A Contraction Approach to Model-based Reinforcement Learning

Ting-Han Fan **Peter J. Ramadge**
Department of Electrical and Computer Engineering, Princeton University

Abstract

Despite its experimental success, Model-based Reinforcement Learning still lacks a complete theoretical understanding. To this end, we analyze the error in the cumulative reward using a contraction approach. We consider both stochastic and deterministic state transitions for continuous (non-discrete) state and action spaces. This approach doesn't require strong assumptions and can recover the typical quadratic error to the horizon. We prove that branched rollouts can reduce this error and are essential for deterministic transitions to have a Bellman contraction. Our analysis of policy mismatch error also applies to Imitation Learning. In this case, we show that GAN-type learning has an advantage over Behavioral Cloning when its discriminator is well-trained.

1 Introduction

Reinforcement learning (RL) has attracted much attention recently due to its ability to learn good policies for sequential systems. However, most RL algorithms have a high sample complexity of environment queries (typically in the order of millions). This sample complexity hinders the deployment of RL in practical systems. An intuitive potential solution is to learn an accurate model of the environment's outcome, hence reducing the demand for environment queries. This approach leads to a dichotomy of RL algorithms: training without an environment model is called model-free RL, and training with an environment model is called model-based RL. Model-free RL is often faulted for low exploitation of environment queries, while the performance of model-based RL suffers under model inaccuracy.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Model-based Reinforcement Learning (MBRL) is non-trivial since RL's sequential nature allows errors to propagate to future time-steps. This fact leads to the planning horizon dilemma (Wang et al., 2019); a long horizon incurs a large cumulative error, while a short horizon results in shortsighted decisions. We need to understand this trade-off better as it is currently one of the fundamental limitations of model-based RL.

Most prior error analyses impose a strong assumption in their proofs; e.g., Lipschitz value function (Luo et al., 2019; Xiao et al., 2019; Yu et al., 2020) or maximum model error (Janner et al., 2019). In general, the value function is unlikely to be Lipschitz because its gradient w.r.t. the state can be very large. This event happens when a state perturbation is applied at the stability-instability boundary of a control system, resulting in a large change in value (performance) from a small change in state. For instance, if one perturbs a robot's leg, it may fall and, as a result, receive many negative future rewards.

To mitigate the cumulative reward error, Janner et al. (2019) experimentally shows that branched rollouts (short model rollouts initialized by previous real rollouts) help reduce this error and improve experimental results. However, the effectiveness of branched rollouts remains unclear since the experiments of Janner et al. (2019) use deterministic transitions (MuJoCo (Todorov et al., 2012)). However, their error analysis only applies to stochastic transitions and contains unclear reasoning, see §3. Ideally, we need an analysis framework that applies to both stochastic and deterministic transitions.

Our main contribution is a contraction-based approach to analyze the error of MBRL that applies to both stochastic and deterministic transitions without strong assumptions. Prior work typically makes strong assumptions such as a Lipschitz value function or a maximum model error. To avoid these assumptions and maintain generality, we: (a) provide an analysis framework that applies to both (absolutely continuous) stochastic and deterministic transitions, (b) mostly uses constants in expectation, and (c) does not require a Lipschitz assumption on value functions. Our

results also contribute to theoretical explanations of some techniques in deep RL. We prove that branched rollouts significantly reduce the cumulative reward error for both stochastic and deterministic transitions. In particular, branched rollouts are vital for deterministic transitions to have a Bellman contraction. Although prior work also claimed to have a similar conclusion, the analysis is unclear (see §3) and, in any case, does not apply to the deterministic environment in their experiment. Our approach also helps analyze Imitation Learning. We show a GAN-type learning method like Generative Adversarial Imitation Learning (Ho and Ermon, 2016) is potentially preferable to a supervised learning method like Behavioral Cloning (Ross et al., 2011; Syed and Schapire, 2010) when the discriminator is well-trained.

The primary intuition of our analysis comes from the MBRL problem's asymmetry: the policy mismatch and model mismatch errors, or the terms on the RHS of Eq. (8), are the errors of interest and are symmetric when interchanging transitions or policies. However, the objects that control the errors of interest (can be directly made small in training) are asymmetric. At some point, we have to bridge from symmetry to asymmetry. We show that the Bellman flow operator is the key to this bridge. If the Bellman flow operator is a contraction w.r.t. a metric, we can analyze the error of MBRL under that metric regardless of the asymmetry. When we do not have a Bellman contraction, we provide another way inspired by Syed and Schapire (2010) to analyze the problem and identify the impact of asymmetry. The resulting insight suggests the potential usefulness of the Ensemble Method (Kurutach et al., 2018).

Prior work has done extensive experiments on branched rollouts (Janner et al., 2019), Generative Adversarial Imitation Learning (Ho and Ermon, 2016) and the Ensemble Method (Kurutach et al., 2018). Since the empirical evidence in the literature is clear, this work does not include additional experiments. Instead, we focus on providing an improved theoretical understanding of existing empirical results.

2 Preliminaries

Consider an infinite-horizon Markov Decision Process (MDP) represented by $\langle \mathcal{S},\ \mathcal{A},\ T,\ r,\ \gamma \rangle$. Here $\mathcal{S}$, $\mathcal{A}$ are finite-dimensional continuous state and action spaces, $T(s'|s,a)$ is the transition density of s' given (s,a), $r(s,a)$ is the reward function, and $\gamma \in (0,1)$ is the discount factor. We use $\overline{T}$ to denote a deterministic transition with the density $T(s'|s,a) = \delta(s' - \overline{T}(s,a))$.

Given an initial state distribution ρ_0, the goal of reinforcement learning is to learn a (stochastic) policy π that maximizes the γ-discounted cumulative reward $R_\gamma(\rho_0, \pi, T)$, or equivalently, the expected cumulative reward, denoted as $R(\rho_{T,\gamma}^{\rho_0,\pi})$, under the *normalized* occupancy measure $\rho_{T,\gamma}^{\rho_0,\pi}$.

$$\begin{aligned} R_\gamma(\rho_0,\pi,T) &= \mathbb{E}\left[\textstyle\sum_{i=0}^{\infty} \gamma^i r(s_i,a_i) \Big| \rho_0,\pi,T\right] \\ &= \tfrac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_{T,\gamma}^{\rho_0,\pi}}[r(s,a)] \\ &= R(\rho_{T,\gamma}^{\rho_0,\pi}). \\ \rho_{T,\gamma}^{\rho_0,\pi}(s,a) &= (1-\gamma)\textstyle\sum_{i=0}^{\infty}\gamma^i f_i(s,a|\rho_0,\pi,T), \end{aligned} \tag{1}$$

where $f_i(s,a|\rho_0,\pi,T)$ is the density of (s,a) at step i under (ρ_0,π,T). Because (ρ_0,π,T,γ) uniquely determines the occupancy measure, we use $R(\rho_{T,\gamma}^{\rho_0,\pi})$ as an alternative expression for $R_\gamma(\rho_0,\pi,T)$. When ρ_0, γ are fixed, we simplify the notation to $R(\pi,T)$ and ρ_T^π.

2.1 Bellman Flow Operator

In Eq. (1), because each $f_i(s,a|\rho_0,\pi,T)$ uses the same policy, $f_i(s,a)$ and $\rho_{T,\gamma}^{\rho_0,\pi}(s,a)$ can be factored as $f_i(s)\pi(a|s)$ and $\rho_{T,\gamma}^{\rho_0,\pi}(s)\pi(a|s)$. This allows us to mainly focus on the state distributions. In particular, we define the normalized state occupancy measure $\rho_{T,\gamma}^{\rho_0,\pi}(s)$ as the marginal of $\rho_{T,\gamma}^{\rho_0,\pi}(s,a)$ and show (Fact 4, Appendix) that it satisfies a fixed-point equation characterized by a Bellman flow operator $B_{\pi,T}(\cdot)$.

$$\begin{aligned} \rho_{T,\gamma}^{\rho_0,\pi}(s) &\triangleq (1-\gamma)\textstyle\sum_{i=0}^{\infty}\gamma^i f_i(s|\rho_0,\pi,T) \\ &= B_{\pi,T}(\rho_{T,\gamma}^{\rho_0,\pi}(s)), \end{aligned} \tag{2}$$

where $B_{\pi,T}(\cdot)$ under (ρ_0,π,T) and γ is defined as:

$$\begin{aligned} B_{\pi,T}(\rho(s)) \triangleq &(1-\gamma)\rho_0(s) \\ &+ \gamma \textstyle\int T(s|s',a')\pi(a'|s')\rho(s')ds'da' \end{aligned} \tag{3}$$

$B_{\pi,T}(\cdot)$ is a γ-contraction w.r.t. total variation distance (see Appendix). Hence, $B_{\pi,T}(\cdot)$ has a unique fixed point, and by Eq. (2), this point is $\rho_{T,\gamma}^{\rho_0,\pi}(s)$. This result foreshadows the utility of the Bellman flow operator for analyzing state occupancy measures. Indeed, Lemma 1 exploits the Bellman flow operator to upper bound the distance between state distributions. This is useful for analyzing MBRL. In passing, we note that previous work (Syed et al., 2008) has made distinct use of a Bellman flow constraint.

2.2 Model-based RL

We study the model-based RL procedure shown in Algorithm 1, and its variants (e.g., branched rollouts). Line 3 deals with the storage of a dataset $\mathcal{D}$ of real transitions. Observe that $\mathcal{D}_{i-1}$ is generated by (ρ_0,π_{i-1},T) and that $\mathcal{D}$ aggregrates the $\mathcal{D}_{i-1}$'s. The policy that generates $\mathcal{D}$, which we

Algorithm 1 Model-based RL Algorithm

Require: Dataset $\mathcal{D} = \emptyset$, policy π_0, learned transition $\hat{T}$.

1: **for** $i = 1, 2, ...$ **do**
2: Sample $\mathcal{D}_{i-1} = \{s_t, a_t, s'_t\}$ from real transition T and policy π_{i-1}.
3: $\mathcal{D} \leftarrow \text{Truncate}(\mathcal{D} \cup \mathcal{D}_{i-1})$
4: Fit $\hat{T}$ using samples in $\mathcal{D}$.
5: $\pi_i = \arg\max_{\pi \in B_{\pi_D}} R(\pi, \hat{T})$
6: **end for**

call the sampling policy π_D, is a mixture of previous policies. If $\mathcal{D}_{i-1}$'s have equal sizes, $\pi_D(a|s) = \sum_{j=i-q}^{i-1} \pi_j(a|s)\rho_T^{\pi_j}(s) / \sum_{j=i-q}^{i-1} \rho_T^{\pi_j}(s)$ with q being the truncation level. The larger q is, the more dependent the sampling policy is on previous policies. To facilitate subsequent supervised learning, we need $\mathcal{D}$ to be sufficiently large. However, the larger dataset implies the stronger dependence of the sampling policy on previous policies. For technical reasons (see the final paragraph of the section), we need the current policy π and the sampling policy π_D to be sufficiently close. Hence we expect a tight truncation, e.g., $\mathcal{D} = \mathcal{D}_{i-1} \cup \mathcal{D}_{i-2}$.

Line 4 is a supervised learning task. The objective function is usually the log-likelihood for stochastic transitions or the ℓ_2 error for deterministic transitions. For stochastic transitions, maximizing likelihood is equivalent to minimizing KL divergence. Hence, by Pinsker's Inequality, the total variation distance

$$\epsilon_{T,\hat{T}}^{\pi_D} = \mathbb{E}_{(s,a)\sim\rho_T^{\pi_D}} D_{TV}(T(\cdot|s,a) \,\|\, \hat{T}(\cdot|s,a))$$

is small. For deterministic transitions, the objective is to minimize $\epsilon_{\ell_2} = \mathbb{E}_{(s,a)\sim\rho_T^{\pi_D}} \|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\|_2$.

Line 5 is to maximize model-based cumulative reward $R(\pi, \hat{T})$ under the learned transition $\hat{T}$. Still, the overall goal is to maximize the true cumulative reward $R(\pi, T)$. Note that

$$\begin{aligned} R(\pi_i, T) - R(\pi_{i-1}, T) &= \underbrace{R(\pi_i, \hat{T}) - R(\pi_{i-1}, \hat{T})}_{\text{m.-b. policy improvement}} \\ &+ \underbrace{R(\pi_i, T) - R(\pi_i, \hat{T}) + R(\pi_{i-1}, \hat{T}) - R(\pi_{i-1}, T)}_{\text{reward errors}}. \end{aligned} \tag{4}$$

Hence Line 5 makes an improvement on $R(\pi, T)$ (Eq. (4)$>$ 0) if the error in cumulative reward $|R(\pi, T) - R(\pi, \hat{T})|$ is small and the model-based policy improvement $R(\pi_i, \hat{T}) - R(\pi_{i-1}, \hat{T})$ is large. However, the model-based policy improvement is often theoretically intractable. This is because the policy optimization is usually conducted by deep RL algorithms (Fujimoto et al., 2018; Haarnoja et al., 2018) but state-of-the-art provable RL algorithms are still limited to linear function approximation (Jin et al., 2020; Duan et al., 2020). Therefore, we assume the model-based policy improvement is sufficiently large and focus on the error in the cumulative reward.

The desired closeness between π and π_D is achieved by Line 3's truncation and Line 5's constraint to a local ball B_{π_D} of π_D. Such closeness of policies is commonly used in the literature (Luo et al., 2019; Janner et al., 2019; Yu et al., 2020). Indeed, since $\hat{T}$ is fitted under π_D ($\mathcal{D}$'s distribution is $\rho_T^{\pi_D}$), if π and π_D are far apart, we cannot expect $\hat{T}$ behave like T under π. Practically, this is not a strong assumption, because we can algorithmically enforce closeness between π and π_D by constraining the KL divergence between π and π_D. Since it is much easier to control the policy error, the model error is the dominating error in MBRL. Hence we focus on the dependency of the cumulative model error w.r.t. the horizon.

3 Related Work

There have been many experimental studies of model-based RL. Evidence in Gu et al. (2016) and Nagabandi et al. (2018) suggests that for continuous control tasks, vanilla MBRL (Sutton, 1991) hardly surpasses model-free RL, unless using a linear transition model or a hybrid model-based and model-free algorithm. To enhance the applicability of MBRL, the Ensemble Method is widely adopted in the literature, since it helps alleviate overfitting in a neural network (NN) model. Instances of this approach include an ensemble of deterministic NN transition models (Kurutach et al., 2018), an ensemble of probabilistic NN transition models (Chua et al., 2018) with model predictive control (Camacho and Bordons Alba, 2013) or ensembles of deterministic NN for means and variances of rollouts with different horizons (Buckman et al., 2018). In addition to training multiple models, Clavera et al. (2018) leverages meta-learning to train a policy that can quickly adapt to new transition models. Wang et al. (2019) provides useful benchmarks of various model-based RL methods.

On the theoretical side, for stochastic state transitions the error in the cumulative reward is quadratic in the length of model rollouts. Specifically, Janner et al. (2019, Theorem A.1) provides the bound

$$R(\pi, T) - R(\pi, \hat{T}) \geq -\frac{2\gamma r^{\max}}{(1-\gamma)^2}(\epsilon_m + 2\epsilon_\pi) - \frac{4\epsilon_\pi r^{\max}}{1-\gamma}, \tag{5}$$

where $\epsilon_m = \max_t \mathbb{E}_{s,a\sim\rho_{\pi_D,t}} D_{TV}(T(\cdot|s,a) \,\|\, \hat{T}(\cdot|s,a))$, $\epsilon_\pi = \max_s D_{TV}(\pi_D(\cdot|s) \,\|\, \pi(\cdot|s))$ and $\rho_{\pi_D,t}$ is the density of (s,a) at step t following (ρ_0, π_D, T). For deterministic state transitions and an L-Lipschitz value function $V(s)$, Luo et al. (2019, Proposition 4.2) shows

$$\Big|R(\pi,T)-R(\pi,\hat{T})\Big| \leq \frac{\gamma}{1-\gamma} L \underset{\substack{s\sim\rho_T^{\pi_D}\\ a\sim\pi(\cdot|s)}}{\mathbb{E}} \|\bar{T}(s,a)-\hat{\bar{T}}(s,a)\| + 2\frac{\gamma^2}{(1-\gamma)^2}\delta \ \mathrm{diam}_{\mathcal{S}}, \tag{6}$$

where $\delta = \mathbb{E}_{s\sim\rho_T^{\pi}} \sqrt{D_{KL}(\pi(\cdot|s)\,\|\,\pi_D(\cdot|s))}$ and diam_S is the diameter of $\mathcal{S}$.

In practice, we enforce π_D and π to be close, so the model error terms dominate in Eq. (5) and (6). Eq. (6) looks sharper since the model error is correlated with a linear rather than quadratic term of the expected rollout length $(1-\gamma)^{-1}$. However, since the value function represents the cumulative reward, it's Lipschitz constant (assuming it exists) can be $O((1-\gamma)^{-1})$. So it is hard to compare Eq (5) and (6). While a Lipschitz value function is commonly assumed in the literature (Luo et al., 2019; Xiao et al., 2019; Yu et al., 2020), in practice, this property is hard to verify, and, even if it holds, the constant can be very large. To avoid strong assumptions, we do not assume a Lipschitz value function. In addition, we enhance the results of Janner et al. (2019), by showing their constants "in maxima" can be replaced by constants "in expectation".

A major contribution of Janner et al. (2019) is the use of branched rollouts generated by $(\rho_T^{\pi_D}, \pi, \hat{T})$. By Theorem 4.3 in (Janner et al., 2019), branched rollouts of length k satisfy

$$R(\pi,T) - R^{\text{branch}}(\pi) \geq -2r^{\max}\Big[\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{\gamma^k\epsilon_\pi}{1-\gamma} + \frac{k\epsilon_m}{1-\gamma}\Big], \tag{7}$$

with the same constants as Eq. (5). Eq. (7) implies that if the model is almost perfect ($\epsilon_m \approx 0$), the error is dominated by the policy error $\gamma^k\epsilon_\pi$. Since $\gamma < 1$, the minimal error is attained at large k. This suggests that a near perfect model helps correct the error due to the mismatch of the sampling policy and current policy. Still, with an almost perfect model, the problem is reduced to off-policy RL, which always suffers from policy mismatch error (Duan et al., 2020). Attaining the minimal error at large branch length k also contradicts the fact that the error accumulates over the trajectory (Wang et al., 2019; Xiao et al., 2019). The error propagation in the MBRL system implies the compounding error always increases with length. If we accept Eq. (7), there is still an important gap since Eq. (7) is for stochastic transitions, but the experiments in Janner et al. (2019) used deterministic transitions. Our analysis shows the error of branched rollouts for both stochastic and deterministic transitions increases in the expected branched length $(1-\beta)^{-1}$. So we always favor short lengths and are free from the issues mentioned above.

4 Main Result

As discussed in §2.2, we focus on the error in the cumulative reward $|R(\pi,T)-R(\pi,\hat{T})|$ in MBRL settings. To do so we use the triangle inequality:

$$|R(\pi,T)-R(\pi,\hat{T})| \leq \underbrace{|R(\pi,T)-R(\pi_D,T)|}_{\text{controlled by } \epsilon^T_{\pi_D,\pi}} + \underbrace{|R(\pi_D,T)-R(\pi_D,\hat{T})|}_{\text{controlled by } \epsilon^{\pi_D}_{T,\hat{T}} \text{ or } \epsilon_{\ell_2}} + \underbrace{|R(\pi_D,\hat{T})-R(\pi,\hat{T})|}_{\text{controlled by } \epsilon^{\hat{T}}_{\pi_D,\pi}}. \tag{8}$$

The error terms on the RHS of Eq. (8) result from policy mismatch (1st and 3rd terms) and transition mismatch (2nd term). Moreover, since these errors are controlled by the discrepancies between $T, \hat{T}$ and between π, π_D, the errors can be made small in the MBRL training (see the discussion in §2.2).

The discrepancy between policies π_D and π is measured by the total variation (TV) distance:

$$\epsilon^T_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho_T^{\pi_D}} D_{TV}\left(\pi_D(\cdot|s) \parallel \pi(\cdot|s)\right)$$
$$\text{and} \quad \epsilon^{\hat{T}}_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho_{\hat{T}}^{\pi}} D_{TV}(\pi_D(\cdot|s) \parallel \pi(\cdot|s)).$$

The expectation in $\epsilon^T_{\pi_D,\pi}$ is over $\rho_T^{\pi_D}$ and in $\epsilon^{\hat{T}}_{\pi_D,\pi}$ over $\rho_{\hat{T}}^{\pi}$. This allows us to measure policy discrepancy under the real dataset $\mathcal{D}$ (distributed as $\rho_T^{\pi_D}$) and the simulated environment (distributed as $\rho_{\hat{T}}^{\pi}$).

The discrepancies between real and learned transitions T, $\hat{T}$ are measured by (a) TV distance for stochastic transitions and (b) ℓ_2 error for deterministic ones.

$$\text{(a) } \epsilon^{\pi_D}_{T,\hat{T}} = \mathbb{E}_{(s,a)\sim\rho_T^{\pi_D}} D_{TV}\left(T(\cdot|s,a) \parallel \hat{T}(\cdot|s,a)\right)$$
$$\text{and (b) } \epsilon_{\ell_2} = \mathbb{E}_{(s,a)\sim\rho_{\overline{T}}^{\pi_D}} \|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\|_2.$$

From the RHS of Eq. (8), the policy mismatch errors (1st and 3rd terms) are invariant under exchange of π and π_D. We hence call these terms "symmetric in (π,π_D)". Similarly, the transition mismatch error (2nd term) is symmetric in $(T,\hat{T})$. However, the terms that control them, $\epsilon^T_{\pi_D,\pi}$, $\epsilon^{\hat{T}}_{\pi_D,\pi}$, and $\epsilon^{\pi_D}_{T,\hat{T}}$, ϵ_{ℓ_2} are asymmetric; the first two in (π,π_D) and the last in $(T,\hat{T})$. To bridge these symmetric and asymmetric quantities, we establish the following:

$$|R(\rho_1)-R(\rho_2)| \underset{(*)}{\leq} C\times\{D_{TV}(\rho_1||\rho_2) \text{ or } W_1(\rho_1\,\|\,\rho_2)\} \underset{(**)}{\leq} C'\times\{\epsilon^T_{\pi_D,\pi}, \epsilon^{\hat{T}}_{\pi_D,\pi}, \epsilon^{\pi_D}_{T,\hat{T}}, \text{ or } \epsilon_{\ell_2}\}. \tag{9}$$

Eq. (9) outlines the proof technique using notations in Eq. (1), with C, C' being underdetermined constants. Inequality $(*)$ upper bounds the cumulative reward error by one of the symmetric quantities w.r.t. occupancy measures $(\rho_1(s,a), \rho_2(s,a))$: TV distance for stochastic transitions or 1-Wasserstein distance (Villani, 2008) for deterministic transitions. Inequality $(**)$ upper bounds the symmetric quantities by one of the asymmetric ones, using the contraction of the Bellman flow operator (if it holds).

While the Bellman flow operator is a contraction w.r.t. TV distance, this may not hold w.r.t. W_1 distance. We address this situation in §4.3.2. Although we use W_1 distance as an intermediate step to analyze deterministic transitions, we finally upper bound W_1 distance by ℓ_2 error, as outlined by Inequality $(**)$. This avoids the need to minimize W_1 error using a Wasserstein GAN (Arjovsky et al., 2017) or other optimization techniques (Peyré and Cuturi, 2019).

In the following subsections, we first analyze the policy error, then the transition error when we have: (1) absolutely continuous stochastic transitions, (2) deterministic transitions with strong continuity, and (3) deterministic transitions with weak continuity. Cases (1) and (2) have Bellman contractions yielding sharp two-sided bounds. Case (3) uses a bounding technique inspired by Syed and Schapire (2010, Lemma 2) to establish a one-sided bound. By combining the policy error with the transition errors, we obtain corresponding MBRL errors. Full proofs are in the Appendix.

4.1 Symmetry Bridge Lemma and Policy Mismatch Error

We start by introducing a key lemma. Then we will use it to analyze the policy mismatch error.

4.1.1 Symmetry Bridge Lemma

Lemma 1 (following Conrad (2014, Corollary 2.4)) is a key to analyze both policy mismatch and transition mismatch errors through contractions.

Lemma 1. *Let B be a Bellman flow operator with fixed-point $\rho^\star$ and ρ be a state distribution. If B is a η-contraction w.r.t. some metric $\|\cdot\|$, then*

$$\|\rho - \rho^\star\| \leq \|\rho - B(\rho)\| / (1-\eta).$$

Recall from inequality $(**)$ of Eq. (9), we need to bridge from symmetric quantities to asymmetric ones. Lemma 1 constructs a bridge for this purpose. The LHS is symmetric (invariant under exchange) in $(\rho,\ \rho^\star)$. Also, the RHS is asymmetric in $(\rho,\ \rho^\star)$ because B is associated with $\rho^\star$. Hence, one can upper bound symmetric quantities using asymmetric ones if the contraction is given.

4.1.2 Policy Mismatch Error

The policy mismatch error is analyzed by Eq. (9). Note that the discrepancy between policies is measured by TV distance and that the Bellman flow operator is a contraction w.r.t. TV distance. Lemma 1 establishes the inequality $(**)$ of Eq. (9). The next Lemma is used to verify inequality $(*)$.

Lemma 2. *If $0 \leq r(s,a) \leq r^{\max}$, then*

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

The following theorem establishes the upper bounds of policy mismatch errors (1^{st} and 3^{rd} terms of Eq. (8)). Lemmas 1 and 2 are used in its proof.

Theorem 1. *If $0 \leq r(s,a) \leq r^{\max}$ and $\epsilon^T_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\pi_D}_T}[D_{TV}(\pi_D(\cdot|s) \,\|\, \pi(\cdot|s))]$, then*

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

Proof Sketch. By Lemma 2, it is enough to upper bound $D_{TV}(\rho^{\pi_D}_T(s,a) || \rho^{\pi}_T(s,a))$:

$$\begin{aligned}
&D_{TV}(\rho^{\pi_D}_T(s,a) \| \rho^{\pi}_T(s,a)) \\
\leq &D_{TV}(\rho^{\pi_D}_T(s)\pi_D(a|s) \| \rho^{\pi_D}_T(s)\pi(a|s)) \\
&+ D_{TV}(\rho^{\pi_D}_T(s)\pi(a|s) \| \rho^{\pi}_T(s)\pi(a|s)) \\
\leq &\epsilon^T_{\pi_D,\pi} + \frac{1}{1-\gamma} D_{TV}(B^{\pi_D}_T(\rho^{\pi_D}_T(s)) \| B^{\pi}_T(\rho^{\pi_D}_T(s))) \\
\leq &\epsilon^T_{\pi_D,\pi} + \frac{\gamma}{1-\gamma} \epsilon^T_{\pi_D,\pi},
\end{aligned}$$

where the second inequality follows from Lemma 1 and the fixed-point property. □

4.1.3 Application to Imitation Learning

We now make the following interesting side observation. Imitation learning (Syed and Schapire, 2010; Ho and Ermon, 2016) is matching the demonstrated policy and the generator policy. Because Theorem 1 is about policy mismatch error, it is applicable to imitation learning. Observe that the objective of GAIL is JS (Jensen-Shannon) divergence when its discriminator is well-trained and that Behavior Cloning's objective is KL divergence. We can use Pinsker's Inequality to upper bound these divergences and translate Theorem 1 and Lemma 2, yielding:

Corollary 1 (Error of Behavioral Cloning)**.** *Let π_D and π be the expert and agent policy. If $0 \leq r(s,a) \leq r^{\max}$ and $\mathbb{E}_{s\sim\rho^{\pi_D}_T} D_{KL}(\pi_D(\cdot|s) \| \pi(\cdot|s)) \leq \epsilon_{BC}$, then*

$$|R(\pi_D, T) - R(\pi, T)| \leq \sqrt{\epsilon_{BC}/2}\ r^{\max} \Big(\tfrac{1}{1-\gamma} + \tfrac{\gamma}{(1-\gamma)^2} \Big).$$

Corollary 2 (Error of GAIL)**.** *Let π_D and π be the expert and agent policy, respectively. If $0 \le r(s,a) \le r^{\max}$ and $D_{JS}(\rho^{\pi_D}_T \| \rho^{\pi}_T) \le \epsilon_{GAIL}$, then*

$$|R(\pi_D, T) - R(\pi, T)| \le \sqrt{2\epsilon_{GAIL}}\, r^{\max}/(1-\gamma).$$

Observe that Behavioral Cloning's error is quadratic w.r.t. the expected horizon $(1-\gamma)^{-1}$ while GAIL's is linear. This suggests that when the discriminator is well-trained, GAN-style imitation learning, like GAIL, has an advantage.

4.2 MBRL with Stochastic Transitions

If the true transitions are stochastic, we can learn $\hat{T}$ by maximizing the likelihood, or equivalently by minimizing the KL divergence. To ensure the KL divergence is defined on a continuous state space, we assume the transition probability is absolutely continuous w.r.t. the state space, i.e., there is a density function and hence no discrete or singular continuous measures (Hewitt and Ross, 1963). The following theorem then follows by the proof of Theorem 1.

Theorem 2. *If $0 \le r(s,a) \le r^{\max}$ and $\epsilon^{\pi_D}_{T,\hat{T}} = \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_T}[D_{TV}(T(\cdot|s,a)\|\hat{T}(\cdot|s,a))]$, then*

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

Theorems 1 and 2 yield the following result for MBRL with absolutely continuous stochastic transitions.

Corollary 3. *Assume $0 \le r(s,a) \le r^{\max}$ and let*

$$\epsilon^{\pi_D}_{T,\hat{T}} \triangleq \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_T} D_{TV}(T(\cdot|s,a) \parallel \hat{T}(\cdot|s,a))$$
$$\epsilon^{T}_{\pi_D,\pi} \triangleq \mathbb{E}_{s\sim\rho^{\pi_D}_T} D_{TV}(\pi_D(\cdot|s) \parallel \pi(\cdot|s))$$
$$\epsilon^{\hat{T}}_{\pi_D,\pi} \triangleq \mathbb{E}_{s\sim\rho^{\pi}_{\hat{T}}} D_{TV}(\pi_D(\cdot|s) \parallel \pi(\cdot|s)).\ \textit{Then}$$

$$|R(\pi,T) - R(\pi,\hat{T})| \le (\epsilon^{\pi_D}_{T,\hat{T}} + \epsilon^{T}_{\pi_D,\pi} + \epsilon^{\hat{T}}_{\pi_D,\pi})\tfrac{r^{\max}\gamma}{(1-\gamma)^2} + (\epsilon^{T}_{\pi_D,\pi} + \epsilon^{\hat{T}}_{\pi_D,\pi})\tfrac{r^{\max}}{(1-\gamma)}.$$

Comparing the result in Corollary 3 with the prior results in Eq. (5), we sharpen the bounds by changing the constants from maxima to expected values.

4.2.1 MBRL with Branched Rollouts

Corollary 3 indicates that the model error term $\epsilon^{\pi_D}_{T,\hat{T}} r^{\max}\gamma/(1-\gamma)^2$ is quadratic w.r.t. the expected rollout length $(1-\gamma)^{-1}$, which makes MBRL undesirable for long rollouts and leads to the planning horizon dilemma. An intuitive countermeasure is to use short rollouts that share similar distributions with the long ones. This leads to the idea of branched rollouts. Throughout the rest of the paper, $\beta > 0$ will denote the branched discount factor with $\beta < \gamma$. We define a *branched rollout* with discount factor β, to be a rollout following the laws of $(\rho^{\pi_D}_{T,\gamma},\ \pi,\ \hat{T})$. Intuitively, these are rollouts initialized on the states of previous real long rollouts, $\rho^{\pi_D}_{T,\gamma}(s)$, and then run a few steps under policy π and model $\hat{T}$.

The occupancy measure of branched rollouts is $\rho^{\rho^{\pi_D}_{T,\gamma},\pi}_{\hat{T},\beta}$ where the superscripts $\rho^{\pi_D}_{T,\gamma}, \pi$ indicate the initial state distribution and policy, and the subscripts $\hat{T}, \beta$ indicate the transition and discount factor. Branched rollouts are short by construction, but it is unclear whether their distribution is similar to that of long rollouts. This is verified by the following Lemma.

Lemma 3. *Let $\gamma > \beta$ be the discount factors of long and short rollouts, and π_D and T be the sampling policy and the real transition. Then*

$$D_{TV}(\rho^{\pi_D}_{T,\gamma} \parallel \rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}) \le (1-\gamma)\beta/(\gamma-\beta).$$

By Lemma 3, $\rho^{\pi_D}_{T,\gamma}$ and $\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}$ are close if β is small. Hence once the pairs (π,π_D) and $(T,\hat{T})$ are close, the distribution of branched rollouts is similar to that of long real rollouts, and the error in cumulative reward is small. This is given in detail below.

Corollary 4. *Let $0 \le r(s,a) \le r^{\max}$,*
$\epsilon^{T,\gamma}_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\pi_D}_{T,\gamma}} D_{TV}(\pi_D(\cdot|s) \parallel \pi(\cdot|s))$,
$\epsilon^{\hat{T},\beta}_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\rho^{\pi_D}_{T,\gamma},\pi}_{\hat{T},\beta}} D_{TV}(\pi_D(\cdot|s) \parallel \pi(\cdot|s))$, *and*
$\epsilon^{\pi_D,\beta}_{T,\hat{T}} = \mathbb{E}_{(s,a)\sim\rho^{\rho^{\pi_D}_{T,\gamma},\pi_D}_{T,\beta}} D_{TV}(T(\cdot|s,a) \parallel \hat{T}(\cdot|s,a))$.
Then
$\left|R_\gamma(\rho_0,\pi,T) - \frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi,\hat{T})\right| \le r^{\max}\Big(\frac{\epsilon^{T,\gamma}_{\pi_D,\pi}\gamma}{(1-\gamma)^2} + \frac{(\epsilon^{\pi_D,\beta}_{T,\hat{T}} + \epsilon^{\hat{T},\beta}_{\pi_D,\pi})\beta}{(1-\beta)(1-\gamma)} + \frac{\epsilon^{T,\gamma}_{\pi_D,\pi} + \epsilon^{\hat{T},\beta}_{\pi_D,\pi}}{1-\gamma} + \frac{\beta}{\gamma-\beta}\Big)$.

Proof Sketch. Decompose the error as follows and then apply Theorems 1, 2, and Lemmas 2, 3.

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

Notice $\epsilon^{\pi_D,\beta}_{T,\hat{T}}$ is controlled by supervised learning since it is evaluated on dataset $\mathcal{D}$. Because branched rollouts are shorter than normal rollouts, the branch cumulative reward is rescaled to $\frac{1-\beta}{1-\gamma}R_\beta(\rho^{\pi_D}_{T,\gamma},\pi,\hat{T})$

for comparison to normal rollouts. Compared with Corollary 3, the model error term's dependency on the rollout lengths is reduced from $O((1-\gamma)^{-2})$ to $O((1-\gamma)^{-1}(1-\beta)^{-1})$. This shows that branched rollouts greatly reduce the cumulative reward error.

Corollary 4 shows the error in cumulative reward increases in β, or equivalently in the expected branched length $(1-\beta)^{-1}$. Thus our result is free from the issue of previous work Eq. (7) (Note: $(1-\beta)^{-1}$ corresponds to the branch length k in Eq. (7)). It is tempting to set $\beta = 0$ to minimize the reward error. However, if $\beta = 0$, each branched rollout is composed of a single point drawn from $\rho^{\pi_D}_{T,\gamma}$. This means that the branched rollouts access neither T nor $\hat{T}$, so we will learn a policy that only optimizes on initial states and has no concern for the future. For example, the reward of MuJoCo environment (Todorov et al., 2012) is typically $r(s,a) = \text{velocity}(s) - \|a\|_2^2$. To maximize cumulative reward on branched rollouts with $\beta = 0$, the optimal policy will shortsightedly select $a = 0$ for any s.

The branched rollout makes a trade-off between policy improvement and reward error, as discussed in §2.2. The policy improvement $R_\beta(\rho^{\pi_D}_T, \pi_i, \hat{T}) - R_\beta(\rho^{\pi_D}_T, \pi_{i-1}, \hat{T})$ in branched rollouts benefits from a larger β, while the reward error, as shown in Corollary 4 and 5, favors smaller β. In MuJoCo, Janner et al. (2019, Appendix C) says the branched length is chosen as 2 in early epochs and may stay small or gradually increase to 16 or 26 later. This suggests for continuous-control (MuJoCo) tasks, $\beta \approx 0.9$ is enough to balance policy improvement and reward error.

4.3 MBRL with Deterministic Transitions

We discuss deterministic transitions under (a) strong, and (b) weak Lipschitz assumptions. The main difference is the validity of Lemma 4, which is controlled by the smoothness of the deterministic transition.

4.3.1 Strong Lipschitz Continuity

A major difficulty in analyzing deterministic transitions is that TV distance is not suitable for comparing $\overline{T}$ and $\hat{\overline{T}}$. Indeed, for any fixed (s,a), $D_{TV}(\delta(s' - \overline{T}(s,a)) \parallel \delta(s' - \hat{\overline{T}}(s,a))) = 1$ once $\overline{T}(s,a) \neq \hat{\overline{T}}(s,a)$. Moreover, the model error is controlled by $\epsilon_{\ell_2} = \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_{\overline{T}}} \|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\|_2$, but the ℓ_2 error is not a distance metric for distributions. To control the distance between distributions through an ℓ_2 error, we can select a distance metric for distributions that is upper bounded by ℓ_2 error. The 1-Wasserstein distance is a good candidate:

$$W_1(\rho_1(s) \parallel \rho_2(s)) = \inf_{J(s_1,s_2)\in\Pi(\rho_1,\rho_2)} \mathbb{E}_J \|s_1 - s_2\|_2 \quad (10)$$

where the infimum is over joint distributions $J(s_1, s_2)$ with marginals $\rho_1(s_1)$, $\rho_2(s_2)$. To apply Eq. (9), it is crucial to use a metric under which the Bellman flow operator is a contraction. To ensure this holds for W_1 distance, we make the following Lipschitz assumptions on the transitions and policies.

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 $\text{diam}_{\mathcal{A}}$) 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}, $\Sigma^{1/2}_{\pi_D}$, $\Sigma^{1/2}_{\pi}$ are $L_{\pi,\mu}$, $L_{\pi_D,\mu}$, $L_{\pi,\Sigma}$, $L_{\pi_D,\Sigma}$ Lipschitz w.r.t. states.

In (1.3), $\mathcal{P}_{\mathcal{A}}$ is the projection to $\mathcal{A}$ and in (1.4) $\|\Sigma^{1/2}_\pi(s) - \Sigma^{1/2}_\pi(s')\| \leq L_{\pi,\Sigma} \|s - s'\|_2$. Assumption 1 is easily satisfied in most continuous control tasks, as explained in §4 of the Appendix. The harder one, which will be resolved later, is $\gamma\eta_{\pi,\overline{T}} < 1$ in Lemma 4.

Lemma 4. *If Assumption 1 holds, and* $\eta_{\pi,\overline{T}} = L_{\overline{T},s} + L_{\overline{T},a}(L_{\pi,\mu} + L_{\pi,\Sigma}\sqrt{\dim_{\mathcal{A}}}) < 1/\gamma$, *then* $B_{\pi,\overline{T}}$ *is a* $\gamma\eta_{\pi,\overline{T}}$*-contraction w.r.t. 1-Wasserstein distance.*

To verify there exists a nontrivial system such that the condition $\gamma\eta_{\pi,\overline{T}} < 1$ in Lemma 4 holds under Assumption 1, we consider a continuous-control task. The key term depends on the sample interval Δ. Let $s = [x, v]^\top = [\text{position, velocity}]^\top$, and $a =$ acceleration. By the laws of motion,

$$\begin{aligned} s' = \begin{bmatrix} x' \\ v' \end{bmatrix} &= \begin{bmatrix} x + v\Delta + \frac{1}{2}a\Delta^2 \\ v + a\Delta \end{bmatrix} \\ &= \begin{bmatrix} I & I\Delta \\ 0 & I \end{bmatrix} s + \begin{bmatrix} I\frac{1}{2}\Delta^2 \\ I\Delta \end{bmatrix} a = \overline{T}(s,a) \end{aligned} \quad (11)$$

This shows $L_{\overline{T},s} = 1 + O(\Delta)$, $L_{\overline{T},a} = O(\Delta)$ and $\eta_{\pi,\overline{T}} = 1 + O(\Delta)$. Therefore, we conclude that $\gamma\eta_{\pi,\overline{T}} = \gamma + O(\Delta) < 1$ for small enough Δ.

If Lemma 4 holds for $\hat{\overline{T}}$, we can apply Eq. (9): measure error in W_1 distance, apply contraction on W_1 to get an asymmetric bound (Lemma 1) and then upper bound W_1 distance by ℓ_2 error. This gives the following Theorem for deterministic transitions.

Theorem 3. *Under Lemma 4, if* $r(s,a)$ *is* L_r*-Lipschitz and the* ℓ_2 *error is* ϵ_{ℓ_2}*, then*

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

The typical MuJoCo reward, $r(s,a) = \text{velocity}(s) - \|a\|_2^2$, is Lipschitz if the diameter $\text{diam}_{\mathcal{A}}$ is finite. As MuJoCo also provides the bounds for the action space $\mathcal{A}$, the Lipschitz assumption on $r(s,a)$ is usually satisfied in MuJuCo continuous-control tasks.

Theorem 3, in conjunction with the calculation of η in Eq. (11), indicates that the cumulative model error decreases as the sampling interval Δ becomes smaller. This is because the cumulative model error decreases in $\eta_{\pi,\hat{\overline{T}}}$ and $\eta_{\pi,\hat{\overline{T}}} = 1 + O(\Delta)$. Hence, as expected, the sampling period of a continuous-control task has to be small enough in order to train an MBRL system.

Although Theorem 3 requires Lemma 4's strong assumption, branched rollouts allow this assumption to be satisfied since branched rollouts use a much smaller discount factor. Thus, one might expect a benefit from using branched rollouts with deterministic transitions. This is validated in the following Corollary.

Corollary 5. *Let $r(s,a)$ be L_r-Lipschitz and bounded: $0 \le r(s,a) \le r^{\max}$. Let*

$\epsilon^{\overline{T},\gamma}_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\pi_D}_{\overline{T},\gamma}} D_{TV}(\pi_D(\cdot|s) \,\|\, \pi(\cdot|s))$,

$\epsilon^{\hat{\overline{T}},\beta}_{\pi_D,\pi} = \mathbb{E}_{s\sim\rho^{\rho^{\pi_D}_{\overline{T},\gamma},\pi}_{\hat{\overline{T}},\beta}} D_{TV}(\pi_D(\cdot|s) \,\|\, \pi(\cdot|s))$,

$\epsilon_{\ell_2,\beta} = \mathbb{E}_{(s,a)\sim\rho^{\rho^{\pi_D}_{\overline{T},\gamma},\pi_D}_{\overline{T},\beta}} \|\overline{T}(s,a) - \hat{\overline{T}}(s,a)\|_2$. *Then,*

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

Corollary 5 shows an additional benefit of branched rollouts: to ensure $\beta\eta_{\pi_D,\hat{\overline{T}}} < 1$, by choosing a small β (say 0.9). This suggests that branched rollouts are particularly useful for deterministic transitions. Such a suggestion on branched length (or equivalently, the branched discount factor β) supports the experimental success of Janner et al. (2019) and their choice of hyperparameter, as mentioned in the last paragraph of § 4.2. Also, this result is for deterministic transitions, so this resolves an open issue in Janner et al. (2019), as they proved for stochastic transitions but experimented with deterministic transitions.

4.3.2 Weak Lipschitz Continuity

When Lemma 4 is invalid, there is no Bellman contraction, and we cannot use the bounding principle in Eq. (9). We provide another way to analyze the error, giving a weaker one-sided bound.

We cannot expect much when $L_{\hat{\overline{T}},s} \gg 1$, since the rollout diverges when being repeatedly applied to $\hat{\overline{T}}$, with the error growing exponentially w.r.t. rollout length. Hence in this subsection we assume $L_{\hat{\overline{T}},s} \le 1 + (1-\gamma)\iota$ with $\iota < 1$; i.e., the Lipschitzness of the transition w.r.t. state is slightly higher than 1. The longer the expected length $(1-\gamma)^{-1}$, the smoother $\hat{\overline{T}}$ should be.

The following theorem reveals the impact of asymmetry when there is no Bellman contraction.

Theorem 4. *Let $0 \le r(s,a) \le r^{\max}$ and $\epsilon_{\ell_2} = \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_{\overline{T}}} \|\overline{T} - \hat{\overline{T}}\|_2$. Assume that:*

(a) $\hat{\overline{T}}(s,a)$, $r(s,a)$, $\pi_D(a|s)$ are Lipschitz in s for any a with constants $L_{\hat{\overline{T}},s}$, $L_{r,s}$, $L_{\pi_D,s}$.

(b) $L_{\hat{\overline{T}},s} \le 1 + (1-\gamma)\iota$ with $\iota < 1$.

(c) The action space is bounded: $\text{diam}_{\mathcal{A}} < \infty$. Then,

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

Theorem 4 is a one-sided bound resulting from the asymmetry of $\epsilon_{\ell_2} = \mathbb{E}_{(s,a)\sim\rho^{\pi_D}_{\overline{T}}} \|\overline{T} - \hat{\overline{T}}\|_2$: $\mathbb{E}$ is taken biasedly on $\rho^{\pi_D}_{\overline{T}}$, so we can only upper bound $R(\pi_D,\overline{T})$ by $R(\pi_D,\hat{\overline{T}}) + O((1-\gamma)^{-5/2})$. The resulting MBRL error only ensures that a policy that works well on $\overline{T}$ also works on $\hat{\overline{T}}$, but not the other way around. This one-sided nature may allow $\hat{\overline{T}}$ to overfit the data. This supports the use of the Ensemble Method (Kurutach et al., 2018) to mitigate model bias by training multiple independent models.

Theorem 4 only indicates the consequence of the ϵ_{ℓ_2} objective's asymmetry. However, this is avoidable. As discussed in Corollary 5, branched rollouts provide a Bellman contraction and hence two-sided bounds.

5 Conclusion

Using a Bellman flow contraction w.r.t. distance metrics of probability distributions, we have provided results on the cumulative reward error in MBRL for both stochastic and deterministic transitions. In particular, absolutely continuous stochastic transitions and deterministic transitions with strong Lipschitz continuity have Bellman contractions. This result suggests that MBRL is better suited to these situations. The difficulty of dealing with deterministic transitions that do not yield a Bellman contraction arises from the objective function's asymmetry. Finally, we prove that branched rollouts can significantly reduce the error of MBRL and allow a Bellman contraction under deterministic transitions.

References

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

Buckman, J., D. Hafner, G. Tucker, E. Brevdo, and H. Lee
2018. Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Pp. 8224–8234. Curran Associates, Inc.

Camacho, E. F. and C. Bordons Alba
2013. *Model Predictive Control*. Springer Science & Business Media.

Chua, K., R. Calandra, R. McAllister, and S. Levine
2018. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Pp. 4754–4765. Curran Associates, Inc.

Clavera, I., J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel
2018. Model-based reinforcement learning via meta-policy optimization. In *Conference on Robot Learning (CORL)*.

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

Duan, Y., Z. Jia, and M. Wang
2020. Minimax-optimal off-policy evaluation with linear function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, Pp. 2701–2709, Virtual. PMLR.

Fujimoto, S., H. van Hoof, and D. Meger
2018. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, J. G. Dy and A. Krause, eds., volume 80 of *Proceedings of Machine Learning Research*, Pp. 1582–1591. PMLR.

Gu, S., T. Lillicrap, I. Sutskever, and S. Levine
2016. Continuous deep q-learning with model-based acceleration. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, P. 2829–2838. JMLR.org.

Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine
2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, J. G. Dy and A. Krause, eds., volume 80 of *Proceedings of Machine Learning Research*, Pp. 1856–1865. PMLR.

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

Ho, J. and S. Ermon
2016. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., Pp. 4565–4573. Curran Associates, Inc.

Janner, M., J. Fu, M. Zhang, and S. Levine
2019. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems 32*, Pp. 12498–12509. Curran Associates, Inc.

Jin, C., Z. Yang, Z. Wang, and M. I. Jordan
2020. Provably efficient reinforcement learning with linear function approximation. volume 125 of *Proceedings of Machine Learning Research*, Pp. 2137–2143. PMLR.

Kurutach, T., I. Clavera, Y. Duan, A. Tamar, and P. Abbeel
2018. Model-ensemble trust-region policy optimization. In *International Conference on Learning Representations*.

Luo, Y., H. Xu, Y. Li, Y. Tian, T. Darrell, and T. Ma
2019. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *International Conference on Learning Representations*.

Nagabandi, A., G. Kahn, R. S. Fearing, and S. Levine
2018. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Pp. 7559–7566.

Peyré, G. and M. Cuturi
2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Ross, S., G. Gordon, and D. Bagnell
2011. A reduction of imitation learning and structured prediction to no-regret online learning. volume 15 of *Proceedings of Machine Learning Research*, Pp. 627–635, Fort Lauderdale, FL, USA. JMLR Workshop and Conference Proceedings.

Sutton, R. S.
1991. Dyna, an integrated architecture for learn-

ing, planning, and reacting. *SIGART Bull.*, 2(4):160–163.

Syed, U., M. Bowling, and R. E. Schapire
2008. Apprenticeship learning using linear programming. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, P. 1032–1039, New York, NY, USA. Association for Computing Machinery.

Syed, U. and R. E. Schapire
2010. A reduction from apprenticeship learning to classification. In *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds., Pp. 2253–2261. Curran Associates, Inc.

Todorov, E., T. Erez, and Y. Tassa
2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Pp. 5026–5033.

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

Wang, T., X. Bao, I. Clavera, J. Hoang, Y. Wen, E. Langlois, S. Zhang, G. Zhang, P. Abbeel, and J. Ba
2019. Benchmarking model-based reinforcement learning. arXiv:1907.02057.

Xiao, C., Y. Wu, C. Ma, D. Schuurmans, and M. Müller
2019. Learning to combat compounding-error in model-based reinforcement learning. arXiv:1912.11206.

Yu, T., G. Thomas, L. Yu, S. Ermon, J. Zou, S. Levine, C. Finn, and T. Ma
2020. Mopo: Model-based offline policy optimization. arXiv:2005.13239.