_{\overline{T},\gamma}} D_{TV}(\pi_D(\cdot|s)||\pi(\cdot|s))$,*
$\epsilon^{\hat{\overline{T}},\beta}_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\pi_D}_{\hat{\overline{T}},\beta},\pi} D_{TV}(\pi_D(\cdot|s)||\pi(\cdot|s))$ *and* $\epsilon_{\ell_2,\beta} = \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_{\overline{T},\gamma},\pi_D} \left\|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\right\|_2$*. Then,*

$$\begin{aligned}\left|R_\gamma(\rho_0, \pi, \overline{T}) - \frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{\overline{T},\gamma}, \pi, \overline{T})\right| &\leq r^{\max}\Big(\frac{\epsilon^{\overline{T},\gamma}_{\pi_D,\pi}\gamma}{(1-\gamma)^2} + \frac{\epsilon^{\hat{\overline{T}},\beta}_{\pi_D,\pi}\beta}{(1-\beta)(1-\gamma)} + \frac{\epsilon^{\overline{T},\gamma}_{\pi_D,\pi} + \epsilon^{\hat{\overline{T}},\beta}_{\pi_D,\pi}}{1-\gamma} + \frac{\beta}{\gamma-\beta}\Big) \\ &\quad + (1 + L_{\pi_D,\mu} + L_{\pi_D,\Sigma}\sqrt{\dim_{\mathcal{A}}})L_r \frac{\beta\epsilon_{\ell_2,\beta}}{(1-\gamma)(1-\beta\eta_{\pi_D,\hat{\overline{T}}})}\end{aligned}$$

Proof. Modifying the proof of Corollary 4 with Theorem 3, the result follows. □

A.5 MBRL with Deterministic Transition and Weak Lipschitz Continuity

Theorem 4 (One-sided Error of Deterministic Transitions)**.** *Let $\overline{T}$, $\hat{\overline{T}}$, r, π_D be deterministic real transition, deterministic learned transtion, reward and sampling policy. Suppose $0 \leq r(s,a) \leq r^{\max}$. $\hat{\overline{T}}(s,a)$, $r(s,a)$ and $\pi_D(a|s)$ are Lipschitz in s for any a with constants $(L_{\hat{\overline{T}}}, L_r, L_{\pi_D})$. Assume that $L_{\hat{\overline{T}}} \leq 1 + (1-\gamma)\iota$ with $\iota < 1$ and that the action space is bounded:* $\text{diam}_{\mathcal{A}} < \infty$*. If the training loss in ℓ_2 error is ϵ_{ℓ_2}, then*

$$R(\pi_D, \overline{T}) - R(\pi_D, \hat{\overline{T}}) \leq \frac{1+\gamma}{(1-\gamma)^2}\sqrt{2\epsilon_{\ell_2} r^{\max} L_r} + \frac{1+O(\iota)}{(1-\gamma)^{5/2}} r^{\max}\sqrt{2\epsilon_{\ell_2} L_{\pi_D}\text{diam}_{\mathcal{A}}}.$$

Proof. Recall the ℓ_2 error is $\mathbb{E}_{(s,a)\sim\rho^{\pi_D}_{\overline{T}}}\left[\left\|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\right\|_2\right] = \epsilon_{\ell_2}$. By Markov's Inequality, for any $\delta > 0$,

$$\mathbb{P}_{(s,a)\sim\rho^{\pi_D}_{\overline{T}}}\Big(\left\|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\right\|_2 < \delta\Big) > 1 - \frac{\epsilon_{\ell_2}}{\delta} \tag{8}$$

Eq. (8) means for a length $H \sim \text{Geometric}(1-\gamma)$ rollout $\{s_t, a_t\}_{t=1}^H$ generated by $(\rho_0, \pi_D, \overline{T})$, $\left\|\overline{T}(s_t,a_t) - \hat{\overline{T}}(s_t,a_t)\right\|_2 < \delta$ with probability greater than $1 - \frac{\epsilon_{\ell_2}}{\delta}$.

Following this idea, *we say a rollout is consistent to $\hat{\overline{T}}$, if for each t,* $\left\|s_{t+1} - \hat{\overline{T}}(s_t,a_t)\right\|_2 < \delta$; in other words, a rollout is consistent to $\hat{\overline{T}}$ if for each time step, the state transition is similar to what $\hat{\overline{T}}$ does. Let $P_{\overline{T}}$ be the probability measure induced on the rollout following the real transition $\overline{T}$. The cumulative reward is bounded by

$$\begin{aligned}R(\pi_D, \overline{T}) &= \int_{\text{traj}} R(\text{traj}) dP_{\overline{T}} = \int_{\text{traj consistent}} R(\text{traj}) dP_{\overline{T}} + \int_{\text{traj inconsistent}} R(\text{traj}) dP_{\overline{T}} \\ &\leq \int_{\text{traj consistent}} R(\text{traj}) dP_{\overline{T}} + \frac{\epsilon_{\ell_2}}{\delta}\mathbb{E}[H^2] r^{\max}.\end{aligned} \tag{9}$$

The inequality holds because for a rollout generated by $\overline{T}$ with length H, the probability that it is inconsistent to $\hat{\overline{T}}$ is at most $\frac{\epsilon_{\ell_2}}{\delta}H$ by Eq. (8) and the union bound over $\{s_t, a_t\}_{t=1}^H$. Also, the maximum reward of such rollout is $Hr^{\max}$.

Now, we'd like to change from $P_{\overline{T}}$ to $P_{\hat{\overline{T}}}$ with the Lipschitz assumptions above. It suffices to reset the states $\{s_i\}_{i=1}^H$ so that the transition obeys $\hat{\overline{T}}$. Suppose the new states are

$$s'_1 = s_1, \qquad s'_i = \hat{\overline{T}}(s'_{i-1}, a_{i-1}),\ \forall\ i \geq 2. \tag{10}$$

By the Lipschitzness of $\hat{\overline{T}}$, triangle inequlaity and $\hat{\overline{T}}$-consistency, the distance between s_i and s'_i obeys

$$\begin{aligned}
&\|s_1 - s'_1\|_2 = 0\\
&\|s_i - s'_i\|_2 \leq \left\|s_i - \hat{\overline{T}}(s_{i-1}, a_{i-1})\right\|_2 + \left\|\hat{\overline{T}}(s_{i-1}, a_{i-1}) - \hat{\overline{T}}(s'_{i-1}, a_{i-1})\right\|_2 \leq \delta + L_{\hat{\overline{T}}}\left\|s_{i-1} - s'_{i-1}\right\|_2,\ \forall i \geq 2.
\end{aligned}$$

That is,

$$\|s_i - s'_i\|_2 \leq \delta \sum_{j=0}^{i-2} L_{\hat{\overline{T}}}^j = \delta \frac{L_{\hat{\overline{T}}}^{i-1} - 1}{L_{\hat{\overline{T}}} - 1}, \quad \forall i \geq 2. \tag{11}$$

The difference of cumulative reward between $\text{traj} = \{s_i, a_i\}_{i=1}^H$ and $\text{traj}' = \{s'_i, a_i\}_{i=1}^H$ satisfies

$$\begin{aligned}
R(\text{traj}) = \sum_{i=1}^H r(s_i, a_i) &\leq r(s'_1, a_1) + \sum_{i=2}^H r(s'_i, a_i) + L_r \|s_i - s'_i\|_2 \overset{(11)}{\leq} R(\text{traj}') + \delta L_r \sum_{i=2}^H \frac{L_{\hat{\overline{T}}}^{i-1} - 1}{L_{\hat{\overline{T}}} - 1}\\
&\overset{(13)}{\leq} R(\text{traj}') + \delta L_r (H^2/2 + (\mathbb{E}H)^2 O(\iota)),
\end{aligned} \tag{12}$$

where (13) results from imposing $L_{\hat{\overline{T}}} = 1 + \iota(1-\gamma) = 1 + \frac{\iota}{\mathbb{E}H}$ into the exponential:

$$\begin{aligned}
\sum_{i=2}^H \frac{L_{\hat{\overline{T}}}^{i-1} - 1}{L_{\hat{\overline{T}}} - 1} &= \frac{1}{L_{\hat{\overline{T}}} - 1}\Big(\frac{L_{\hat{\overline{T}}}^H - L_{\hat{\overline{T}}}}{L_{\hat{\overline{T}}} - 1} - H + 1\Big) = \frac{(1 + \frac{\iota}{\mathbb{E}H})^H - \iota\frac{H}{\mathbb{E}H} - 1}{\frac{\iota^2}{(\mathbb{E}H)^2}} \leq \frac{e^{\iota\frac{H}{\mathbb{E}H}} - \iota\frac{H}{\mathbb{E}H} - 1}{\frac{\iota^2}{(\mathbb{E}H)^2}}\\
&= \frac{(\iota\frac{H}{\mathbb{E}H})^2/2 + O(\iota^3)}{\frac{\iota^2}{(\mathbb{E}H)^2}} = \frac{H^2}{2} + (\mathbb{E}H)^2 O(\iota)
\end{aligned} \tag{13}$$

Because the transitions are deterministic, $\{s'_i\}_{i=1}^H$ are constant given $s_1,\ a_1, ..., a_H$, which means the randomness depends on $s_1,\ a_1, ..., a_H$ (with $\{s'_i\}_{i=1}^H$ being the conditions of π_D), and the density satisfies

$$\begin{aligned}
P_{\hat{\overline{T}}}(\text{traj}') &= \rho_0(s'_1)\pi_D(a_1|s'_1)\prod_{i=2}^H \pi_D(a_i|s'_i) \geq \rho_0(s_1)\pi_D(a_1|s_1)\prod_{i=2}^H \left(\pi_D(a_i|s_i) - L_{\pi_D}\|s_i - s'_i\|_2\right)\\
&\overset{(11)}{\geq} \rho_0(s_1)\pi_D(a_1|s_1)\prod_{i=2}^H \left(\pi_D(a_i|s_i) + \delta L_{\pi_D}\frac{L_{\hat{\overline{T}}}^{i-1} - 1}{L_{\hat{\overline{T}}} - 1}\right) \geq P_{\overline{T}}(\text{traj})\Big(1 - \sum_{i=2}^H \frac{\delta L_{\pi_D}}{\pi_D(a_i|s_i)}\frac{L_{\hat{\overline{T}}}^{i-1} - 1}{L_{\hat{\overline{T}}} - 1}\Big)
\end{aligned} \tag{14}$$

Then, conditioning on the length of rollout being H, the integral term in Eq. (9) is bounded by

$$
\begin{aligned}
&\int_{\text{traj consistent}|H} R(\text{traj}) dP_{\overline{T}} = \int_{s_1,a_1,\ldots,a_H \text{ consis.}} R(\text{traj}) P_{\overline{T}}(\text{traj}) ds_1 da_1 ... da_H \\
\overset{(14)}{\leq} &\int_{s_1,a_1,\ldots,a_H \text{ consis.}} R(\text{traj}) \Big(P_{\hat{\overline{T}}}(\text{traj}') + P_{\overline{T}}(\text{traj}) \sum_{i=2}^{H} \frac{\delta L_{\pi_D}}{\pi_D(a_i|s_i)} \frac{L_{\hat{\overline{T}}}^{i-1} - 1}{L_{\hat{\overline{T}}} - 1} \Big) ds_1 da_1 ... da_H \\
\leq &\int_{s_1,a_1,\ldots,a_H \text{ consis.}} R(\text{traj}) P_{\hat{\overline{T}}}(\text{traj}') + \int_{s_1,a_1,\ldots,a_H} R(\text{traj}) P_{\overline{T}}(\text{traj}) \sum_{i=2}^{H} \frac{\delta L_{\pi_D}}{\pi_D(a_i|s_i)} \frac{L_{\hat{\overline{T}}}^{i-1} - 1}{L_{\hat{\overline{T}}} - 1} \\
\overset{(12)}{\leq} &\int_{s_1,a_1,\ldots,a_H \text{ consis.}} \big(R(\text{traj}') + \delta L_r (H^2/2 + (\mathbb{E}H)^2 O(\iota)) \big) P_{\hat{\overline{T}}}(\text{traj}') ds_1 da_1 ... da_H + \\
&\int_{s_1,a_1,\ldots,a_H} H r^{\max} P_{\overline{T}}(\text{traj}) \sum_{i=2}^{H} \frac{\delta L_{\pi_D}}{\pi_D(a_i|s_i)} \frac{L_{\hat{\overline{T}}}^{i-1} - 1}{L_{\hat{\overline{T}}} - 1} ds_1 da_1 ... da_H \\
\overset{(13)}{\leq} &\int_{s_1,a_1,\ldots,a_H} \big(R(\text{traj}') + \delta L_r (H^2/2 + (\mathbb{E}H)^2 O(\iota)) \big) P_{\hat{\overline{T}}}(\text{traj}') + H r^{\max} \delta L_{\pi_D} \text{diam}_{\mathcal{A}} (H^2/2 + (\mathbb{E}H)^2 O(\iota)) \\
\leq &R(\pi_D, \hat{\overline{T}}) + \delta L_r (H^2/2 + (\mathbb{E}H)^2 O(\iota)) + \delta L_{\pi_D} r^{\max} \text{diam}_{\mathcal{A}} (H^3/2 + H(\mathbb{E}H)^2 O(\iota))
\end{aligned}
\tag{15}
$$

Combining Eq. (9) (15), by choosing

$$
\delta = \sqrt{\frac{2\epsilon_{\ell_2} r^{\max} \mathbb{E}[H^2]}{L_r \mathbb{E}[H^2] + \mathbb{E}[H]^2 O(\iota) + L_{\pi_D} r^{\max} \text{diam}_{\mathcal{A}} \big(\mathbb{E}[H^3] + \mathbb{E}[H]^3 O(\iota) \big)}},
$$

we are able to minimize:

$$
\frac{\epsilon_{\ell_2}}{\delta} \mathbb{E}[H^2] r^{\max} + \delta L_r (\mathbb{E}[H^2]/2 + (\mathbb{E}H)^2 O(\iota)) + \delta L_{\pi_D} r^{\max} \text{diam}_{\mathcal{A}} (\mathbb{E}[H^3]/2 + (\mathbb{E}H)^3 O(\iota)),
$$

yielding

$$
\begin{aligned}
&R(\pi_D, \overline{T}) - R(\pi_D, \hat{\overline{T}}) \\
\leq &\mathbb{E}[H^2] \sqrt{\Big(2\epsilon_{\ell_2} r^{\max} \Big) \Big(L_r + \mathbb{E}[H]^2 O(\iota)/E[H^2] + L_{\pi_D} r^{\max} \text{diam}_{\mathcal{A}} \big(\mathbb{E}[H^3]/\mathbb{E}[H^2] + \mathbb{E}[H]^3 O(\iota)/\mathbb{E}[H^2] \big) \Big)} \\
\overset{(a)}{=} &\mathbb{E}[H^2] \sqrt{2\epsilon_{\ell_2} r^{\max} L_r + 2\epsilon_{\ell_2} L_{\pi_D} (r^{\max})^2 \text{diam}_{\mathcal{A}} \big(\mathbb{E}[H^3]/\mathbb{E}[H^2] + \mathbb{E}[H]^3 O(\iota)/\mathbb{E}[H^2] \big)} \\
\overset{(b)}{\leq} &\mathbb{E}[H^2] \sqrt{2\epsilon_{\ell_2} r^{\max} L_r} + \mathbb{E}[H^2] r^{\max} \sqrt{2\epsilon_{\ell_2} L_{\pi_D} \text{diam}_{\mathcal{A}} \big(\mathbb{E}[H^3]/\mathbb{E}[H^2] + \mathbb{E}[H]^3 O(\iota)/\mathbb{E}[H^2] \big)} \\
= &\mathbb{E}[H^2] \sqrt{2\epsilon_{\ell_2} r^{\max} L_r} + r^{\max} \sqrt{2\epsilon_{\ell_2} L_{\pi_D} \text{diam}_{\mathcal{A}}} \sqrt{\mathbb{E}[H^2] \big(\mathbb{E}[H^3] + \mathbb{E}[H]^3 O(\iota) \big)} \\
\overset{(c)}{=} &\frac{1+\gamma}{(1-\gamma)^2} \sqrt{2\epsilon_{\ell_2} r^{\max} L_r} + \frac{\sqrt{1 + 5\gamma + 5\gamma^2 + \gamma^3 + (1+\gamma) O(\iota)}}{(1-\gamma)^{5/2}} r^{\max} \sqrt{2\epsilon_{\ell_2} L_{\pi_D} \text{diam}_{\mathcal{A}}} \\
\overset{(d)}{\leq} &\frac{1+\gamma}{(1-\gamma)^2} \sqrt{2\epsilon_{\ell_2} r^{\max} L_r} + \frac{1 + O(\iota)}{(1-\gamma)^{5/2}} r^{\max} \sqrt{2\epsilon_{\ell_2} L_{\pi_D} \text{diam}_{\mathcal{A}}}.
\end{aligned}
$$

(a) merge the two $O(\iota)$ terms together. (b) uses $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x,\ y \geq 0$. (c) applies the identities $\mathbb{E}[H^2] = \frac{1+\gamma}{(1-\gamma)^2}$, $\mathbb{E}[H^3] = \frac{1+4\gamma+\gamma^2}{(1-\gamma)^3}$. (d) uses $\sqrt{1+x} \leq 1 + x/2$. □

Corollary 6 (One-sided of MBRL with Deterministic Transition and Branched Rollouts)**.** *Let $\gamma > \beta$ be discount factors of long and short rollouts. Let π_D, π, $\overline{T}$ and $\hat{\overline{T}}$ be sampling policy, agent policy, real deterministic transition and deterministic learned transition. Under the assumptions of Theorem 4, let* $\epsilon_{\pi_D,\pi}^{\overline{T},\gamma} = \mathbb{E}_{s \sim \rho_{\overline{T},\gamma}^{\pi_D}} D_{TV}(\pi_D(\cdot|s) || \pi(\cdot|s))$, $\epsilon_{\pi_D,\pi}^{\hat{\overline{T}},\beta} = \mathbb{E}_{s \sim \rho_{\hat{\overline{T}},\beta}^{\rho_{\overline{T},\gamma}^{\pi_D},\pi}} D_{TV}(\pi_D(\cdot|s) || \pi(\cdot|s))$

and $\epsilon_{\ell_2,\beta} = \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_{\overline{T},\gamma},\pi_D} \left\|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\right\|_2$. *Then* $R_\gamma(\rho_0,\pi,\overline{T}) - \frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{\overline{T},\gamma},\pi,\overline{T}) \leq r^{\max}\Big(\frac{\epsilon^{\overline{T},\gamma}_{\pi_D,\pi}\gamma}{(1-\gamma)^2} + \frac{\epsilon^{\hat{\overline{T}},\beta}_{\pi_D,\pi}\beta}{(1-\beta)(1-\gamma)} + \frac{\epsilon^{\overline{T},\gamma}_{\pi_D,\pi}+\epsilon^{\hat{\overline{T}},\beta}_{\pi_D,\pi}}{1-\gamma} + \frac{\beta}{\gamma-\beta}\Big) + \frac{1+\beta}{(1-\beta)(1-\gamma)}\sqrt{2\epsilon_{\ell_2,\beta}r^{\max}L_r} + \frac{1+O(\iota)}{(1-\beta)^{3/2}(1-\gamma)}r^{\max}\sqrt{2\epsilon_{\ell_2,\beta}L_{\pi_D}\mathrm{diam}_{\mathcal{A}}}$.

Proof. Plugging in Theorem 4 with $L_{\hat{\overline{T}}} \leq 1 + (1-\beta)\iota$ to the proof of Corollary 4, the result follows. □

References

Arjovsky, M., S. Chintala, and L. Bottou
2017. Wasserstein gan. arXiv:1701.07875.

Conrad, K.
2014. The contraction mapping theorem. University of Connecticut. Expository paper.

Fujita, Y. and S.-i. Maeda
2018. Clipped action policy gradient. volume 80 of *Proceedings of Machine Learning Research*, Pp. 1597–1606, Stockholmsmässan, Stockholm Sweden. PMLR.

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville
2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Pp. 5767–5777. Curran Associates, Inc.

Hewitt, E. and K. A. Ross
1963. *Abstract Harmonic Analysis I.* Springer-Verlag.

Miyato, T., T. Kataoka, M. Koyama, and Y. Yoshida
2018. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*.

Villani, C.
2008. *Optimal transport – Old and new*, volume 338, Pp. xxii+973.