Combining Parametric and Nonparametric Models for Off-Policy Evaluation — Appendix

A. Model Selection Algorithms

Algorithm 2 Greedy model selection

```
function GreedyMoeModelSelection($(s_t, a_t)$)
    $\hat{\varepsilon}_{\mathrm{t,np}} \leftarrow$ Eq. 9
    $\hat{\varepsilon}_{\mathrm{t,p}} \leftarrow$ Eq. 13
    if $\hat{\varepsilon}_{\mathrm{np}}(x_t^{(n)}, a_t^{(n)}) < \hat{\varepsilon}_{\mathrm{p}}(x_t^{(n)}, a_t^{(n)})$ then
        // Return nonparametric model
        Return $(\hat{f}_{t,np}, \hat{f}_{r,np})$
    else
        // Return parametric model
        Return $(\hat{f}_{t,p}, \hat{f}_{r,p})$
    end if
end function
```

In this section we provide the two algorithms used to choose the model in the MoE simulator. The functions *GreedyMoeModelSelection* in Algorithm 2 and *MctsMoeModelSelection* in Algorithm 3 can be substituted with *ChooseModel* in Algroithm 1 in the main text.

Algorithm 2 is straight forward and simply returns the model with the smaller immediate estimated transition error. This algorithm could also use a weighted sum of both the transition and reward error, but that choice would require choosing a tuning parameter which controls the relative importance of the transitions and rewards accuracy.

Algorithm 3 is based on the standard upper confidence bound for trees (UCT) algorithm (Coulom, 2006; Browne et al., 2012). We note once again that the domain over which the MCTS algorithm plans is not the same domain as the RL environment. The states for the MCTS algorithm are state-action pair in the RL domain, and the actions are choosing either the parametric or nonparametric model.

The value of a rollout for the planner is minus the return error bound derived in Theorem 1 in the main text, $-\delta_g$. Because of the compounding effect of the state error bound, $\delta(t)$, the value of δ for each node must be rolled forward for all nodes which results in the main modifications to the standard UCT algorithm in, mainly in functions *Expand* and *DefaultPolicy*.

A tuning parameter of the UCT algorithm is the exploration constant, c_e, which controls how frequently the algorithm should explore branches which appear not promising if they have not been explored enough. When the rewards are bounded between 0 and 1, a standard choice for c_e is $1/\sqrt{2}$. Because we don't know a priori how large the errors might be, we continuously update the exploration parameter such that $c_e = \max \hat{\varepsilon}_t / \sqrt{2}$.

B. Proof of Lemma 1

We first restate Lemma 1.

Lemma 1 Let $\varepsilon_t(t)$ be the transition estimation error bound for the chosen model at time-step t,

$$\varepsilon_t(t) \geq \Delta(\hat{x}_{t+1}, f_t(\hat{x}_t, a_t)) \tag{19}$$

The state error at time-step t is:

$$\delta(t) := \Delta(x_t, \hat{x}_t) \leq \sum_{t'=0}^{t-1} (L_t)^{t'} \varepsilon_t(t - t' - 1) \tag{20}$$

where L_t is the Lipschitz constant of the transition function, f_t.

Proof. We prove Lemma 1 by induction. The state prediction error at time t is bounded by:

$$\begin{aligned} \delta(t) &:= \Delta(x_t, \hat{x}_t) \\ &\leq \Delta(x_t, f_t(\hat{x}_{t-1}, a_t)) + \Delta(f_t(\hat{x}_{t-1}, a_{t-1}), \hat{x}_t) \\ &\leq L_t \delta(t-1) + \varepsilon_t(t-1), \end{aligned} \tag{21}$$

Where the first inequality is a consequence of the triangle inequality. By definition, $\delta(1) \leq \varepsilon(0)$. Therefore

$$\begin{aligned} \delta(t) &:= \Delta(x_t, \hat{x}_t) \\ &\leq L_t \delta(t-1) + \varepsilon_t(t-1) \\ &\leq L_t(L_t \delta(t-2) + \varepsilon_t(t-2)) + \varepsilon(t-1) \\ &\dots \\ &\leq \sum_{t'=0}^{t-1} (L_t)^{t'} \varepsilon_t(t - t' - 1), \end{aligned} \tag{22}$$

completing the proof.

Algorithm 3 MCTS-MoE model selection

```
function MctsMoeModelSelection($(s_t, a_t)$)
    create root node $\nu_t$ with state $(s_t, a_t)$
    $\delta(\nu_t) \leftarrow 0$ // State error bound for node
    $\delta_g(\nu_t) \leftarrow 0$ // Return error bound for node
    $\tau(\nu) \leftarrow 0$ // Time-steps from root node
    while within computational budget do
        $\nu_l \leftarrow$ TreePoicy$(\nu_t)$
        $V \leftarrow$ DefaultPolicy$(\nu_l)$
        Backup$(\nu_l, V)$
    end while
    Return Model( $\underset{\nu' \in \text{children of } \nu_t}{\arg\max} \widetilde{Q}(\nu')$)
end function

function TreePolicy$(\nu)$
    while $\nu$ is not terminal do
        if $\nu$ not fully expanded then
            Return Expand$(\nu)$
        else
            $\nu \leftarrow \underset{\nu' \in \text{children of } \nu}{\arg\max} \frac{Q(\nu)}{N(\nu)} + c_e \sqrt{\frac{2 \ln N(\nu)}{N(\nu')}}$
        end if
        Return $\nu$
    end while
end function

function Expand$(\nu)$
    add a new child $\nu'$ to $\nu$
    if $\nu$ has no children then
        Model$(\nu') \leftarrow$ GreedyMoeModelSelection$(s(\nu), a(\nu))$
    else
        Model$(\nu') \leftarrow$ model not yet tried in $\nu$
    end if
    $(s(\nu'), a(\nu')) \leftarrow (\hat{f}_{t,\text{Model}(\nu')}, \pi_e(\hat{f}_{t,\text{Model}(\nu')}))$
    $\varepsilon_t(\nu'), \varepsilon_r(\nu') \leftarrow$ ComputeErrors(Model$(\nu')$)
    $N(\nu') \leftarrow 0$ // Times node was visited
    $Q(\nu') \leftarrow 0$ // Total reward of all rollouts through node
    $\widetilde{Q}(\nu') \leftarrow 0$ // Rollout with highest reward for node
    $\tau(\nu') \leftarrow \tau(\nu) + 1$
    $\delta(\nu') \leftarrow L_t \cdot \delta(\nu) + \varepsilon_t(\nu')$
    $\delta_g(\nu') \leftarrow \delta_g(\nu) + \gamma^{\tau(\nu')} (\varepsilon_r(\nu') + L_t \cdot \delta(\nu'))$
    Return $\nu'$
end function

function DefaultPolicy$(\nu)$
    $(s^*, a^*) \leftarrow (s(\nu), a(\nu))$
    $\tau^* \leftarrow \tau(\nu)$
    $\delta^* \leftarrow \delta(\nu)$
    $\delta_g^* \leftarrow \delta_g(\nu)$
    while s in not terminal do
        Model $\leftarrow$ GreedyMoeModelSelection$(s, a)$
        $s \leftarrow \hat{f}_{t,Model}(s, a)$
        $\varepsilon_t^*, \varepsilon_r^* \leftarrow$ ComputeErrors(Model)
        $a \leftarrow \pi_e(s)$
        $\tau^* \leftarrow \tau^* + 1$
        $\delta^* \leftarrow L_t \cdot \delta^* + \varepsilon_t^*$
        $\delta_g^* \leftarrow \delta_g^* + \gamma^{\tau^*} (\varepsilon_r^* + L_t \cdot \delta^*)$
    end while
    Return $-\delta_g^*$
end function

function Backup$(\nu, V)$
    while $\nu$ is not null do
        $N(\nu) \leftarrow N(\nu) + 1$
        $Q(\nu) \leftarrow Q(\nu) + V$
        $\widetilde{Q}(\nu) \leftarrow \max(\widetilde{Q}(\nu), V)$
        $\nu \leftarrow$ parent of $\nu$
    end while
end function

function ComputeErrors(Model)
    if Model = parametric then
        $\varepsilon_t \leftarrow$ Eq. 13
        $\varepsilon_r \leftarrow$ Eq. 14
    else
        // Model = nonparametric
        $\varepsilon_t \leftarrow$ Eq. 9
        $\varepsilon_r \leftarrow$ Eq. 10
    end if
    Return $\varepsilon_t, \varepsilon_r$
end function
```

C. Proof of Consistency

In this section we are going to prove MoE simulator (Algorithm 1) with MCTS model selection is a consistent estimator i.e. the return error goes to zero when the number of samples collected from behavior policy goes to infinity. We assume the planning error of MCTS is bounded by $\epsilon_{\text{planning}}$ where the objective of planning is to maximize:

$$-L_r \sum_{t=0}^{T-t_0} \gamma^t \sum_{t'=0}^{t-1} (L_t)^{t'} \hat{\varepsilon}_t(t_0+t-t'-1) - \sum_{t=0}^{T-t_0} \gamma^t \hat{\varepsilon}_r(t_0+t) \tag{23}$$

for any input state action pair (s_{t_0}, a_{t_0}).

Assumption 1. *(Coverage of behavior policy) For a data set $\mathcal{D}$ with n samples collected from behavior policy and any given state x and action a, let rad_n be $\min_{x_t^{(i)} \in \mathcal{D}, a_t^{(i)}=a} \Delta(x, x_t^{(i)})$. Then $\lim_{n\to\infty} rad_n = 0$.*

Assumption 2. *(Coverage of radius C) There exist an N such that for any $n > N$, for any n sample collected from behavior policy and any state x and action a, the chosen radius C satisfy that there is at least one sample in data set is within distance C of x and matches the action a.*

Assumption 3. *(Lipschitz continuity of parametric model) Functions $\hat{f}_t$ and $\hat{f}_r$ in parametric model class are L-Lipschitz with $L < \infty$.*

Lemma 2. *Under assumptions 1 and 3, Let n be the number of samples collected from behavior policy. For any x:*

$$\lim_{n\to\infty} \varepsilon_{t,np}(x) = 0, \quad \lim_{n\to\infty} \varepsilon_{r,np}(x) = 0$$
$$\lim_{n\to\infty} \hat{\varepsilon}_{t,np}(x) = 0, \quad \lim_{n\to\infty} \hat{\varepsilon}_{r,np}(x) = 0$$

Proof. Let $x_t^{(i)}$ be the state closest to x whose action $a_t^{(i)}$ equals a.

$$\varepsilon_{t,np}(x) = \Delta(f_t(x,a), f_t(x_t^{(i)}, a)) \tag{24}$$
$$\leq L_t \Delta(x, x_t^{(i)}) \leq L_t \text{rad}_n \tag{25}$$
$$\varepsilon_{r,np}(x) = \Delta(f_r(x,a), f_r(x_t^{(i)}, a)) \tag{26}$$
$$\leq L_r \Delta(x, x_t^{(i)}) \leq L_r \text{rad}_n \tag{27}$$

Thus $0 \leq \lim_{n\to\infty} \varepsilon_{t,np}(x) \leq L_t \lim_{n\to\infty} \text{rad}_n = 0$. So $\lim_{n\to\infty} \varepsilon_{t,np}(x) = 0$, similarly $\lim_{n\to\infty} \varepsilon_{r,np}(x) = 0$. For the estimated error:

$$\hat{\varepsilon}_{t,np}(x) = \hat{L}_t \Delta(x, x_t^{(i)}) \leq \hat{L}_t \text{rad}_n \tag{28}$$
$$= \max_{i \neq j} \frac{\Delta(x_{t'+1}^{(i)}, x_{t''+1}^{(j)})}{\Delta(x_{t'}^{(i)}, x_{t''}^{(j)})} \text{rad}_n \tag{29}$$
$$\leq L_t \text{rad}_n \tag{30}$$

Similarly, we have $\lim_{n\to\infty} \hat{\varepsilon}_{t,np}(x) = 0$ and $\lim_{n\to\infty} \hat{\varepsilon}_{r,np}(x) = 0$ □

A direct conclusion following from this claim and Theorem 1 is that the non-parametric model is a consistent estimator.

Lemma 3. *Let $L_{\hat{f}_t}$ be the Lipschitz constant of the parametric model $\hat{f}_t$, and $L_{\hat{f}_r}$ be the Lipschitz constant of $\hat{f}_r$.*

$$\varepsilon_{t,p}(x) \leq \hat{\varepsilon}_{t,p}(x) + L_t rad_n + L_{\hat{f}_t} rad_n \tag{31}$$
$$\varepsilon_{r,p}(x) \leq \hat{\varepsilon}_{r,p}(x) + L_r rad_n + L_{\hat{f}_r} rad_n \tag{32}$$

Proof. Let $x_t^{(i)}$ be the state closest to x whose action $a_t^{(i)}$ equals a.

$$\varepsilon_{t,p}(x) = \Delta(f_t(x,a), \hat{f}_t(x,a)) \tag{33}$$
$$\begin{aligned} \leq \quad & \Delta(f_t(x,a), f_t(x_t^{(i)}, a)) \\ & + \Delta(f_t(x_t^{(i)}, a), \hat{f}_t(x_t^{(i)}, a)) \\ & + \Delta(\hat{f}_t(x_t^{(i)}, a), \hat{f}_t(x,a)) \end{aligned} \tag{34}$$
$$\begin{aligned} \leq \quad & L_t \text{rad}_n + \Delta(f_t(x_t^{(i)}, a), \hat{f}_t(x_t^{(i)}, a)) \\ & + L_{\hat{f}_t} \text{rad}_n \end{aligned} \tag{35}$$

Since the closest sample $x_t^{(i)}$ is within distance C of the state of interest x by Assumption 2,

$$\Delta\left(f_t(x_t^{(i)}, a), \hat{f}_t(x_t^{(i)}, a)\right) = \Delta\left(\hat{f}_t(x_t^{(i)}, a), x_{t+1}^{(i)}\right) \tag{36}$$
$$\leq \max \Delta\left(\hat{f}_t(x_{t'}^{(i)}, a), x_{t'+1}^{(i)}\right) = \hat{\varepsilon}_{t,p} \tag{37}$$

So we finished the proof for $\varepsilon_{t,p}(x)$. Similarly we can show $\varepsilon_{r,p}(x) \leq \hat{\varepsilon}_{r,p}(x) + L_r \text{rad}_n + L_{\hat{f}_r} \text{rad}_n$ □

Now we are going to prove Theorem 2:

Theorem 2. *(Restated) Under the assumptions 1, 2, 3 in our appendix, assuming planning error $\epsilon_{planning} = o(1)$, the MoE simulator with MCTS model selection is a consistent estimator of policy value of π_e.*

Proof. By assuming the planning error of MCTS is bounded by $\epsilon_{\text{planning}}$, we have that the return of chosen node will be no less than the return of nonparametric model minus $\epsilon_{\text{planning}}$.

$$\max_{\nu' \in \text{children of } \nu} \widetilde{Q}(\nu') \tag{38}$$
$$\begin{aligned} \geq \quad & -L_r \sum_{t=0}^{T-t_0} \gamma^t \sum_{t'=0}^{t-1} (L_t)^{t'} \hat{\varepsilon}_{t,np}(t_0+t-t'-1) \\ & - \sum_{t=0}^{T-t_0} \gamma^t \hat{\varepsilon}_{r,np}(t_0+t) - \epsilon_{\text{planning}} \end{aligned} \tag{39}$$
$$\geq \quad -K \text{rad}_n - \epsilon_{\text{planning}} \tag{40}$$

where K is some constant independent of sample size n. By MCTS algorithm, we have that the return of chosen node is:

$$
\begin{aligned}
& \max_{\nu' \in \text{children of } \nu} \widetilde{Q}(\nu') && (41)\\
= \; & -L_r \sum_{t=0}^{T-t_0} \gamma^t \sum_{t'=0}^{t-1} (L_t)^{t'} \hat{\varepsilon}_{t,\text{MCTS}}(t_0+t-t'-1) \\
& - \sum_{t=0}^{T-t_0} \gamma^t \hat{\varepsilon}_{r,\text{MCTS}}(t_0+t) && (42)
\end{aligned}
$$

where $\varepsilon_{t,\text{MCTS}}(t)$ and $\varepsilon_{r,\text{MCTS}}(t)$ is the transition and reward error of the model selected by MCTS MoE model selection algorithm at each planning step. Thus

$$
\begin{aligned}
L_r\gamma\hat{\varepsilon}_{t,\text{MCTS}}(t_0) + \hat{\varepsilon}_{r,\text{MCTS}}(t_0) &\le -\widetilde{Q}(\nu') \\
&\le K\text{rad}_n + \epsilon_{\text{planning}} \qquad (43)
\end{aligned}
$$

Then we can bound the estimated one step transition and reward error of the chosen model by

$$
\begin{aligned}
\hat{\varepsilon}_{t,\text{MCTS}}(t_0) &\le K'(\text{rad}_n + \epsilon_{\text{planning}}) && (44)\\
\hat{\varepsilon}_{r,\text{MCTS}}(t_0) &\le K'(\text{rad}_n + \epsilon_{\text{planning}}) && (45)
\end{aligned}
$$

where K' is some other constant independent with sample size n. Now we need to bound the true one step transition and reward error of the chosen model $\varepsilon_{t,\text{MCTS}}(t_0)$ and $\varepsilon_{t,\text{MCTS}}(t_0)$. By Lemma 2 we know that we can bound it for non-parametric model for any state:

$$
\varepsilon_{t,np}(x) \le L_t\text{rad}_n, \quad \varepsilon_{r,np}(x) \le L_r\text{rad}_n \qquad (46)
$$

and Lemma 3 show that

$$
\begin{aligned}
\varepsilon_{t,p}(x) &\le \hat{\varepsilon}_{t,p}(x) + L_t\text{rad}_n + L_{\hat{f}_t}\text{rad}_n && (47)\\
\varepsilon_{r,p}(x) &\le \hat{\varepsilon}_{r,p}(x) + L_r\text{rad}_n + L_{\hat{f}_r}\text{rad}_n && (48)
\end{aligned}
$$

Then for both model we have that

$$
\begin{aligned}
\varepsilon_t(x) &\le \hat{\varepsilon}_t(x) + K''\text{rad}_n && (49)\\
\varepsilon_r(x) &\le \hat{\varepsilon}_r(x) + K''\text{rad}_n && (50)
\end{aligned}
$$

for some constant K''. Therefore for the chosen model, we can bound its one step transition error and reward error.

$$
\begin{aligned}
\varepsilon_{t,\text{MCTS}}(x) &\le \hat{\varepsilon}_{t,\text{MCTS}}(x) + K''\text{rad}_n \\
&= O(\text{rad}_n) + O(\epsilon_{\text{planning}}) && (51)\\
\varepsilon_{r,\text{MCTS}}(x) &\le \hat{\varepsilon}_{r,\text{MCTS}}(x) + K''\text{rad}_n \\
&= O(\text{rad}_n) + O(\epsilon_{\text{planning}}) && (52)
\end{aligned}
$$

Combining this with Theorem 1, we have that the total error of return could be bounded by $O(\text{rad}_n) + O(\epsilon_{\text{planning}})$. Thus, if $O(\epsilon_{\text{planning}}) = o(1)$, the total return error will also be bounded by $o(1)$ and MoE simulator with MCTSmodel selection is a consistent estimator. □

D. Consistency of MCTS-MoE Under Weaker Conditions

In our proof of theorem 2, we assume that the planning error $\epsilon_{\text{planning}}$ will converge to zero. If that is not true, we can still prove the consistency result with a slightly different variant of Algorithm 2. Consider if the condition in line 4 of Algorithm 2 changes to:

$$
\varepsilon_{t,p}(x) + \alpha_r\varepsilon_{r,p}(x) \le \hat{\varepsilon}_{t,p}(x) + \alpha_r\hat{\varepsilon}_{r,p}(x), \qquad (53)
$$

where the coefficient α_r is a constant factor just determined by the scale of reward and transition function. Then we can show a new theorem about Algorithm 1 with both greedy and MCTS model selection are consistent estimators i.e. the return error goes to zero when the number of samples collected from behavior policy goes to infinity. We keep the same assumptions (Assumption 1, 2, 3) for other parts of algorithm as last section.

Lemma 4. *MoE simulator with greedy model selection is a consistent estimator of the policy value of π_e.*

Theorem 3. *MoE simulator with MCTS model selection is a consistent estimator of the policy value of π_e.*

Proof sketch: Notice that only when $\hat{\varepsilon}_{t,p}(x) + \alpha_r\hat{\varepsilon}_{r,p}(x) \le \hat{\varepsilon}_{t,np}(x) + \alpha_r\hat{\varepsilon}_{r,np}(x)$ we will select parametric model. Then Lemma 4 can be proved by showing the greedy model is consistent since the nonparametric model is consistent. Thus we can further prove Theorem 3 by show that the MCTS policy will always choose a model better than greedy selection since greedy selection is the default roll out policy and the environment is deterministic.

We now show the proofs formally. Proof of Lemma 4:

Proof. We are going to show that the error of the return goes to zero as the number of samples goes to infinity. According to Theorem 1, we only need to show that $\varepsilon_{t,\text{greedy}}(t)$ and $\varepsilon_{r,\text{greedy}}(t)$ goes to zero for any time t where greedy $\in \{p, np\}$ is the model selected by greedy MoE model selection algorithm at time step t.

We showed in Lemma 2 that the non-parametric model error $\varepsilon_{t,np}(x)$ and $\varepsilon_{r,np}(x)$ goes to zero when n goes to infinity. Now we are going to show that we will select a parametric model at a given state x only if $\varepsilon_{t,p}(x) + \varepsilon_{r,p}(x)$ will also go to zero.

According to the greedy model selection algorithm, we will only select the parametric model when

$$
\hat{\varepsilon}_{t,p}(x) + \alpha_r\hat{\varepsilon}_{r,p}(x) \le \hat{\varepsilon}_{t,np}(x) + \alpha_r\hat{\varepsilon}_{r,np}(x)
$$

, where the coefficient α_r is a constant factor determined by the scale of reward and transition function. According to

Lemma 3,

$$\begin{aligned} & \varepsilon_{t,p}(x) + \alpha_r \varepsilon_{r,p}(x) && (54)\\ \leq \;& \hat{\varepsilon}_{t,p}(x) + \alpha_r \hat{\varepsilon}_{r,p}(x) + O(\text{rad}_n) && (55)\\ \leq \;& \hat{\varepsilon}_{t,np}(x) + \alpha_r \hat{\varepsilon}_{r,np}(x) + O(\text{rad}_n) && (56)\\ = O(\text{rad}_n) & && (57) \end{aligned}$$

Since $\lim_{n\to\infty} \text{rad}_n = 0$, for any chosen model at time step t, $\varepsilon_{t,\text{greedy}}(t)$ and $\varepsilon_{r,\text{greedy}}(t)$ is also $o(1)$. The proof follows from then applying Theorem 1 □

Proof of Theorem 3

Proof. According to the MCTS MoE model selection algorithm, for any input (s_{t_0}, a_{t_0}) we will at least have one roll-out trajectory following by the greedy MoE model selection. So the return of the chosen node is at least larger than this:

$$\begin{aligned} & \max_{\nu' \in \text{children of } \nu} \widetilde{Q}(\nu') && (58)\\ \geq \;& -L_r \sum_{t=0}^{T-t_0} \gamma^t \sum_{t'=0}^{t-1} (L_t)^{t'} \hat{\varepsilon}_{t,\text{greedy}}(t_0 + t - t' - 1) \\ & - \sum_{t=0}^{T-t_0} \gamma^t \hat{\varepsilon}_{r,\text{greedy}}(t_0 + t) && (59)\\ \geq \;& -K \text{rad}_n && (60) \end{aligned}$$

where K is some constant independent of sample size n. This follows from the fact that the estimated error of greedy selected model can be bounded by the estimated error of non-parametric model, and further bounded by $O(\text{rad}_n)$. By the MCTS algorithm, we have that the return of chosen node can be expressed as:

$$\begin{aligned} & \max_{\nu' \in \text{children of } \nu} \widetilde{Q}(\nu') && (61)\\ = \;& -L_r \sum_{t=0}^{T-t_0} \gamma^t \sum_{t'=0}^{t-1} (L_t)^{t'} \hat{\varepsilon}_{t,\text{MCTS}}(t_0 + t - t' - 1) \\ & - \sum_{t=0}^{T-t_0} \gamma^t \hat{\varepsilon}_{r,\text{MCTS}}(t_0 + t) && (62) \end{aligned}$$

where $\varepsilon_{t,\text{MCTS}}(t)$ and $\varepsilon_{r,\text{MCTS}}(t)$ are the transition and reward error of the model selected by MCTS MoE model selection algorithm. Thus

$$L_r \gamma \hat{\varepsilon}_{t,\text{MCTS}}(t_0) + \hat{\varepsilon}_{r,\text{MCTS}}(t_0) \leq -\widetilde{Q}(\nu') \leq K \cdot \text{rad}_n \qquad (63)$$

Thus, there exist another constant K' such that the one step transition and reward error of the chosen model satisfy that

$$\begin{aligned} \hat{\varepsilon}_{t,\text{MCTS}}(t_0) &\leq K' \text{rad}_n && (64)\\ \hat{\varepsilon}_{r,\text{MCTS}}(t_0) &\leq K' \text{rad}_n && (65) \end{aligned}$$

Now we need to bound the true one step transition and reward error of the chosen model $\varepsilon_{t,\text{MCTS}}(t_0)$ and $\varepsilon_{t,\text{MCTS}}(t_0)$. By Lemma 2 we know that we can bound it for non-parametric model for any state:

$$\varepsilon_{t,np}(x) \leq L_t \text{rad}_n, \quad \varepsilon_{r,np}(x) \leq L_r \text{rad}_n \qquad (66)$$

and Lemma 3 show that

$$\begin{aligned} \varepsilon_{t,p}(x) &\leq \hat{\varepsilon}_{t,p}(x) + L_t \text{rad}_n + L_{\hat{f}_t} \text{rad}_n && (67)\\ \varepsilon_{r,p}(x) &\leq \hat{\varepsilon}_{r,p}(x) + L_r \text{rad}_n + L_{\hat{f}_r} \text{rad}_n && (68) \end{aligned}$$

Then for both model we have that

$$\begin{aligned} \varepsilon_t(x) &\leq \hat{\varepsilon}_t(x) + K'' \text{rad}_n && (69)\\ \varepsilon_r(x) &\leq \hat{\varepsilon}_r(x) + K'' \text{rad}_n && (70) \end{aligned}$$

for some constant K''. Therefore for the chosen model, we can bound its one step transition error and reward error.

$$\begin{aligned} \varepsilon_{t,\text{MCTS}}(x) &\leq \hat{\varepsilon}_{t,\text{MCTS}}(x) + K'' \text{rad}_n = O(\text{rad}_n) && (71)\\ \varepsilon_{r,\text{MCTS}}(x) &\leq \hat{\varepsilon}_{r,\text{MCTS}}(x) + K'' \text{rad}_n = O(\text{rad}_n) && (72) \end{aligned}$$

Combining this with Theorem 1, we have that the total error of return could be bounded by $O(\text{rad}_n)$ and goes to zero as n goes to infinity. □

E. Evaluation of Model Error Estimators

In this section we empirically investigate the quality of the estimators we use for the error of the transition function, by analyzing their performance on the example presented in section 5.1. In Figure 1a we plot the true error of the non-parametric model as a function of coordinate for the action ”North”, and compare it with the estimate from Equation 9 in the main text, shown in Figure 1b. Figures 1e and 1f are the equivalent figures for the parametric model. Comparing the errors shown in Figures 1a and 1e indicates whether the parametric or nonparametric model should be selected, and the correct selection based on the true errors is presented in Figure 1i. Similarly by comparing the errors presented in Figures 1b and 1f, we present in Figure 1j which model our MoE model would actually select. Finally, in Figure 1m we compare Figures 1i and 1j to show if the MoE model would make the correct choice in which model to use. Similar analyses is presented on the right half of Figure 1 for the ”East” action.

We see that the nonparametric model has small error for the areas where trajectories in the data pass through, and the error increases with distance from clusters of observations. The simple parametric model, on the other hand, has errors which are uncorrelated with the density of observations (In all other domains we will present in this paper this will not be the case, as we will learn the parametric model from the data, and therefore expect the parametric model to be more accurate in regions where we have observations of transitions). Our estimates for the error follow this general trend, and more importantly they properly identify the model with the smaller error over most of the space (Figures 1m and 1n).

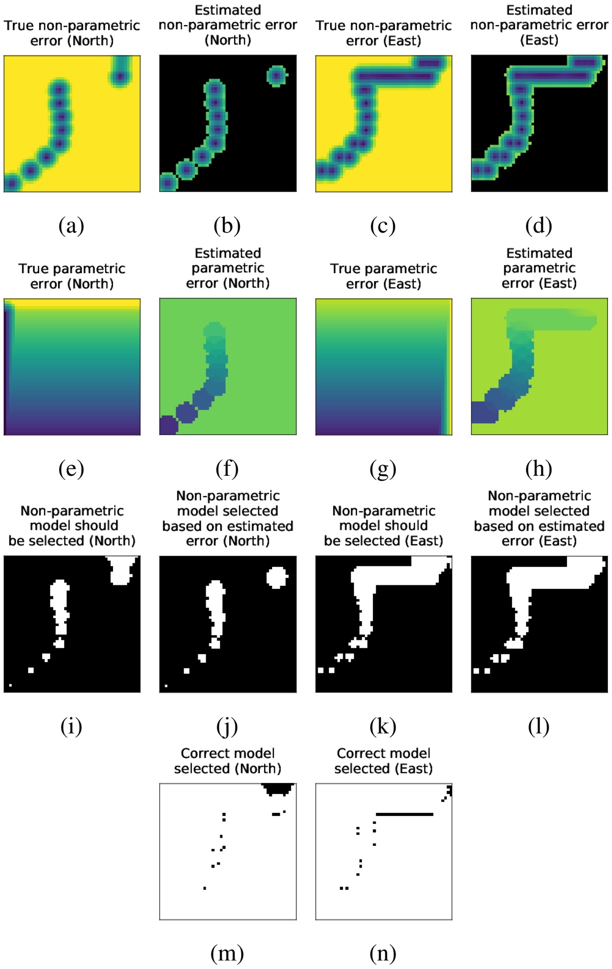


Figure 1. **Empirical evaluation of error estimates for model errors.** For the 2D gridworld example described in Section 5.1 the estimators we use for the model errors resemble the true errors of both the parametric and non-parametric models (a-h). More importantly, these estimators allow the MoE model to correctly select the model with the lower prediction error on the transition (i-n). (All heatmaps figures are presented in the same color scale)

F. Experimental Details

The dynamics of the cancer domain follow the ODEs presented in (Ribba et al., 2012) which model the response of cancer cells to treatment. The state space consists of 4 features representing cell counts and medication concentrations, and each time step represent a month in which a clinician may choose between administering a particular treatment or avoiding treatment. The reward at each time step is the total change in diameter of cancerous cells. To learn the parametric model we fit a linear regression model to predict the dynamics of the states given each action.

The HIV domain is described in Ernst et al. (2006), and consists of 6 parameters describing the state of the patient and 4 possible actions. As the reward function we use the reward described in Ernst et al. (2006). As the parametric model we use a feed-forward neural network with two layers, each consisting of 50 hidden units and a $\tanh$ activation function.

Evaluation and behavior policies. For the cancer domain, we test an evaluation policy which treats the patient every month for 10 months, and then stops treatment. As behavior policy we use an ϵ-greedy version of the evaluation policy. For each value of ϵ we run 500 experiments in which we generate 10 trajectories for learning the models.

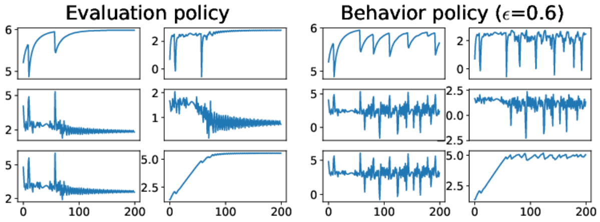
In the HIV domain, we use fitted Q iterations to learn an optimal policy. Under this policy — whose trajectory is shown in Figure 2a as the time evolution of the 6 state dimensions — patients start in a state with a high viral load, which decreases over roughly 70 treatment steps. After the patient is brought to a steady state with low viral load, the continued treatment keeps the patient stabilized. As a behavior policy, we use a policy which is identical to the evaluation policy when the patient is far away from the stable state, and switch to an ϵ-greedy policy around the steady state. This can be thought of as a likely real world scenario where clinicians know how to treat severely ill patients, but are less certain about how to keep them stable in the long run when their condition is not critical. More explicitly, the behavior policy follows the evaluation policy for $\log E < 4$, where E is the number of immune effectors, whose evolution is shown in the bottom right plot in Figure 2a , and switches to ϵ-greedy when $\log E > 4$. For each value of ϵ we run 100 experiments in which 5 trajectories are generated and used for learning the models.

F.1. Comparison with IS methods

In section 6.2 we compared the parametric and nonparametric models, as well as our greedy MoE model to two common importance sampling estimators. In Table 1 here we provide additional results for more importance sampling based estimators - standard importance sampling (IS), weighted importance sampling (WIS), per-decision importance sampling (PDIS), consistent weighted per-decision importance sampling (CWPDIS), doubly robust (DR) and weighted doubly robust (WDR) (Precup, 2000; Jiang & Li, 2016; Thomas & Brunskill, 2016; Thomas, 2015). The DR and WDR estimators require independent estimates of state values, which we obtain using the parametric model. These

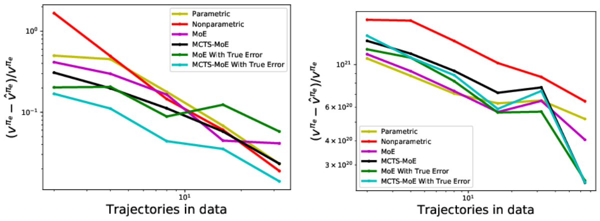
Table 1. $\sqrt{\mathbb{E}[(v^{\pi_e} - \hat{v}^{\pi_e})^2]}/v^{\pi_e}$; $(\epsilon = 0.4)$

	M_{p}	M_{np}	M_{MoE}	$M_{\text{MCTS-MoE}}$	IS	WIS	PDIS	CWPDIS	DR	WDR
Cancer	0.021	0.027	0.020	**0.019**	1.0	1.0	0.55	0.22	0.87	0.22
HIV	0.65	0.88	0.64	**0.63**	1.0	1.0	0.66	0.99	89.2	0.99



(a) HIV (b) HIV

Figure 2. **Trajectories in the true environment generated by the evaluation and behavior policies.** The behavior policy is similar to the evaluation policy for the initial part of the trajectory (roughly for the first 70 steps) and becomes ε-greedy near the steady state, as can be seen by the more erratic nature of the trajectories for late time steps.



(a) Cancer (b) HIV

Figure 3. **Empirical check of consistency.** For both medical simulators the value estimation error decreases as the number of observed trajectories is increased. For both domains we the behavior policy is the ϵ-greed policy with $\epsilon = 0.4$.

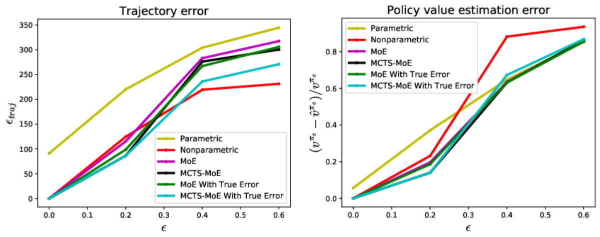
results demonstrate that for regimes with limited amount of data, even for moderate trajectory lengths (30 steps for the cancer simulator), all IS based estimators fail due to extremely small effective sample sizes (Liu et al., 2018; Gottesman et al., 2018), and therefore we must resort to model based estimators.

F.2. Empirical test for consistency

In this section we empirically test the consistency of the MoE simulators and demonstrate in Figure 3 that as the number of observed trajectories increases, the value estimation error for both domains decreases across all models. In the cancer domain we see that with access to the true error, the MCTS-MoE consistently outperforms all other methods. For the HIV domain we observe that minimizing the trajectory simulation accuracy does not imply minimizing the value estimation error due to improper choice of metric, as discussed in Appendix F.3.

F.3. Effect of the metric on value estimation for HIV

When presenting the results for the HIV simulator, we noted that for high values of randomness in the behavior policy, the MCTS-MoE is outperformed by the parametric model and the greedy MoE, despite performing well in terms of the trajectory error. We argued this effect can be attributed to the distance metric used to quantify the transition error, which does not take into account the fact that some dimensions are more strongly correlated with the reward than others. This claim is further supported by the observation that in the regime where the MCTS-MoE performs poorly in terms of value estimation, the nonparametric performs significantly worse than all other methods, despite performing reasonably well in terms of trajectory error.



(a) HIV (b) HIV

Figure 4. **Effect of the metric.** By replacing the Euclidean distance with a metric that puts more weight on state dimensions which are strongly correlated with reward, the value estimation performance of the MCTS-MoE can be improved at the cost of trajectory error.

To further investigate the effect of the metric we ran our experiments again but used a metric which gives 20 times more weight to the 6^{th} dimension in the state space. This dimension (bottom right plot in Figures 2a) and 2b) represents the number of immune effectors in the patient's body and is most strongly correlated with the reward. In Figure 4 we present the results for OPE on the HIV simulator with this new metric and demonstrate that indeed using this new metric improves the performance of the MCTS-MoE simulator in terms of value estimation, at the cost of degrading the trajectory error.

References

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

Coulom, R. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pp. 72–83. Springer, 2006.

Ernst, D., Stan, G.-B., Goncalves, J., and Wehenkel, L. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Decision and Control, 2006 45th IEEE Conference on*, pp. 667–672. IEEE, 2006.

Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., Zhang, L., Ding, Y., Wihl, D., Peng, X., et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pp. 652–661. JMLR. org, 2016.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Ribba, B., Kaloshi, G., Peyre, M., Ricard, D., Calvez, V., Tod, M., Bernard, B. C., Idbaih, A., Psimaras, D., Dainese, L., et al. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. *Clinical Cancer Research*, pp. clincanres–0084, 2012.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

Thomas, P. S. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Libraries, 2015.