A. BMPO Performance Guarantee

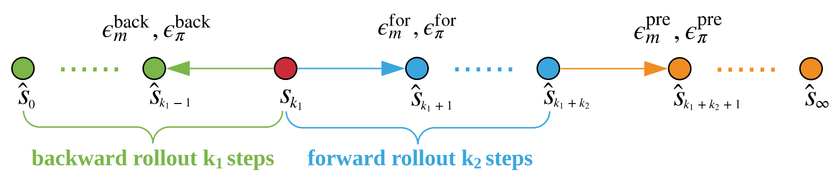

Figure 6. Bidirectional rollout.

Lemma A.1. *(Bidirectional Branched Rollout Returns Bound). Let η_1, η_2 be the expected returns of two bidirectional branched rollouts. Out of the branch, we assume that the expected total variation distance between these two dynamics at each timestep t is bounded as $\max_t E_{(s,a)\sim p_1^t(s,a)} D_{TV}\left(p_1^{\text{pre}}\left(s'|s,a\right)\|p_2^{\text{pre}}\left(s'|s,a\right)\right) \leq \epsilon_m^{\text{pre}}$, similarly, the forward branch dynamic bounded as $\max_t E_{(s,a)\sim p_1^t(s,a)} D_{TV}(p_1^{\text{for}}\left(s'|s,a\right)\|p_2^{\text{for}}\left(s'|s,a\right)) \leq \epsilon_m^{\text{for}}$, and the backward branch dynamic bounded as $\max_t E_{(s',a)\sim p_1^t(s',a)} D_{TV}(p_1^{\text{back}}\left(s|s',a\right)\|p_2^{\text{back}}\left(s|s',a\right)) \leq \epsilon_m^{\text{back}}$. Likewise, the total variation distance of policy is bounded by $\epsilon_\pi^{\text{pre}}$, $\epsilon_\pi^{\text{for}}$ and $\epsilon_\pi^{\text{back}}$, respectively (as Figure 6 shows). Then the returns are bounded as*

$$
\begin{aligned}
|\eta_1 - \eta_2| \leq 2r_{\max}\Bigg[\frac{\gamma^{k_1+k_2+1}}{(1-\gamma)^2}\left(\epsilon_m^{\text{pre}}+\epsilon_\pi^{\text{pre}}\right) + \frac{\gamma^{k_1+k_2}}{1-\gamma}\epsilon_\pi^{\text{pre}} + \frac{1-\gamma^{k_1}}{1-\gamma}\left(k_1\left(\epsilon_m^{\text{back}}+\epsilon_\pi^{\text{back}}\right)+\epsilon_\pi^{\text{back}}\right) \\
+\frac{\gamma^{k_1}}{1-\gamma}\left(k_2\left(\epsilon_m^{\text{for}}+\epsilon_\pi^{\text{for}}\right)+\epsilon_\pi^{\text{for}}\right)\Bigg].
\end{aligned} \tag{11}
$$

Proof. Lemma B.1 and Lemma B.2 imply that state marginal error at each timestep can be bounded by the divergence at the current timestep plus the state marginal error at the next (Lemma B.1), or previous (Lemma B.2) timestep. And by employing Lemma B.3, we can convert the (s,a) joint distribution to marginal distributions. Thus, letting $d_1(s,a)$ and $d_2(s,a)$ denote the state-action marginals, we can write:
For $t \leq k_1$:

$$
\begin{aligned}
D_{TV}\left(d_1^t(s,a)\|d_2^t(s,a)\right) &\leq D_{TV}\left(d_1^t(s)\|d_2^t(s)\right) + \max_{s'} D_{TV}\left(\pi_1(a|s')\|\pi_2(a|s')\right) \\
&\leq (k_1 - t)\left(\epsilon_m^{\text{back}}+\epsilon_\pi^{\text{back}}\right)+\epsilon_\pi^{\text{back}} \leq k_1\left(\epsilon_m^{\text{back}}+\epsilon_\pi^{\text{back}}\right)+\epsilon_\pi^{\text{back}}
\end{aligned} \tag{12}
$$

Similarly, for $k_1 < t \leq k_1 + k_2$:

$$
D_{TV}\left(d_1^t(s,a)\|d_2^t(s,a)\right) \leq (t-k_1)\left(\epsilon_m^{\text{for}}+\epsilon_\pi^{\text{for}}\right)+\epsilon_\pi^{\text{for}} \leq k_2\left(\epsilon_m^{\text{for}}+\epsilon_\pi^{\text{for}}\right)+\epsilon_\pi^{\text{for}} \tag{13}
$$

And for $t > k_1 + k_2$:

$$
D_{TV}\left(d_1^t(s,a)\|d_2^t(s,a)\right) \leq (t-k_1-k_2)\left(\epsilon_m^{\text{pre}}+\epsilon_\pi^{\text{pre}}\right)+k_2\left(\epsilon_m^{\text{for}}+\epsilon_\pi^{\text{for}}\right)+\epsilon_\pi^{\text{pre}}+\epsilon_\pi^{\text{for}} \tag{14}
$$

We can now bound the difference in occupancy measures by averaging the state marginal error over time, weighted by the discount:

$$
\begin{aligned}
D_{TV}\left(d_1(s,a)\|d_2(s,a)\right) \leq & (1-\gamma)\sum_{t=0}^{\infty}\gamma^t D_{TV}\left(d_1^t(s,a)\|d_2^t(s,a)\right) \\
\leq & (1-\gamma)\sum_{t=0}^{k_1}\gamma^t\left(k_1\left(\epsilon_m^{\text{back}}+\epsilon_\pi^{\text{back}}\right)+\epsilon_\pi^{\text{back}}\right) \\
& +(1-\gamma)\sum_{t=k_1}^{k_1+k_2}\gamma^t\left(k_2\left(\epsilon_m^{\text{for}}+\epsilon_\pi^{\text{for}}\right)+\epsilon_\pi^{\text{for}}\right) \\
& +(1-\gamma)\sum_{t=k_1+k_2}^{\infty}\gamma^t\left((t-k_1-k_2)\left(\epsilon_m^{\text{pre}}+\epsilon_\pi^{\text{pre}}\right)+k_2\left(\epsilon_m^{\text{for}}+\epsilon_\pi^{\text{for}}\right)+\epsilon_\pi^{\text{pre}}+\epsilon_\pi^{\text{for}}\right) \\
= & \left(k_1\left(\epsilon_m^{\text{back}}+\epsilon_\pi^{\text{back}}\right)+\epsilon_\pi^{\text{back}}\right)\left(1-\gamma^{k_1}\right)+\left(k_2\left(\epsilon_m^{\text{for}}+\epsilon_\pi^{\text{for}}\right)+\epsilon_\pi^{\text{for}}\right)\left(\gamma^{k_1}\right) \\
& +\frac{\gamma^{k_1+k_2+1}}{1-\gamma}\left(\epsilon_m^{\text{pre}}+\epsilon_\pi^{\text{pre}}\right)+\gamma^{k_1+k_2}\epsilon_\pi^{\text{pre}}
\end{aligned}
$$

Multiplying this bound by $\frac{2r_{\max}}{1-\gamma}$ to convert the occupancy measure difference into a returns bound completes the proof. □

Theorem A.1. *(BMPO Return Discrepancy Upper Bound) Assume that the expected total variation distance between the learned forward model $\hat{p}$ and the true dynamics p at each timestep t is bounded as $\max_t E_{(s,a)\sim\pi_t}\left[D_{TV}\left(p\left(s'|s,a\right)\|\hat{p}\left(s'|s,a\right)\right)\right] \leq \epsilon_m^{\text{for}}$. Similarly, the error of backward model $\hat{q}$ is bounded as $\max_t E_{(s',a)\sim\pi_t}\left[D_{TV}\left(q\left(s|s',a\right)\|\hat{q}\left(s|s',a\right)\right)\right] \leq \epsilon_m^{\text{back}}$ and the variation between current policy and the behavioral policy is bounded as $\max_s D_{TV}\left(\pi_D(a|s)\|\pi(a|s)\right) \leq \epsilon_\pi$. Assume $\epsilon_m^{\text{for}} \approx \epsilon_m^{\text{back}} = \epsilon_m$ and $\epsilon_\pi^{\text{back}} = 0$, then under a branched rollouts scheme with a backward branch length of k_1 and a forward branch length of k_2, the returns are bounded as:*

$$
\left|\eta[\pi]-\eta^{\text{branch}}[\pi]\right| \leq 2r_{\max}\left[\frac{\gamma^{k_1+k_2+1}\epsilon_\pi}{(1-\gamma)^2}+\frac{\gamma^{k_1+k_2}\epsilon_\pi}{(1-\gamma)}+\frac{max(k_1,k_2)}{1-\gamma}\left(\epsilon_m\right)\right]. \tag{15}
$$

Proof. Using Lemma A.1, out of the branch, we only suffer from error of executing old policy π_D, so, set $\epsilon_\pi^{\text{pre}} = \epsilon_\pi$ and $\epsilon_m^{\text{pre}} = 0$. Then in the branched rollout, we execute current policy, so the only error comes from using the learned model to simulate. Set $\epsilon_\pi^{\text{for}} = \epsilon_\pi^{\text{back}} = 0$ and $\epsilon_m^{\text{for}} = \epsilon_m^{\text{back}} = \epsilon_m$. Plugging these in Lemma B.1 we can get:

$$
\begin{aligned}
\left|\eta[\pi]-\eta^{\text{branch}}[\pi]\right| &\leq 2r_{\max}\left[\frac{\gamma^{k_1+k_2+1}\epsilon_\pi}{(1-\gamma)^2}+\frac{\gamma^{k_1+k_2}\epsilon_\pi}{(1-\gamma)}+\frac{k_1(1-\gamma^{k_1})+k_2(\gamma^{k_1})}{1-\gamma}\left(\epsilon_m\right)\right] \\
&\leq 2r_{\max}\left[\frac{\gamma^{k_1+k_2+1}\epsilon_\pi}{(1-\gamma)^2}+\frac{\gamma^{k_1+k_2}\epsilon_\pi}{(1-\gamma)}+\frac{max(k_1,k_2)(1-\gamma^{k_1}+\gamma^{k_1}))}{1-\gamma}\left(\epsilon_m\right)\right] \\
&\leq 2r_{\max}\left[\frac{\gamma^{k_1+k_2+1}\epsilon_\pi}{(1-\gamma)^2}+\frac{\gamma^{k_1+k_2}\epsilon_\pi}{(1-\gamma)}+\frac{max(k_1,k_2)}{1-\gamma}\left(\epsilon_m\right)\right]
\end{aligned} \tag{16}
$$

□

B. Useful Lemmas

In this section, we give proofs of the lemmas used before.

Lemma B.1. *(Backward State Marginal Distance Bound). Suppose the expected total variation distance between two backward dynamics is bounded as $\max_t E_{(s',a)\sim p_1^t}\left[D_{TV}\left(p_1\left(s|s',a\right)\|p_2\left(s|s',a\right)\right)\right] \leq \epsilon_m^{\text{back}}$ and the backward policy divergences are bounded as $\max_{s'} D_{TV}\left(\pi_1(a|s')\|\pi_2(a|s')\right) \leq \epsilon_\pi^{\text{back}}$. Then the state marginal distance at timestep t can be bounded as:*

$$
D_{TV}\left(p_1^t(s)\|p_2^t(s)\right) \leq \epsilon_m^{\text{back}}+\epsilon_\pi^{\text{back}}+D_{TV}\left(p_1^{t+1}(s)\|p_2^{t+1}(s)\right). \tag{17}
$$

Proof. Let the total variation distance of state at time t be denoted as $\epsilon_t = D_{TV}\left(p_1^t(s)\|p_2^t(s)\right)$.

$$
\begin{aligned}
\left|p_1^t(s)-p_2^t(s)\right| =& \Big|\sum_{s',a} p_1\left(s_t=s|s',a\right)p_1^{t+1}\left(s',a\right)-p_2\left(s_t=s|s',a\right)p_2^{t+1}\left(s',a\right)\Big| \\
\le& \sum_{s',a}\left|p_1\left(s_t=s|s',a\right)p_1^{t+1}\left(s',a\right)-p_2\left(s_t=s|s',a\right)p_2^{t+1}\left(s',a\right)\right| \\
=& \sum_{s',a}\left|p_1\left(s_t=s|s',a\right)p_1^{t+1}\left(s',a\right)-p_2\left(s_t=s|s',a\right)p_1^{t+1}\left(s',a\right)\right. \\
& \left.+p_2\left(s_t=s|s',a\right)p_1^{t+1}\left(s',a\right)-p_2\left(s_t=s|s',a\right)p_2^{t+1}\left(s',a\right)\right| \\
\le& \sum_{s',a} p_1^{t+1}\left(s',a\right)\left|p_1\left(s|s',a\right)-p_2\left(s|s',a\right)\right|+p_2\left(s|s',a\right)\left|p_1^{t+1}\left(s',a\right)-p_2^{t+1}\left(s',a\right)\right| \\
=& E_{s',a\sim p_1^{t+1}}\left[\left|p_1\left(s|s',a\right)-p_2\left(s|s',a\right)\right|\right]+\sum_{s',a} p_2\left(s|s',a\right)\left|p_1^{t+1}\left(s',a\right)-p_2^{t+1}\left(s',a\right)\right|
\end{aligned}
$$

$$
\begin{aligned}
\epsilon_t = D_{TV}\left(p_1^t(s)\|p_2^t(s)\right) =& \frac{1}{2}\sum_s\left|p_1^t(s)-p_2^t(s)\right| \\
\le& \frac{1}{2}\sum_s\Big(E_{s',a\sim p_1^{t+1}}\left[\left|p_1\left(s|s',a\right)-p_2\left(s|s',a\right)\right|\right]+\sum_{s',a} p_2\left(s|s',a\right)\left|p_1^{t+1}\left(s',a\right)-p_2^{t+1}\left(s',a\right)\right|\Big) \\
=& E_{s',a\sim p_1^{t+1}}\left[D_{TV}\left(p_1\left(s|s',a\right)\|p_2\left(s|s',a\right)\right)\right]+D_{TV}\left(p_1^{t+1}\left(s',a\right)\|p_2^{t+1}\left(s',a\right)\right) \\
\le& \epsilon_m^{\text{back}}+D_{TV}\left(p_1^{t+1}\left(s'\right)\|p_2^{t+1}\left(s'\right)\right)+\max_{s'} D_{TV}\left(p_1(a|s')\|p_2(a|s')\right) \\
=& \epsilon_m^{\text{back}}+\epsilon_\pi^{\text{back}}+D_{TV}\left(p_1^{t+1}(s)\|p_2^{t+1}(s)\right)
\end{aligned}
$$

□

Lemma B.2. *(Forward State Marginal Distance Bound) ((Janner et al., 2019), Lemma B.2, B.3). Suppose the expected TVD between two forward dynamics is bounded as* $\max_t E_{(s,a)\sim p_1^t}\left[D_{TV}\left(p_1\left(s'|s,a\right)\|p_2\left(s'|s,a\right)\right)\right] \le \epsilon_m^{\text{for}}$ *and the forward policy divergences are bounded as* $\max_{s'} D_{TV}\left(\pi_1(a|s)\|\pi_2(a|s)\right) \le \epsilon_\pi^{\text{for}}$*. Then the state marginal distance at timestep t can be bounded as:*

$$
D_{TV}\left(p_1^t(s)\|p_2^t(s)\right) \le \epsilon_m^{\text{for}}+\epsilon_\pi^{\text{for}}+D_{TV}\left(p_1^{t-1}(s)\|p_2^{t-1}(s)\right). \tag{18}
$$

Lemma B.3. *(TVD Of Joint Distributions) ((Janner et al., 2019), Lemma B.1). 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 distributions as:*

$$
D_{TV}\left(p_1(x,y)\|p_2(x,y)\right) \le D_{TV}\left(p_1(x)\|p_2(x)\right)+\max_x D_{TV}\left(p_1(y|x)\|p_2(y|x)\right). \tag{19}
$$

C. Environment Settings

In this section, we provide a comparison of the environment settings used in our experiments. Among them, 'Hopper-NT' and 'Walker2d-NT' refer to the settings in Langlois et al. (2019) and others are the standard version.

Table 1. Observation and action dimension, and task horizon of the environments used in our experiments.

Environment Name	Observation Space Dimension	Action Space Dimension	Steps Per Epoch
Pendulum	3	1	200
Hopper	11	3	1000
Hopper-NT	11	3	1000
Walker2d	17	6	1000
Walker2d-NT	17	6	1000
Ant	27	8	1000

Table 2. Reward function and termination states condition of the environments used in our experiments. θ_t denotes the joint angle, x_t denotes the position in x direction, a_t denotes the action control input, and z_t denotes the height.

Environment Name	Reward Function	Termination States Condition
Pendulum	$-\theta_t^2 - 0.1\dot{\theta}_t^2 - 0.001 \left\| a_t \right\|_2^2$	None
Hopper	$\dot{x}_t - 0.001 \left\| a_t \right\|_2^2 + 1$	$z_t \leq 0.7$ or $\theta_t \geq 0.2$
Hopper-NT	$\dot{x}_t - 0.1 \left\| a_t \right\|_2^2 - 3.0 \times (z_t - 1.3)^2 + 1$	None
Walker2d	$\dot{x}_t - 0.001 \left\| a_t \right\|_2^2 + 1$	$z_t \leq 0.8$ or $z_t \geq 2.0$ or $\lvert\theta_t\rvert \geq 1.0$
Walker2d-NT	$\dot{x}_t - 0.1 \left\| a_t \right\|_2^2 - 3.0 \times (z_t - 1.3)^2 + 1$	None
Ant	$\dot{x}_t - 0.5 \left\| a_t \right\|_2^2 + 1$	$z_t \leq 0.2$ or $z_t \geq 1.0$

D. Hyperparameters

Table 3. Hyperparameter settings for BMPO. $x \to y$ over epochs $a \to b$ means clipped linear function, *i.e.* for epoch e, $f(e) = clip((x + \frac{e-a}{b-a} \cdot (x-y)), x, y)$. Other hyperparameters not listed here are the same as those in MBPO (Janner et al., 2019).

Environment Name	k_1	k_2	β	MPC Horizon	Epochs
Pendulum	$1 \to 5$ over epochs $1 \to 5$	$1 \to 5$ over epochs $1 \to 5$	$0.01 \to 0$ over epochs $0 \to 10$	6	20
Hopper	$1 \to 15$ over epochs $20 \to 150$	$1 \to 15$ over epochs $20 \to 150$	$0.004 \to 0.003$ over epochs $20 \to 30$	6	100
Hopper-NT	$1 \to 15$ over epochs $20 \to 150$	$1 \to 15$ over epochs $20 \to 150$	0.01	6	100
Walker2d	1	1	$0.01 \to 0$ over epochs $0 \to 100$	1	200
Walker2d-NT	1	1	0.01	0	200
Ant	1	$1 \to 25$ over epochs $20 \to 100$	0.003	0	300

E. Computing Infrastructure

In this section, we provide a description of the computing infrastructure used to run all the experiments in Table 4. We also show the computation time comparison between our algorithm and the MBPO baseline in Table 5.

Table 4. Computing infrastructure.

CPU	GPU	Memory
AMD2990WX	RTX2080TI$\times$4	256GB

Table 5. Computation time in hours for one experiment.

	Pendulum	Hopper	Hopper-NT	Walker2d	Walker2d-NT	Ant
BMPO	0.49	16.34	17.98	27.24	27.34	71.51
MBPO	0.41	10.33	11.12	22.26	21.32	57.42