How to Select the Appropriate One From the Trained Models for Model-Based OPE

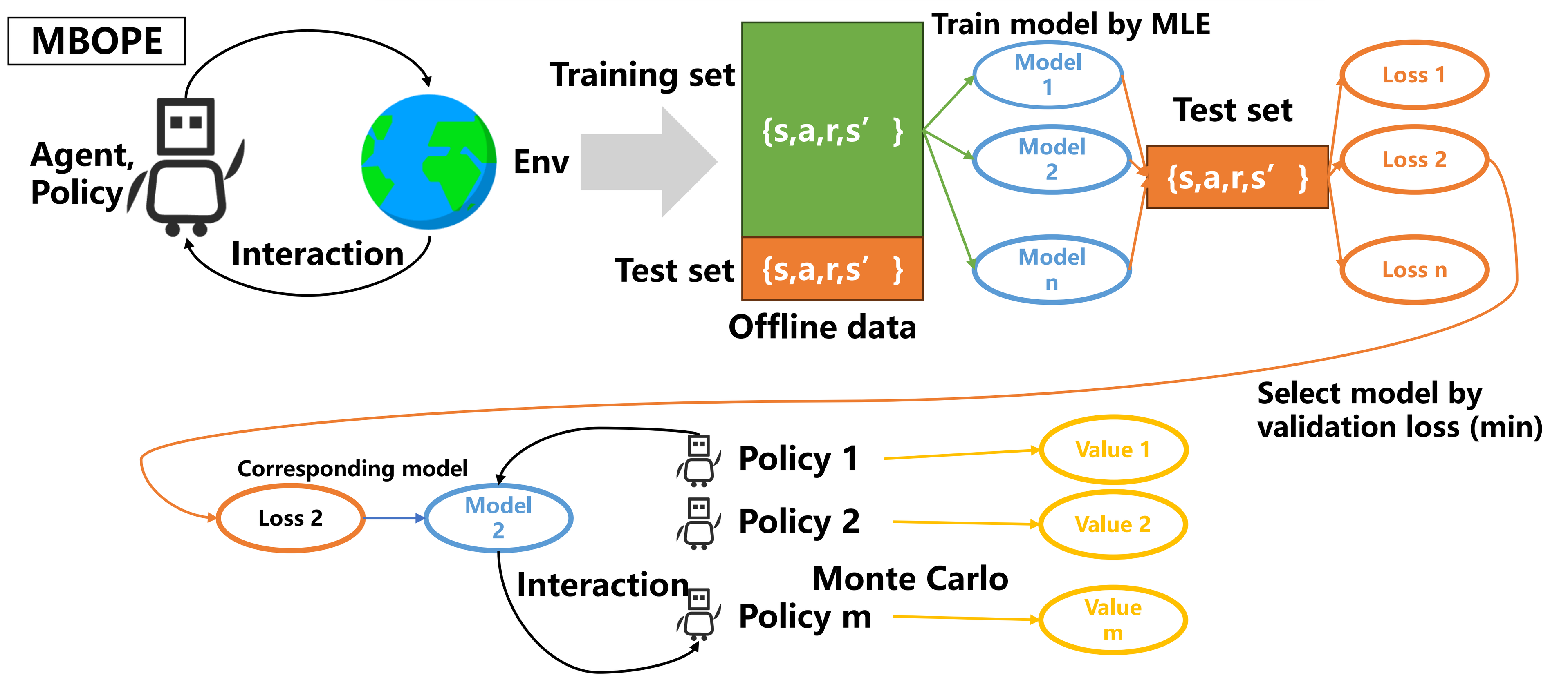
Chongchong Li[1], Yue Wang[1], Zhi-Ming Ma[2], and Yuting Liu[1]
[1]Beijing Jiaotong University, [2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences

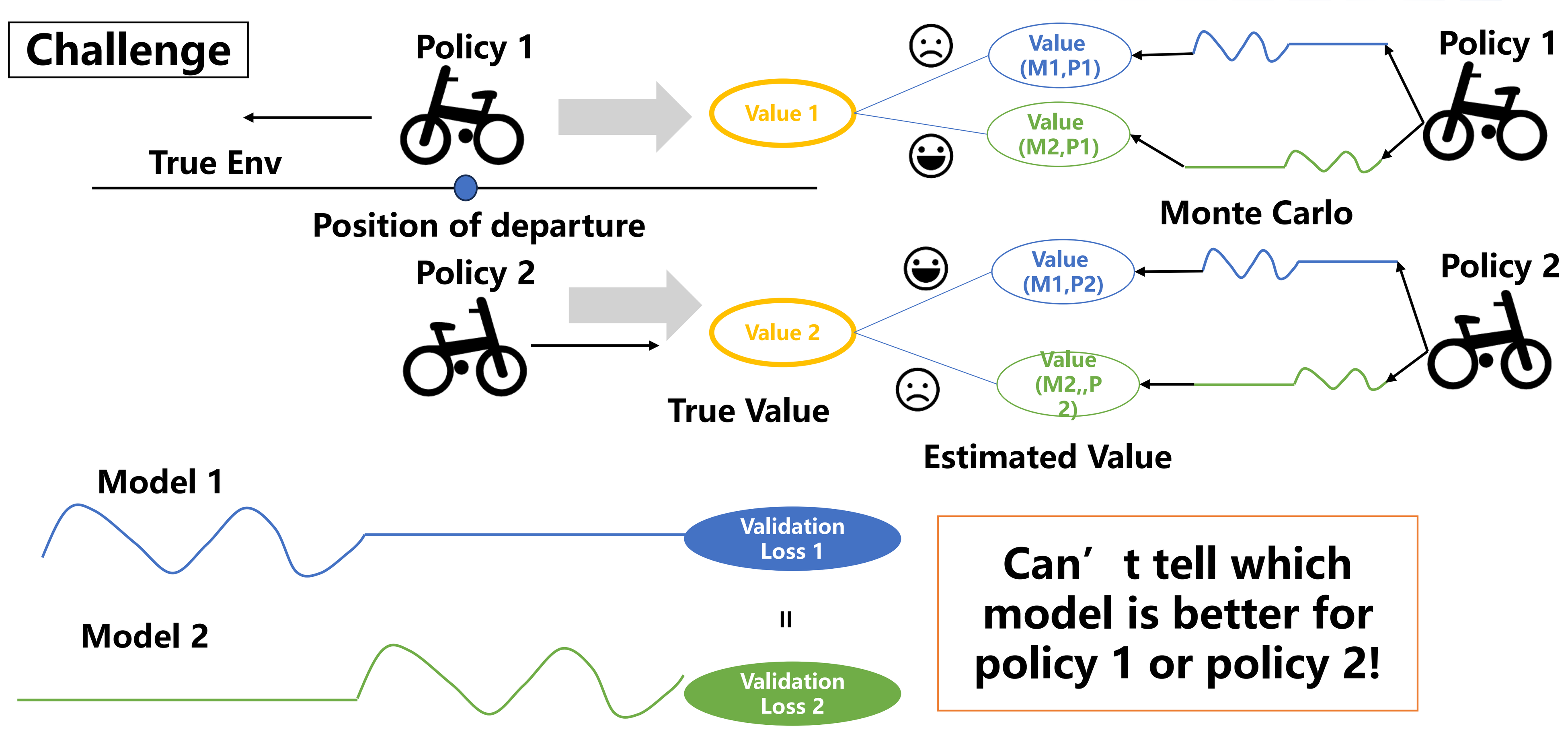
CICAI 2023
CAAI International Conference on Artificial Intelligence

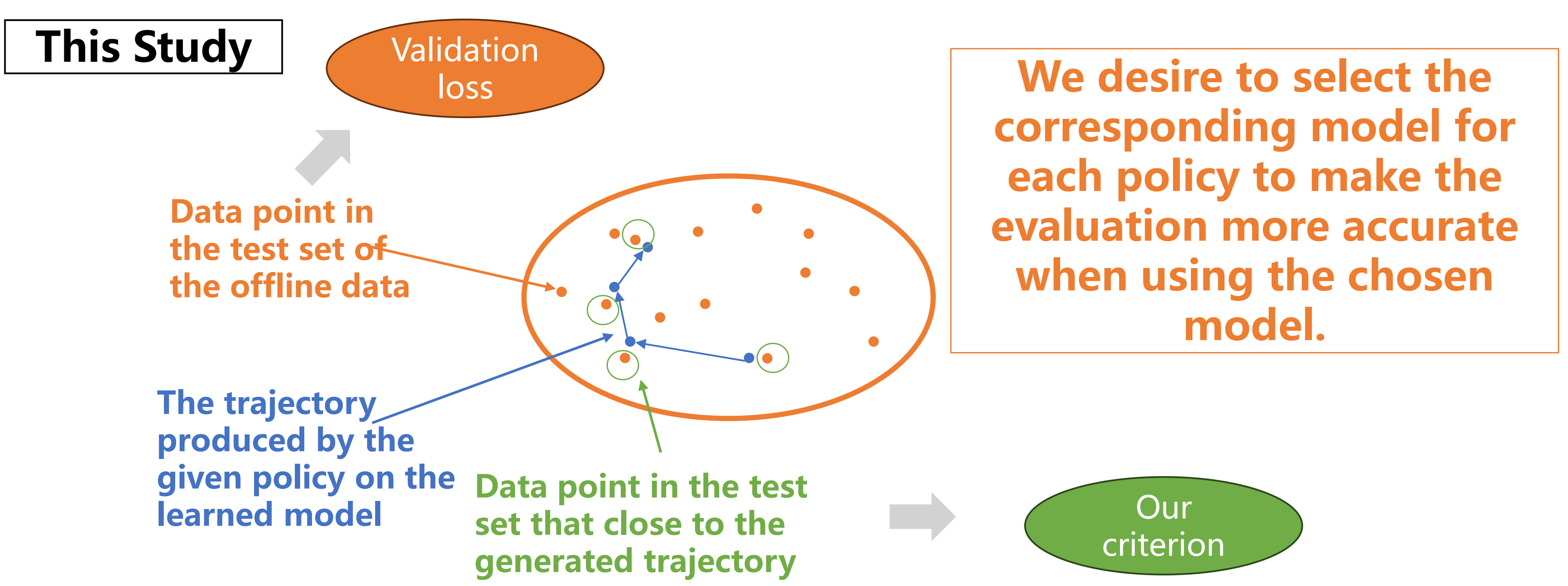
1. Motivation

- **Model-Based Offline Policy Evaluation (MBOPE) :**
Approximate the value of a given policy directly by estimated transition and reward functions of the environment + Monte Carlo.
- **A challenge remains in selecting an appropriate model from the trained models:**
Traditional method: train a set of models (ensemble) + select by comparing the validation loss. The local errors of the model and the degree of fit with the policy to be evaluated are ignored.
- **This study:**
Explore the criterion for selecting models from trained models.

2. Graphical Representation







3. Theoretical Analysis

Theorem 1 (the upper bound of the discrepancy between the actual value and the approximated value calculated using the learned model.)

Let η_π be the true value of the policy π, $\hat{\eta}_\pi$ be the estimated value using the learned model, then we have:

$$|\eta_\pi - \hat{\eta}_\pi| \le C \cdot \mathbb{E}_{t\sim Gemo(\gamma)} \mathbb{E}_{s' \sim P_{mix}^t, a' \sim \pi(s')} \mathcal{L}(s', a'),$$

Where $P_{mix}^t = \beta P_\pi^t + (1-\beta)\hat{P}_\pi^t$, $C = \frac{2\gamma r_{max}}{(1-\gamma)^2} \cdot \mathcal{L}(s', a')$, the error of the model at (s', a') is $D_{TV}(P(s|s', a') || \hat{P}(s|s', a'))$ and $Gemo(\gamma)$ is a geometric distribution with parameter γ.

- The error can be upper bounded by the expected error of the learned model over the distribution of trajectories produced by that policy.
- The error depends on how the agent generates trajectories on the learned model and actual environment. The error of the learned model at these generated data points also plays a role.
- The geometric distribution shows that the error of the estimated value of the given policy is more relevant to the front part of the resulting trajectory.
- In this study β is set to zero to provide a more convenient bound since it is not possible to gather trajectories using the given policy on offline policy evaluation tasks.

4. Proposed Method

Geometric Loss Criterion

$$C \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{H} g_t \left(d\left((\hat{s}_t^i, a_t^i), (s_t^{i*}, a_t^{i*}) \right) + l(s_t^{i*}, a_t^{i*}) \right)$$

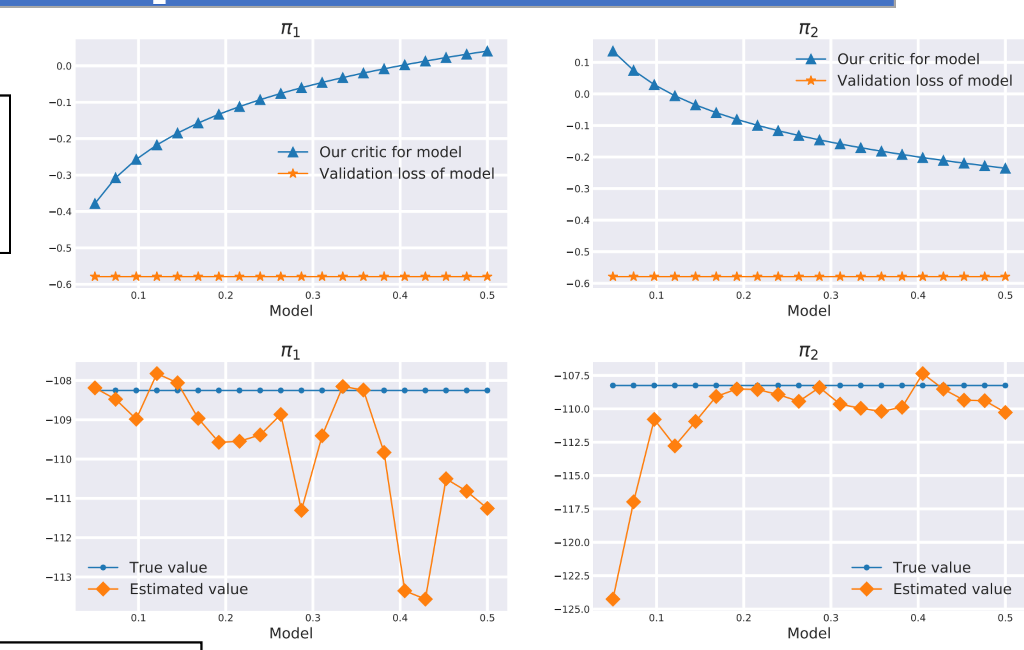
where C is $\frac{2\gamma r_{max}}{(1-\gamma)^2}$. The variable g_t represents the probability of the geo metric distribution for sampling t. And d is the distance (MSE) of the generated data and the corresponding nearest data point in validation. l is the prediction loss of the learned model on (s_t^{i*}, a_t^{i*}).

Algorithm 1 A Criterion for Choosing the Trained Model

```
Require: the learned models {M̂_k}_{k=1}^K, policy π_j, discount factor γ, horizon H, set of
    initial states S_0, batch size N, r_max which is the maximum of the reward in offline
    data set.
for k=1,...,K do
    B_k ← 0
    for i = 1,2,...,N do
        B_k^i ← 0
        Sample initial state s_0 from S_0.
        ŝ_0^i = s_0
        for t = 0,1,...,H−1 do
            Sample action using policy π, a_t^i ~ π(ŝ_t^i)
            Sample next state and reward using M̂_k: ŝ_{t+1}^i, r̂_{t+1}^i ~ M̂_k(ŝ_t, a_t^i)
            Find the nearest data of (ŝ_t^i, a_t^i) in validation: (ŝ_t^{i*}, a_t^{i*})
            Calculated the distance d of (ŝ_t^i, a_t^i) and (ŝ_t^{i*}, a_t^{i*}).
            Find the model prediction error l on (ŝ_t^{i*}, a_t^{i*}) in the validation set
            B_k^i ← B_k^i + (1−γ)γ^t (d + l)
        end for
        B_k^i ← B_k^i · 2γr_max/(1−γ)^2
    end for
    B_k = (1/N) Σ_{i=1}^N B_k^i
end for
k* = arg min_k B_k
return the index k* of the best model for π_j
```

5. Experimental Results

Synthetic (bicycle)



Benchmark (DM control)

- **Env**: Deepmind control-suite
- **Offline data**: RL Unplugged
- **Policy data**: 96 policies from BC, D4PG, CRR, RABM
- **Models**: 48 models for each env(different hyper-parameters)
- **Compare**: select by validation loss and by our critic

Correlation coefficient results

Correlation coefficient results

Environment	Validation loss	Ours w/o d	Ours
cartpole swingup	-0.01±0.29	0.28±0.15	**0.45±0.20**
cheetah run	0.21±0.14	0.18±0.15	**0.33±0.13**
finger turn hard	-0.30±0.16	**-0.17±0.09**	-0.19±0.09
fish swim	0.15±0.12	0.12±0.10	**0.42±0.26**
walker stand	-0.20±0.07	-0.20±0.07	**0.18±0.21**
walker walk	0.15±0.24	**0.16±0.28**	**0.16±0.29**
humanoid run	-0.06±0.04	-0.06±0.04	**0.01±0.03**

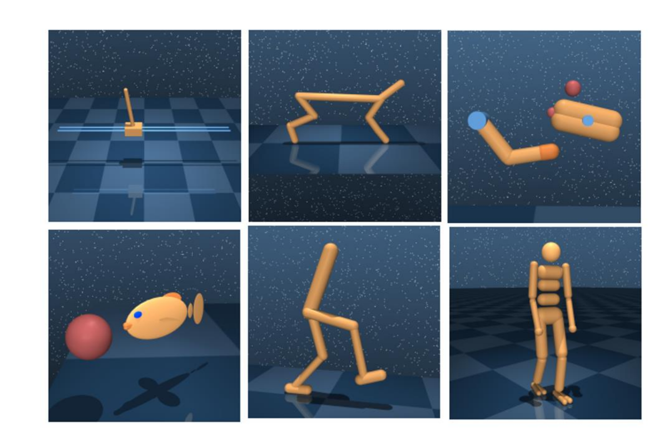
Average Absolute Error results

Environment	Validation loss	Ours w/o d	Ours
cartpole swingup	26.9±15.1	24.2±19.5	**17.5±24.2**
cheetah run	13.4±8.37	13.2±8.37	**8.52±8.49**
finger turn hard	**31.0±19.4**	35.5±24.6	34.2±29.0
fish swim	27.5±14.2	27.7±14.3	**20.1±15.5**
walker stand	66.5±27.5	**55.5±27.6**	**58.5±27.3**
walker walk	66.7±28.3	**57.1±30.5**	**59.6±30.0**
humanoid run	34.3±22.4	**32.2±24.8**	35.4±27.1

Spearman' s rank correlation between ground truth values and the estimated ones.

Environment	Validation loss	Ours w/o d	Ours
cartpole swingup	0.71	0.72	**0.73**
cheetah run	0.55	0.53	**0.60**
finger turn hard	**0.08**	-0.13	-0.03
fish swim	0.35	0.37	**0.50**
walker stand	0.32	0.14	**0.41**
walker walk	0.20	**0.38**	0.24
humanoid run	-0.58	**0.49**	0.24

Env



chongchongli@bjtu.edu.cn or ytliu@bjtu.edu.cn