When to Trust Your Model: Model-Based Policy Optimization

Chen Yen Ju
Department of Applied Mathematics
National Yang Ming Chiao Tung University
nctu8888.sc07@nycu.edu.tw

1 Introduction

Please provide a clear but brief overview of the selected paper. You may want to discuss the following aspects:

- The main research challenges tackled by the paper

  Reinforcement learning algorithms can be easily divide into two cases, that is model-based approaches and model-free approaches.

  Model-free method interact with the environment directly while model-based method aims to build a predictive model of an environment. Moreover, there will also be some trade-off among these two approaches. Using model-free method seems easier when the environment is complicated because of the difficulty of building a predictive model. And Using model-free method have shown promise as a general-purpose tool (Mnih et al. [2015]; Lillicrap et al. [2015]; Haarnoja et al. [2018]). But it also learn slowly especially when data collection is an arduous process, leading to the cost of efficiency.

  Therefore, people started to develop the part of model-based method due to their comparatively fast learning. However, the difficulty of model-based method will be the model accuracy. Once the model accuracy is worse, it is just like learning a method under the different environment, which will lead to a useless learning process. And because of this difficulty, model-based method often perform worse asymptotically than their model-free counterparts.

  So, how could we make great use of model-based method will be the main research challenges in this paper.

- The high-level technical insights into the problem of interest

  Since the main research challenges in this paper is to make great use of model-based method, this paper investigate the way about how to guarantee the monotonic improvement, i.e. make sure that model is surely improving during training epoch.

  Hence, this paper investigate the relationship between the returns from the true MDP and the predictive MDP, and construct an inequality between those two returns with the help of "model error, ϵ_m" and "policy error, ϵ_π".

  After that, we can clearly take model error into consideration. To say more, if I want to improve the policy and the model simultaneously, then what I have to do is to avoid the model error influencing the whole training step. So if I can improve the return greater

than some number (we call it C in this paper, it's a function of ϵ_m and ϵ_π), than I can ease the bias of model generated data. (See more in equation 2)

- The main contributions of the paper (compared to the prior works)

  The main contributions of the paper is a practical algorithm built on the insights we've mentioned above. And we call this algorithm model-based policy optimization (MBPO), that makes limited use of a predictive model to achieve pronounced improvements in performance compared to other model-based approaches. To say more, this algorithm disentangle the the task horizon and model horizon by querying the model only for short rollouts. And at last, the paper investigate different strategies of model usage and find that the careful use of large amount of short model-generated rollouts can make the most benefit to a algorithm.

- Your personal perspective on the proposed method

  I think that the main contribution of this paper is not only the MBPO algorithm but also the behind derivation of the relationship between the returns from the true MDP and the predictive MDP. This will lead to a great improvement of model-based method and allow others to take more consideration into model-based approaches.

2 Problem Formulation

Please present the formulation in this section. You may want to cover the following aspects:

- Your notations (e.g. MDPs, value functions, function approximators,...etc)

  We consider a Markov decision process (MDP), defined by the tuple $(\boldsymbol{S}, \boldsymbol{A}, p, r, \gamma, \rho_0)$. $\boldsymbol{S}$ and $\boldsymbol{A}$ are the state and action spaces, respectively, and $\gamma \in (0, 1)$ is the discount factor. The dynamics or transition distribution are denoted as $p(s'|s, a)$, the initial state distribution as $p_0(s)$, and the reward function as $r(s, a)$. The goal of reinforcement learning is to find the optimal policy π^* that maximizes the expected sum of discounted rewards, denoted by η:

$$\pi^* = \underset{\pi}{argmax}\, E_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \tag{1}$$

- The optimization problem of interest

  The goal of this paper is to outline a principled framework in which we can provide performance guarantees for model-based algorithms. To show monotonic improvement for a model-based method, we wish to construct a bound of the following form:

$$\eta\left[\pi\right] \geq \hat{\eta}\left[\pi\right] - C. \tag{2}$$

  $\eta\left[\pi\right]$ denotes the returns of the policy in the true MDP, whereas $\hat{\eta}\left[\pi\right]$ denotes the returns of the policy under our predictive model.
  Moreover, the notation C will be the function of "model error, ϵ_m" and "policy error, ϵ_π" we've mentioned in the Lemmas and Theorems.

  Such a statement guarantees that, as long as we improve by at least C under the model, we can guarantee improvement on the true MDP.

  In my humble opinion, the guarantee improvement can be shown by the inequality

below:

$$\begin{aligned}\eta\left[\pi_{t+k}\right] &\overset{(2)}{\geq} \hat{\eta}_{t+k}\left[\pi\right] - C \\ &\geq \hat{\eta}_{t+k-1}\left[\pi\right] \\ &\geq \hat{\eta}_{t+k-2}\left[\pi\right] + C \\ &\geq \hat{\eta}_{t+k-3}\left[\pi\right] + 2C \\ &\geq \dots \\ &\geq \hat{\eta}_{t}\left[\pi\right] + (k-1)C.\end{aligned} \tag{3}$$

Hence once we can improve by at least C under the model, from the above equation we may find that the lower bound of returns in the true MDP will keep growing greater and greater, leading to the guarantee improvement we've mentioned above.

- The technical assumptions

  The dynamics $p(s'|s,a)$ are assumed to be unknown. Model-based reinforcement learning methods aim to construct a model of the transition distribution, that is, to construct the transition matrix $p_\theta(s'|s,a)$, using data collected from interaction with the environment. We additionally assume that the reward function has unknown form, and predict reward function under state s and action a.

  The upper bound of model error, ϵ_m, exists as a constant, by (Shalev-Shwartz and Ben-David [2014];)

- Preliminaries

  Before we start to conduct the theoretical proof, I think we have to get more sense about how MBPO works and what we want to show:

  First, let's take a look at the Algorithm 1 below, it's the original form about MBPO. And from the algorithm we can find that MBPO is just an optimization algorithm by training predictive model and policy simultaneously. The main key is at row 5, if we combine row 5 and equation 2 and 3, than we can understand the main target of this paper, which is improvement guarantee. Moreover, we can also find that the value function C we are looking for is the function of ϵ_m and ϵ_π, and its explicit form is derived in "Theorem A.1".

Algorithm 1 Monotonic Model-Based Policy Optimization

1: Initialize policy $\pi(a|s)$, predictive model $p_\theta(s', r|s, a)$, empty dataset $\mathcal{D}$.
2: **for** N epochs **do**
3: Collect data with π in real environment: $\mathcal{D} = \mathcal{D} \cup \{(s_i, a_i, s'_i, r_i)\}_i$
4: Train model p_θ on dataset $\mathcal{D}$ via maximum likelihood: $\theta \leftarrow \text{argmax}_\theta \mathbb{E}_\mathcal{D}[\log p_\theta(s', r|s, a)]$
5: Optimize policy under predictive model: $\pi \leftarrow \text{argmax}_{\pi'} \hat{\eta}[\pi'] - C(\epsilon_m, \epsilon_\pi)$
6: **end for**

Although "Theorem A.1" provides a useful explicit form of function C, it will still be some problem that "What if the model error, ϵ_m, is too high that we cannot find a policy satisfying $|\eta[\pi] - \hat{\eta}[\pi]| \geq C$?"

Then the original MBPO algorithm might not work under this condition. Hence the algorithm must be modified to solve the ϵ_m problem. Consider a infinite Markov Decision Process, since we are using model-based method, there will be model error ϵ_m during every training step. And the infinite sequence will make the model error ϵ_m tempt to compound, leading to the tragedy we've mentioned above (that is, ϵ_m problem).

To solve this problem, we introduce the notion of a branched rollout, which takes only k step instead of going through the entire trajectory. More precisely, we start with the

state sampled from the previous state distribution and run k steps under the learned model. And the corresponding algorithm is listed below.

Under such a scheme, the model error ϵ_m won't be compounded with infinitely sequence, and the new C function can be found using "Theorem A.3".

Algorithm 2 Model-Based Policy Optimization with Deep Reinforcement Learning

Initialize policy π_ϕ, predictive model p_θ, environment dataset $\mathcal{D}_{\text{env}}$, model dataset $\mathcal{D}_{\text{model}}$
for N epochs **do**
  Train model p_θ on $\mathcal{D}_{\text{env}}$ via maximum likelihood
  for E steps **do**
    Take action in environment according to π_ϕ; add to $\mathcal{D}_{\text{env}}$
    for M model rollouts **do**
      Sample s_t uniformly from $\mathcal{D}_{\text{env}}$
      Perform k-step model rollout starting from s_t using policy π_ϕ; add to $\mathcal{D}_{\text{model}}$
    end for
    for G gradient updates **do**
      Update policy parameters on model data: $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi, \mathcal{D}_{\text{model}})$
    end for
  end for
end for

Finally, the paper has conducted some experiments, trying to modify the bounds, i.e. the value of C that we got from "Theorem A.3".

The paper found that whenever the amount of the collecting data is large enough, the derivative of model error on the distribution of the current policy w.r.t the policy error will approach 0.

And since we can adopted linear function to approximate the model error on the distribution of the current policy, $\hat{\epsilon}_{m'}$, that is, $\hat{\epsilon}_{m'} \approx \epsilon_m + \epsilon_\pi \frac{d\epsilon_{m'}}{d\epsilon_\pi}$. So we can neglected the last term since it approaches zero to get: $\hat{\epsilon}_{m'} \approx \epsilon_m$.

This result implies that we can substitute the last term $(\epsilon_m + 2\epsilon_\pi)$ with $\epsilon_{m'}$ since the policy error term will be 0 under the distribution of the current policy. At last, we used "Theorem A.2" to modify the bound in order to motivate model usage.

3 Theoretical Analysis

Please present the theoretical analysis in this section. Moreover, please formally state the major theoretical results using theorem/proposition/corollary/lemma environments. Also, please clearly highlight the errors that you found as well as your new proofs or extensions (if any).

We are going to prove three theorems in this section. And before we prove these theorems, we have to prove some useful lemmas first:

Lemma B.1 (TVD of Joint Distributions).
Suppose we have two distributions $p_1(x, y) = p_1(x)p_1(y|x)$ and $p_2(x, y) = p_2(x)p_2(y|x)$. We can bound the total variation distance of the joint as:

$$D_{TV}(p_1(x, y) \parallel p_2(x, y)) \leq D_{TV}(p_1(x) \parallel p_2(x)) + \max_x D_{TV}(p_1(y|x) \parallel p_2(y|x))$$

Alternatively, we have a tighter bound in terms of the expected TVD of the conditional:

$$D_{TV}(p_1(x, y) \parallel p_2(x, y)) \leq D_{TV}(p_1(x) \parallel p_2(x)) + E_{x\sim p_1}[D_{TV}(p_1(y|x) \parallel p_2(y|x))]$$

Used Properties.

Conditional probability:

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{4}$$

Triangle inequality:

$$|a + b| \leq |a| + |b| \tag{5}$$

Property of summation:

$$\sum_{x,y} f(x)g(y|x) = \sum_x \left[f(x) \sum_y g(y|x) \right] \tag{6}$$

Summation of conditional probability:

$$\sum_y p(y|x) = \sum_y \frac{p(x, y)}{p(x)} = \frac{p(x)}{p(x)} = 1 \tag{7}$$

Expectation and Maximum:

$$\begin{aligned} E_{x\sim p_1}[f(x)] &= \sum_x p_1(x)f(x) \\ &\leq \sum_x p_1(x) \max(f(x)) \\ &= \max(f(x)) \sum_x p_1(x) \\ &= \max(f(x)) \cdot 1 = \max(f(x)) \end{aligned} \tag{8}$$

Proof.

$$
\begin{aligned}
& D_{TV}(p_1(x,y) \parallel p_2(x,y)) \\
&= \frac{1}{2}\sum_{x,y} |p_1(x,y) - p_2(x,y)| && \text{by definition of TVD} \\
&= \frac{1}{2}\sum_{x,y} |p_1(x)p_1(y|x) - p_2(x)p_2(y|x)| && \text{by equation 4} \\
&= \frac{1}{2}\sum_{x,y} |p_1(x)p_1(y|x) - p_1(x)p_2(y|x) + (p_1(x) - p_2(x))p_2(y|x)| && \text{by Addition and Subtraction} \\
&\le \frac{1}{2}\sum_{x,y} [|p_1(x)(p_1(y|x) - p_2(y|x))| + |p_2(y|x)(p_1(x) - p_2(x))|] && \text{by equation 5} \\
&= \frac{1}{2}\sum_{x,y} p_1(x)\,|p_1(y|x) - p_2(y|x)| + \frac{1}{2}\sum_{x,y} p_2(y|x)\,|p_1(x) - p_2(x)| \\
&= \frac{1}{2}\sum_{x,y} p_1(x)\,|p_1(y|x) - p_2(y|x)| + \frac{1}{2}\sum_{x} |p_1(x) - p_2(x)| \sum_{y} p_2(y|x) && \text{by equation 6} \\
&= \frac{1}{2}\sum_{x,y} p_1(x)\,|p_1(y|x) - p_2(y|x)| + \frac{1}{2}\sum_{x} |p_1(x) - p_2(x)| && \text{by equation 7} \\
&= E_{x\sim p_1}[D_{TV}(p_1(y|x) \parallel p_2(y|x))] + D_{TV}(p_1(x) \parallel p_2(x)) && \text{by definition of Expectation} \\
&\le \max_x D_{TV}(p_1(y|x) \parallel p_2(y|x)) + D_{TV}(p_1(x) \parallel p_2(x)) && \text{by equation 8}
\end{aligned}
\tag{9}
$$

Lemma B.2 (Markov chain TVD bound, time-varying).
Suppose the expected KL-divergence between two transition distributions is bounded as $\max_t E_{s\sim p_1^t(s)} D_{KL}(p_1(s'|s) \parallel p_2(s'|s)) \le \delta$, and the initial state distributions are the same – $p_1^{t=0}(s) = p_2^{t=0}(s)$. Then the distance in the state marginal is bounded as:

$$D_{TV}(p_1^t(s) \parallel p_2^t(s)) \le t\delta$$

("Possible" Error on paper : they used "KL-divergence" instead of "TVD", but from the proof 14, we can tell that δ_t is defined as $\frac{1}{2}E_{s'\sim p_1(s_{t-1})}\sum_s |p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')| = \frac{1}{2}E_{s'\sim p_1(s_{t-1})} D_{TV}(p_1(s_t = s|s_{t-1} = s')||p_2(s_t = s|s_{t-1} = s'))$. and also δ is defined as $\max_t \delta_t$, hence I think using "TVD" here is more logical although there also exist the inequality between kl-divergence and TVD. And the other reason is that when we using these lemma in the Theorems below, they also define δ by using TVD.)

Used Properties.

Conditional probabilities:

$$
\begin{aligned}
p(s_t = s) &= \sum_{s'} p(s_t = s, s_{t-1} = s') \\
&= \sum_{s'} \frac{p(s_t = s, s_{t-1} = s')}{p(s_{t-1} = s')} p(s_{t-1} = s') \\
&= \sum_{s'} p(s_t = s|s_{t-1} = s')p(s_{t-1} = s')
\end{aligned}
\tag{10}
$$

Triangle inequality (with summation):

$$\begin{aligned}\left|\sum_i (a_i - b_i)\right| &= |(a_1 - b_1) + (a_2 - b_2) + \cdots + (a_n - b_n)| \\ &\leq |(a_1 - b_1)| + |(a_2 - b_2)| + \cdots + |(a_n - b_n)| = \sum_i |(a_i - b_i)|\end{aligned} \tag{11}$$

Changing Expectation and Summation:

$$\begin{aligned}E\left[\sum_{i=1}^{n} X_i\right] &= \int_\Omega \sum_{i=1}^{n} X_i(\omega) P(\omega) \quad \text{by definition of Expectation} \\ &= \sum_{i=1}^{n} \int_\Omega X_i(\omega) P(\omega) \quad \text{provided } X_i \text{ is positive} \\ &= \sum_{i=1}^{n} E\left[X_i\right]\end{aligned} \tag{12}$$

Proof.

$$\begin{aligned}&\left|p_1^t(s) - p_2^t(s)\right| \\ &= |p_1(s_t = s) - p_2(s_t = s)| && \text{Rewriting with other way} \\ &= \left|\sum_{s'} p_1(s_t = s|s_{t-1} = s') p_1(s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s') p_2(s_{t-1} = s')\right| && \text{by equation 10} \\ &\leq \sum_{s'} |p_1(s_t = s|s_{t-1} = s') p_1(s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s') p_2(s_{t-1} = s')| && \text{by equation 11} \\ &= \sum_{s'} |p_1(s_t = s|s_{t-1} = s') p_1(s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s') p_1(s_{t-1} = s') \\ &\qquad + p_2(s_t = s|s_{t-1} = s') p_1(s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s') p_2(s_{t-1} = s')| && \text{by Addition and Subtraction} \\ &\leq \sum_{s'} |p_1(s_{t-1} = s')[p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')]| \\ &\quad + \sum_{s'} |p_2(s_t = s|s_{t-1} = s')[p_1(s_{t-1} = s') - p_2(s_{t-1} = s')]| && \text{by equation 5} \\ &= \sum_{s'} p_1(s_{t-1} = s') \, |p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')| \\ &\quad + \sum_{s'} p_2(s_t = s|s_{t-1} = s') \, |p_1(s_{t-1} = s') - p_2(s_{t-1} = s')| && \text{since probability > 0} \\ &= E_{x \sim p_1(s_{t-1})}[|p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')|] \\ &\quad + \sum_{s'} p_2(s_t = s|s_{t-1} = s') \, |p_1(s_{t-1} = s') - p_2(s_{t-1} = s')| && \text{by definition of Expectation}\end{aligned} \tag{13}$$

$$
\begin{aligned}
\epsilon_t &\overset{def}{=} D_{TV}(p_1^t(s) \parallel p_2^t(s)) \\
&= \frac{1}{2}\sum_s |p_1(s_t = s) - p_2(s_t = s)| && \text{by definition of TVD} \\
&\leq \frac{1}{2}\sum_s E_{s' \sim p_1(s_{t-1})}[|p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')|] \\
&\quad + \frac{1}{2}\sum_s \sum_{s'} p_2(s_t = s|s_{t-1} = s')\,|p_1(s_{t-1} = s') - p_2(s_{t-1} = s')| && \text{by equation 13} \\
&\quad \text{(Error on paper : they used "=", but from equation 13, we can tell that it's "}\leq\text{")} \\
&= \frac{1}{2}E_{s' \sim p_1(s_{t-1})}\sum_s |p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')| \\
&\quad + \frac{1}{2}\sum_s \sum_{s'} p_2(s_t = s|s_{t-1} = s')\,|p_1(s_{t-1} = s') - p_2(s_{t-1} = s')| && \text{by equation 12} \\
&= \frac{1}{2}E_{s' \sim p_1(s_{t-1})}\sum_s |p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')| \\
&\quad + \frac{1}{2}\sum_{s'} \left[|p_1(s_{t-1} = s') - p_2(s_{t-1} = s')| \sum_s p_2(s_t = s|s_{t-1} = s')\right] && \text{by equation 6} \\
&= \frac{1}{2}E_{s' \sim p_1(s_{t-1})}\sum_s |p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')| \\
&\quad + \frac{1}{2}\sum_{s'} [|p_1(s_{t-1} = s') - p_2(s_{t-1} = s')| \cdot 1] && \text{by equation 7} \\
&= \frac{1}{2}E_{s' \sim p_1(s_{t-1})}\sum_s |p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')| \\
&\quad + \frac{1}{2}\sum_{s'} |p_1(s_{t-1} = s') - p_2(s_{t-1} = s')| \\
&\overset{def}{=} \delta_t + \epsilon_{t-1} \\
&\quad \text{(where we define } \delta_t = \frac{1}{2}E_{s' \sim p_1(s_{t-1})}\sum_s |p_1(s_t = s|s_{t-1} = s') - p_2(s_t = s|s_{t-1} = s')|\text{)} \\
&= \ldots && \text{by doing iteratively} \\
&= \sum_{i=1}^{t} \delta_i + \epsilon_0 \\
&\quad \text{(Assuming we are not modeling the initial state distribution, so } \epsilon = 0\text{)} \\
&= \sum_{i=1}^{t} \delta_i \\
&\quad \text{(Error on paper : they mis-wrote the range of i as "0" and the subscript of } \delta \text{ as "t")} \\
&\leq \sum_{i=1}^{t} \delta && \text{by definition of } \delta \\
&= t\delta
\end{aligned}
\tag{14}
$$

Lemma B.3 (Branched Returns bound).
Suppose the expected KL-divergence between two dynamics distributions is bounded as $\max_t E_{s\sim p_1^t(s)} D_{KL}(p_1(s',a|s) \parallel p_2(s',a|s)) \leq \epsilon_m$, and $\max_s D_{TV}(\pi_1(a|s) \parallel \pi_2(a|s)) \leq \epsilon_\pi$.
Then the returns are bounded as:

$$|\eta_1 - \eta_2| \leq \frac{2\,\mathrm{R}\,\gamma(\epsilon_\pi + \epsilon_m)}{(1-\gamma)^2} + \frac{2\,\mathrm{R}\,\epsilon_\pi}{1-\gamma}$$

("Possible" Error on paper : Same thing I've argued in "Lemma B.2", I think that the KL-divergence here may also be modified to TVD)
("Possible" Error on paper : In my proof below, I think that the term "ϵ_π" highlighted above is unnecessary)
(Different sign on paper : The sign "R" here means "r_{max}" below, paper didn't use the same sign nor define what "R is.)

Used Properties.

Usage of "Lemma B.2":

$$\begin{aligned}
&D_{TV}(p_1(s_t = s) \parallel p_2(s_t = s)) \\
&\leq t\delta && \text{by "Lemma B.2"} \\
&= t \max_t \frac{1}{2} E_{s'\sim p_1^{t-1}}[D_{TV}((p_1(s_t = s|s_{t-1} = s')) \parallel (p_2(s_t = s|s_{t-1} = s'))] \\
&= t \max_t \frac{1}{2} E_{s'\sim p_1^{t-1}}[\sum_s |((p_1(s_t = s|s_{t-1} = s')) - (p_2(s_t = s|s_{t-1} = s'))|] && \text{by definition of TVD} \\
&= t \max_t \frac{1}{2} E_{s'\sim p_1^{t-1}}[\sum_s |\sum_a ((p_1(s_t = s, a_t = a|s_{t-1} = s')) - (p_2(s_t = s, a_t = a|s_{t-1} = s'))|] && \text{by } \textstyle\sum \text{ all possible action} \\
&\leq t \max_t \frac{1}{2} E_{s'\sim p_1^{t-1}}[\sum_s \sum_a |((p_1(s_t = s, a_t = a|s_{t-1} = s')) - (p_2(s_t = s, a_t = a|s_{t-1} = s'))|] && \text{by equation 11} \\
&\leq t\epsilon_m && \text{by definition of } \epsilon_m
\end{aligned} \tag{15}$$

Infinitely series summation:

$$\begin{aligned}
S &\stackrel{def}{=} 1 + 2\gamma + 3\gamma^2 + 4\gamma^3 + \dots \\
S - 2\gamma \cdot S + \gamma^2 \cdot S &= (1-\gamma)^2 S \\
&= 1 - (n+2)\gamma^{n+1} + (n-1)\gamma n + 2 \\
&= 1 && \text{as n approaches } \infty \\
\sum_t t\gamma^t &= \gamma \cdot \sum_t t\gamma^{t-1} = \frac{\gamma}{(1-\gamma)^2}
\end{aligned} \tag{16}$$

Proof.

Here, η_1 denotes returns of π_1 under dynamics $p_1(s'|s,a)$, and η_2 denotes returns of π_2 under dynamics $p_2(s'|s,a)$.

$$
\begin{aligned}
&|\eta_1 - \eta_2| \\
&= |\sum_{s,a}[(\sum_t \gamma^t r(s,a))p_1(s,a)] \\
&\quad - \sum_{s,a}[(\sum_t \gamma^t r(s,a))p_2(s,a)]| && \text{by definition of Expectation} \\
&= |\sum_{s,a}[\sum_t \gamma^t r(s,a)p_1(s,a)] \\
&\quad - \sum_{s,a}[\sum_t \gamma^t r(s,a)p_2(s,a)]| && \text{by putting } p(s,a) \text{ into } \sum_t \\
&= |\sum_{s,a}[\sum_t \gamma^t r(s,a)p_1(s_t=s,a_t=a) \\
&\qquad\qquad - \gamma^t r(s,a)p_2(s_t=s,a_t=a)]| \\
&= |\sum_{s,a}\sum_t \gamma^t r(s,a)[p_1(s_t=s,a_t=a) - p_2(s_t=s,a_t=a)]| \\
&= |\sum_t \sum_{s,a} \gamma^t r(s,a)[p_1(s_t=s,a_t=a) - p_2(s_t=s,a_t=a)]| \\
&\leq \sum_t \sum_{s,a} \gamma^t r(s,a)|p_1(s_t=s,a_t=a) - p_2(s_t=s,a_t=a)| && \text{by equation 6} \\
&\leq r_{max} \sum_t \gamma^t \sum_{s,a} |p_1(s_t=s,a_t=a) - p_2(s_t=s,a_t=a)| \\
&= r_{max} \sum_t \gamma^t 2 D_{TV}(p_1(s_t=s,a_t=a) \parallel p_2(s_t=s,a_t=a)) && \text{by definition of TVD} \\
&\leq r_{max} \sum_t \gamma^t 2[max_s D_{TV}(p_1(a_t=a|s_t=s) \parallel p_2(a_t=a|s_t=s)) \\
&\qquad\qquad + D_{TV}(p_1(s_t=s) \parallel p_2(s_t=s))] && \text{by "Lemma B.1"} \\
&\leq r_{max} \sum_t \gamma^t 2[t\epsilon_m + \epsilon_\pi] && \text{by equation 15 and definition of } \epsilon_\pi
\end{aligned}
$$

("Possible" Error on paper : they took "$\epsilon_m + \epsilon_\pi$", instead of using "$\epsilon_\pi$" here, but I think that using "ϵ_π" is enough to be the upper bound)

$$
\leq 2r_{max}(\frac{\gamma\epsilon_m}{(1-\gamma)^2} + \frac{\epsilon_\pi}{1-\gamma}) \quad \text{by equation 16 and sum of geometric seq} \tag{17}
$$

Lemma B.4 (Returns bound, branched rollout).
Assume we run a branched rollout of length k. Before the branch ("pre" branch), we assume that the dynamics distributions are bounded as $\max_t E_{s\sim p_1^t(s)} D_{KL}(p_1^{pre}(s',a|s) \parallel p_2^{pre}(s',a|s)) \leq \epsilon_m^{pre}$ and after the branch as $\max_t E_{s\sim p_1^t(s)} D_{KL}(p_1^{post}(s',a|s) \parallel p_2^{post}(s',a|s)) \leq \epsilon_m^{post}$. Likewise, the policy divergence is bounded pre- and post- branch by ϵ_π^{pre} and ϵ_π^{post}, repsectively. Then the K-step returns are bounded as:

$$|\eta_1 - \eta_2| \leq 2r_{max}\left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \frac{k}{1-\gamma}(\epsilon_m^{post} + \epsilon_\pi^{post}) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi^{pre} + \frac{1}{1-\gamma}\epsilon_\pi^{post}\right]$$

("Possible" Error on paper : Same thing I've argued in "Lemma B.2", I think that the KL-divergence here may also be modified to TVD)
("Possible" Error on paper : In my proof below, I think that the term "ϵ_π" highlighted above is unnecessary)

Used Properties.

Usage of similar proof in "Lemma B.3"):

For $t \leq k$:

$$D_{TV}(d_1^t(s,a) \parallel d_2^t(s,a)) \leq t(\epsilon_m^{post} + \epsilon_\pi^{post}) + \epsilon_\pi^{post} \leq k(\epsilon_m^{post} + \epsilon_\pi^{post}) + \epsilon_\pi^{post}$$

("Possible" Error on paper : if my argument is right in "Lemma B.3",then we can also cancel the highlighted term above)
(More Personal Perspective: I think the reason why they enhance the upper bound here is to make it easier for computing the infinite sequence sum below)

(18)

and for $t \geq k$:

$$\begin{aligned} D_{TV}(d_1^t(s,a) \parallel d_2^t(s,a)) &\leq t(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \epsilon_\pi^{pre} + \epsilon_\pi^{post} \\ &= (t-k)(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + k(\epsilon_m^{post} + \epsilon_\pi^{post}) + \epsilon_\pi^{pre} + \epsilon_\pi^{post} \end{aligned}$$

("Possible" Error on paper : if my argument is right in "Lemma B.3",then we can also cancel the highlighted term above)

Proof.

$$D_{TV}(d_1(s,a) \parallel d_2(s,a))$$

$$\leq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t D_{TV}(d_1^t(s,a) \parallel d_2^t(s,a))$$

(by averaging the state marginal error overtime)

("Possible" Error on paper : I think the paper mis-wrote the term "t", since the inequality below doesn't exist "t")

$$\leq (1-\gamma)\sum_{t=0}^{k-1}\gamma^t(k(\epsilon_m^{post}+\epsilon_\pi^{post})+\epsilon_\pi^{post})$$

$$+(1-\gamma)\sum_{t=k}^{\infty}\gamma^t((t-k)(\epsilon_m^{pre}+\epsilon_\pi^{pre})+k(\epsilon_m^{post}+\epsilon_\pi^{post})+\epsilon_\pi^{pre}+\epsilon_\pi^{post}) \quad \text{by equation 18}$$

(Error on paper : the value on $\sum$ on paper is "k", but this will lead to "repeated" computing term "k")

$$= (1-\gamma)\sum_{t=0}^{k-1}\gamma^t k(\epsilon_m^{post}+\epsilon_\pi^{post})$$

$$+(1-\gamma)\sum_{t=0}^{k-1}\gamma^t\epsilon_\pi^{post}$$

$$+(1-\gamma)\sum_{t=k}^{\infty}\gamma^t(t-k)(\epsilon_m^{pre}+\epsilon_\pi^{pre})$$

$$+(1-\gamma)\sum_{t=k}^{\infty}\gamma^t k(\epsilon_m^{post}+\epsilon_\pi^{post})$$

$$+(1-\gamma)\sum_{t=k}^{\infty}\gamma^t(\epsilon_\pi^{pre}+\epsilon_\pi^{post})$$

$$= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t k(\epsilon_m^{post}+\epsilon_\pi^{post})+(1-\gamma)\sum_{t=0}^{k-1}\gamma^t\epsilon_\pi^{post}$$

$$+(1-\gamma)\sum_{t=k}^{\infty}\gamma^t(t-k)(\epsilon_m^{pre}+\epsilon_\pi^{pre})+(1-\gamma)\sum_{t=k}^{\infty}\gamma^t(\epsilon_\pi^{pre}+\epsilon_\pi^{post}) \quad \text{by combining 1-th and 4-th summation}$$

$$= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t k(\epsilon_m^{post}+\epsilon_\pi^{post})+(1-\gamma)\sum_{t=0}^{k-1}\gamma^t\epsilon_\pi^{post}$$

$$+(1-\gamma)\gamma^k\sum_{t=0}^{\infty}\gamma^t t(\epsilon_m^{pre}+\epsilon_\pi^{pre})+(1-\gamma)\gamma^k\sum_{t=0}^{\infty}\gamma^t(\epsilon_\pi^{pre}+\epsilon_\pi^{post}) \quad \text{by changing range in summation}$$

$$= k(\epsilon_m^{post}+\epsilon_\pi^{post})+(1-\gamma^k)\epsilon_\pi^{post}+\frac{\gamma^{k+1}}{1-\gamma}(\epsilon_m^{pre}+\epsilon_\pi^{pre})+\gamma^k(\epsilon_\pi^{pre}+\epsilon_\pi^{post})$$

$$= k(\epsilon_m^{post}+\epsilon_\pi^{post})+\epsilon_\pi^{post}+\frac{\gamma^{k+1}}{1-\gamma}(\epsilon_m^{pre}+\epsilon_\pi^{pre})+\gamma^k\epsilon_\pi^{pre}$$

(Error on paper : paper mis-wrote the highlighted term as $k(\epsilon_m^{post}+\epsilon_\pi^{post}+\epsilon_\pi^{post})$)

(19)

$$
\begin{aligned}
&|\eta_1 - \eta_2| \\
&\leq r_{max} \sum_t \gamma^t 2 D_{TV}(d_1(s_t = s, a_t = a) \parallel d_2(s_t = s, a_t = a)) \qquad \text{similar proof in "Lemma B.3"} \\
&= \frac{2 r_{max}}{1-\gamma} D_{TV}(d_1(s_t = s, a_t = a) \parallel d_2(s_t = s, a_t = a)) \\
&\leq \frac{2 r_{max}}{1-\gamma} [k(\epsilon_m^{post} + \epsilon_\pi^{post}) + \epsilon_\pi^{post} + \frac{\gamma^{k+1}}{1-\gamma}(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \gamma^k \epsilon_\pi^{pre}] \quad \text{by equation 19} \\
&= 2 r_{max} [\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \frac{k}{1-\gamma}(\epsilon_m^{post} + \epsilon_\pi^{post}) \\
&\quad + \frac{\gamma^k}{1-\gamma}\epsilon_\pi^{pre} + \frac{1}{1-\gamma}\epsilon_\pi^{post}]
\end{aligned}
\tag{20}
$$

Now, We are going to prove three theorems with the help of the lemmas we've proved above. And these theorems are mainly about the performance guarantees of MBPO. To say more, we are willing to find to exact value of C that we mentioned in equation 2

Theorem A.1 (MBPO performance bound).
Let the expected total variation between two transition distributions be bounded at each timestep by $\max_t E_{s\sim\pi_{D,t}}[D_{TV}(p(s'|s,a) \parallel \hat{p}(s'|s,a))] \leq \epsilon_m$, and the policy divergences are bounded as $\max_s[D_{TV}(\pi_D(a|s) \parallel \pi(a|s))] \leq \epsilon_\pi$. the returns are bounded as:

$$\eta[\pi] \geq \hat{\eta}[\pi] - \frac{2\gamma r \max(\epsilon_m + 2\epsilon_\pi)}{(1-\gamma)^2} - \frac{4r \max \epsilon_\pi}{(1-\gamma)}$$

("Possible" Error on paper : In my proof of "Lemma B.3", I think that the term "$2\epsilon_\pi$" highlighted above is unnecessary)

Proof.
Let π_D denote the data collecting policy, then by Addition and Subtraction:

$$\eta[\pi] - \hat{\eta}[\pi] = \underbrace{\eta[\pi] - \eta[\pi_D]}_{L_1} + \underbrace{\eta[\pi_D] - \hat{\eta}[\pi]}_{L_2} \tag{21}$$

Hence, if we can find the upper bound of L_1 and L_2, we're done:

$$L_1 \geq -\frac{2\gamma r \max \epsilon_\pi}{(1-\gamma)^2} - \frac{2r \max \epsilon_\pi}{1-\gamma}$$

(We can get the boundary above by applying "Lemma B.3", and since η are both calculate under the true model, hence for the first term, we don't have to consider the model error ϵ_m)
("Possible" Error on paper : Since we are applying "Lemma B.3" to get the lower bound of L_1, and I think that some term in "Lemma B.3" is unnecessary, hence I think that we don't need the whole first term highlighted above.)

$$L_2 \geq -\frac{2\gamma r \max(\epsilon_m + \epsilon_\pi)}{(1-\gamma)^2} - \frac{2r \max \epsilon_\pi}{1-\gamma}$$

(We can get the boundary above by applying "Lemma B.3", but for L_2, we have to consider the model error ϵ_m)
("Possible" Error on paper : Since we are applying "Lemma B.3" to get the lower bound of L_1, and I think that some term in "Lemma B.3" is unnecessary, hence I think that we don't need the whole first term highlighted above.)

$$\begin{aligned}
\eta[\pi] &= \hat{\eta}[\pi] + \underbrace{\eta[\pi] - \eta[\pi_D]}_{L_1} + \underbrace{\eta[\pi_D] - \hat{\eta}[\pi]}_{L_2} && \text{by equation 21} \\
&\geq \hat{\eta}[\pi] - \frac{2\gamma r \max \epsilon_\pi}{(1-\gamma)^2} - \frac{2r \max \epsilon_\pi}{1-\gamma} - \frac{2\gamma r \max(\epsilon_m + \epsilon_\pi)}{(1-\gamma)^2} - \frac{2r \max \epsilon_\pi}{1-\gamma} && \text{by inequality above} \\
&= \hat{\eta}[\pi] - \frac{2\gamma r \max(\epsilon_m + 2\epsilon_\pi)}{(1-\gamma)^2} - \frac{4r \max \epsilon_\pi}{(1-\gamma)}
\end{aligned} \tag{22}$$

Next, we describe bounds for branched rollouts. We define a branched rollout as a rollout which begins under some policy and dynamics (either true or learned), and at some point in time switches to rolling out under a new policy and dynamics for k steps. The point at which the branch is selected is weighted exponentially in time – that is, the probability of a branch point t being selected is proportional to γ^t.

We first present the simpler bound where the model error is bounded under the new policy, which we label as $\epsilon_{m'}$. This bound is difficult to apply in practice as supervised learning will typically control model error under the dataset collected by the previous policy.

Theorem A.2
Let the expected total variation between two the learned model is bounded at each timestep under the expectation of π by $\max_t E_{s\sim\pi_t}[D_{TV}(p(s'|s,a) \parallel \hat{p}(s'|s,a)] \leq \epsilon_{m'}$, and the policy divergences are bounded as $\max_s D_{TV}(\pi_D(a|s) \parallel \pi(a|s)) \leq \epsilon_\pi$. Then under a branched rollouts scheme with a branch length of k, the returns are bounded as:

$$\eta[\pi] \geq \eta^{branch}[\pi] - 2r_{max}\left[\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{\gamma^k\epsilon_\pi}{(1-\gamma)} + \frac{k}{1-\gamma}(\epsilon_{m'})\right]$$

("Possible" Error on paper : In my proof of "Lemma B.4", I think that the term "ϵ_π" highlighted above is unnecessary)

Proof.
As in the proof for "Theorem A.1", the proof for this theorem requires adding and subtracting the correct reference quantity and applying the corresponding returns bound ("Lemma B.4").

The choice of reference quantity is a branched rollout which executes the old policy π_D under the true dynamics until the branch point, then executes the new policy π under the true dynamics for k steps. We denote the returns under this scheme as $\eta^{\pi_D,\pi}$. We can split the returns as follows:

$$\eta[\pi] - \eta^{branch} = \underbrace{\eta[\pi] - \eta^{\pi_D,\pi}}_{L_1} + \underbrace{\eta^{\pi_D,\pi} - \eta^{branch}}_{L_2} \tag{23}$$

Hence, if we can find the upper bound of L_1 and L_2, we're done:

For L_1, L_1 accounts for the error from executing the old policy instead of the current policy. This term only suffers from error before the branch begins, and we can use "Lemma B.4" with $\epsilon_\pi^{pre} \leq \epsilon_\pi$ and all other errors set to 0:

$$\begin{aligned}
|\underbrace{\eta[\pi] - \eta^{\pi_D,\pi}}_{L_1}| &\leq 2r_{max}\left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \frac{k}{1-\gamma}(\epsilon_m^{post} + \epsilon_\pi^{post}) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi^{pre} + \frac{1}{1-\gamma}\epsilon_\pi^{post}\right] \\
&\leq 2r_{max}\left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(0 + \epsilon_\pi) + \frac{k}{1-\gamma}(0+0) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi + \frac{1}{1-\gamma}\cdot 0\right] \\
&= 2r_{max}\left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_\pi) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi\right]
\end{aligned} \tag{24}$$

("Possible" Error on paper : Since we are applying "Lemma B.4" to get the bound of L_1, and I think that some term in "Lemma B.4" is unnecessary, hence I think that we don't need the term highlighted above.)

For L_2, L_2 incorporates model error under the new policy incurred after the branch. Hence this term only suffers from model error after the branch, and we can use "Lemma B.4" with $\epsilon_m^{post} \leq \epsilon_{m'}$ and all other errors set to 0: (Error on paper : they mis-wrote the term of m' as m)

$$\begin{aligned}
|\underbrace{\eta^{\pi_D,\pi} - \eta^{branch}}_{L_2}| &\leq 2r_{max}\left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \frac{k}{1-\gamma}(\epsilon_m^{post} + \epsilon_\pi^{post}) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi^{pre} + \frac{1}{1-\gamma}\epsilon_\pi^{post}\right] \\
&\leq 2r_{max}\left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(0+0) + \frac{k}{1-\gamma}(\epsilon_{m'} + 0) + \frac{\gamma^k}{1-\gamma}\cdot 0 + \frac{1}{1-\gamma}\cdot 0\right] \\
&= 2r_{max}\left[\frac{k}{1-\gamma}\epsilon_{m'}\right]
\end{aligned} \tag{25}$$

$$
\begin{aligned}
\eta[\pi] &= \eta^{branch} + \underbrace{\eta[\pi] - \eta^{\pi_D,\pi}}_{L_1} + \underbrace{\eta^{\pi_D,\pi} - \eta^{branch}}_{L_2} && \text{by equation 23} \\
&\geq \eta^{branch} - 2r_{max}\left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_\pi) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi\right] - 2r_{max}\left[\frac{k}{1-\gamma}\epsilon_{m'}\right] && \text{by inequality above} \\
&= \eta^{branch}[\pi] - 2r_{max}\left[\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{\gamma^k\epsilon_\pi}{(1-\gamma)} + \frac{k}{1-\gamma}(\epsilon_{m'})\right]
\end{aligned}
\tag{26}
$$

The next bound is an analogue of "Theorem A.2" except using model errors under the previous policy π_D rather than the new policy π.

Theorem A.3
Let the expected total variation between two the learned model is bounded at each timestep under the expectation of π by $\max_t E_{s\sim\pi_{D,t}}[D_{TV}(p(s'|s,a) \parallel \hat{p}(s'|s,a))] \leq \epsilon_m$, and the policy divergences are bounded as $\max_s D_{TV}(\pi_D(a|s) \parallel \pi(a|s)) \leq \epsilon_\pi$. Then under a branched rollouts scheme with a branch length of k, the returns are bounded as:

$$\eta[\pi] \geq \eta^{branch}[\pi] - 2r \max \left[\frac{\gamma^{k+1} \epsilon_\pi}{(1-\gamma)^2} + \frac{\gamma^k + 2}{(1-\gamma)} \epsilon_\pi + \frac{k}{1-\gamma}(\epsilon_m + 2\epsilon_\pi) \right]$$

("Possible" Error on paper : In my proof of "Lemma B.4", I think that the term "ϵ_π" highlighted above is unnecessary)

Proof.
As in the proof for "Theorem A.2", the proof for this theorem requires adding and subtracting the correct reference quantity and applying the corresponding returns bound ("Lemma B.4"). The only modification is that we need to bound L_2 in terms of the model error under the π_D rather than π.

Once again, we design a new reference rollout. We use a rollout that executes the old policy π_D under the true dynamics until the branch point, then executes the *old* policy π_D under the model for k steps. We denote the returns under this scheme as $\eta^{\pi_D,\hat{\pi_D}}$. We can split L_2 we used in "Theorem A.2" as follows:

$$\underbrace{\eta^{\pi_D,\pi} - \eta^{branch}}_{L_2} = \underbrace{\eta^{\pi_D,\pi} - \eta^{\pi_D,\hat{\pi_D}}}_{L_3} + \underbrace{\eta^{\pi_D,\hat{\pi_D}} - \eta^{branch}}_{L_4} \tag{27}$$

Once again, we bound both terms L_3 and L_4 using "Lemma B.4".

For L_3, The rollouts in L_3 differ in both model and policy after the branch. This term suffers from error after the branch begins, and we can use "Lemma B.4" with $\epsilon_\pi^{post} \leq \epsilon_\pi$ and $\epsilon_m^{post} \leq \epsilon_m$ and all other errors set to 0:

$$\begin{aligned}
|\underbrace{\eta^{\pi_D,\pi} - \eta^{\pi_D,\hat{\pi_D}}}_{L_3}| &\leq 2r_{max} \left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \frac{k}{1-\gamma}(\epsilon_m^{post} + \epsilon_\pi^{post}) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi^{pre} + \frac{1}{1-\gamma}\epsilon_\pi^{post} \right] \\
&\leq 2r_{max} \left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(0+0) + \frac{k}{1-\gamma}(\epsilon_m + \epsilon_\pi) + \frac{\gamma^k}{1-\gamma} \cdot 0 + \frac{1}{1-\gamma}\epsilon_\pi \right] \\
&= 2r_{max} \left[\frac{k}{1-\gamma}(\epsilon_m + \epsilon_\pi) + \frac{1}{1-\gamma}\epsilon_\pi \right]
\end{aligned} \tag{28}$$

("Possible" Error on paper : Since we are applying "Lemma B.4" to get the bound of L_1, and I think that some term in "Lemma B.4" is unnecessary, hence I think that we don't need the term highlighted above.)

For L_4, The rollouts in L_4 differ only in the policy after the branch (as they both rollout under the model). This term suffers from error after the branch begins, and we can use "Lemma B.4" with $\epsilon_\pi^{post} \leq \epsilon_\pi$ and all other errors set to 0:

$$\begin{aligned}
|\underbrace{\eta^{\pi_D,\hat{\pi_D}} - \eta^{branch}}_{L_4}| &\leq 2r_{max} \left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \frac{k}{1-\gamma}(\epsilon_m^{post} + \epsilon_\pi^{post}) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi^{pre} + \frac{1}{1-\gamma}\epsilon_\pi^{post} \right] \\
&\leq 2r_{max} \left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(0+0) + \frac{k}{1-\gamma}(0 + \epsilon_\pi) + \frac{\gamma^k}{1-\gamma} \cdot 0 + \frac{1}{1-\gamma}\epsilon_\pi \right] \\
&= 2r_{max} \left[\frac{k}{1-\gamma}\epsilon_\pi + \frac{1}{1-\gamma}\epsilon_\pi \right]
\end{aligned} \tag{29}$$

("Possible" Error on paper : Since we are applying "Lemma B.4" to get the bound of L_1, and I think that some term in "Lemma B.4" is unnecessary, hence I think that we don't need the term highlighted above.)

Finally, we can combine the result in "Theorem A.3" and the bound we got above to get the desired result:

$$
\begin{aligned}
\eta[\pi] &= \eta^{branch} + \underbrace{\eta[\pi] - \eta^{\pi_D,\pi}}_{L_1} + \underbrace{\eta^{\pi_D,\pi} - \eta^{branch}}_{L_2} && \text{by equation 23} \\
&= \eta^{branch} + \underbrace{\eta[\pi] - \eta^{\pi_D,\pi}}_{L_1} + \underbrace{\eta^{\pi_D,\pi} - \eta^{\pi_D,\hat{\pi_D}}}_{L_3} + \underbrace{\eta^{\pi_D,\hat{\pi_D}} - \eta^{branch}}_{L_4} \\
&\geq \eta^{branch} - \underbrace{2r_{max}\left[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_\pi) + \frac{\gamma^k}{1-\gamma}\epsilon_\pi\right]}_{L_1} \\
&\quad - \underbrace{2r_{max}\left[\frac{k}{1-\gamma}(\epsilon_m + \epsilon_\pi) + \frac{1}{1-\gamma}\epsilon_\pi\right]}_{L_2} \\
&\quad - \underbrace{2r_{max}\left[\frac{k}{1-\gamma}\epsilon_\pi + \frac{1}{1-\gamma}\epsilon_\pi\right]}_{L3} && \text{by inequality above} \\
&= \eta^{branch}[\pi] - 2r\max\left[\frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{\gamma^k + 2}{(1-\gamma)}\epsilon_\pi + \frac{k}{1-\gamma}(\epsilon_m + 2\epsilon_\pi)\right]
\end{aligned}
\tag{30}
$$

4 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions

  In this paper, it has shown that it is possible to formulate model-based reinforcement learning algorithms with monotonic improvement guarantees. But whether we can found the policy satisfying the bound C comes into question. Although in this paper it has modified the pessimistic bound using linear approximation of model generalization, I think that finding a tighter bound will be the potential future research directions in Model-Based approaches.

- Any technical limitations

  I think that in this paper, there are only few number of technical assumptions that I've mentioned in section 2. And I think that the reason why there is almost "no" technical limitation is that the paper want to find the bound that useful for all Model-Based method and all kinds of state action cases.

- Any latest results on the problem of interest

  NAN

References

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

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.