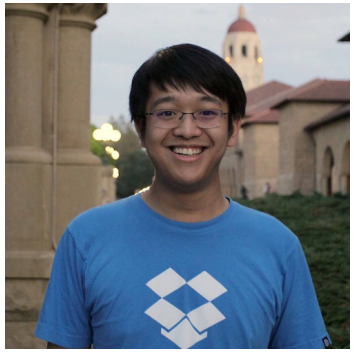
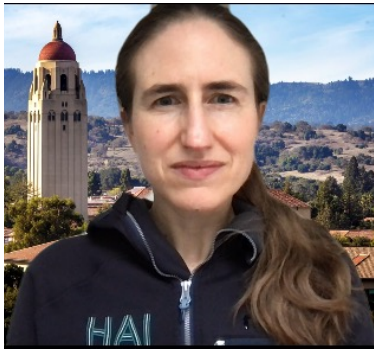
Model-based Offline Reinforcement Learning with Local Misspecification

Kefan Dong*, Yannis Flet-Berliac*, Allen Nie*, Emma Brunskill

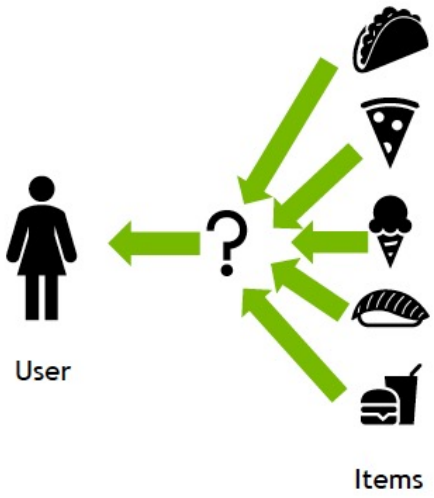


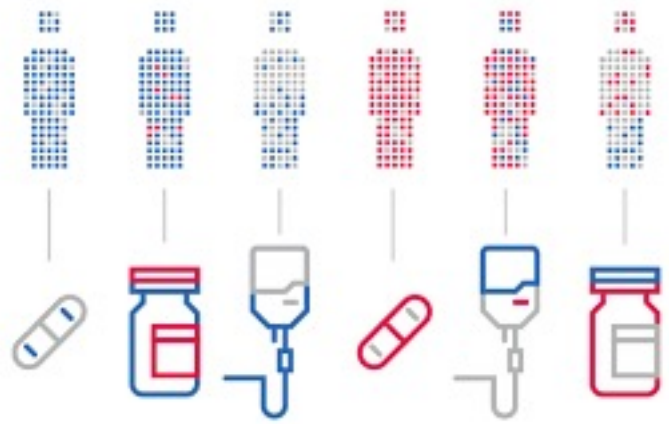


Stanford | ENGINEERING
Computer Science

Offline policy learning: motivation

- There exists lots of historical data for many real-world decision-making problems
  - E.g., recommendation systems, personalized healthcare





- Can we take historical data and output better policies?

Figures from https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/
https://global.colleaga.org/article/explaining-personalized-healthcare-0

Offline policy learning: main challenges

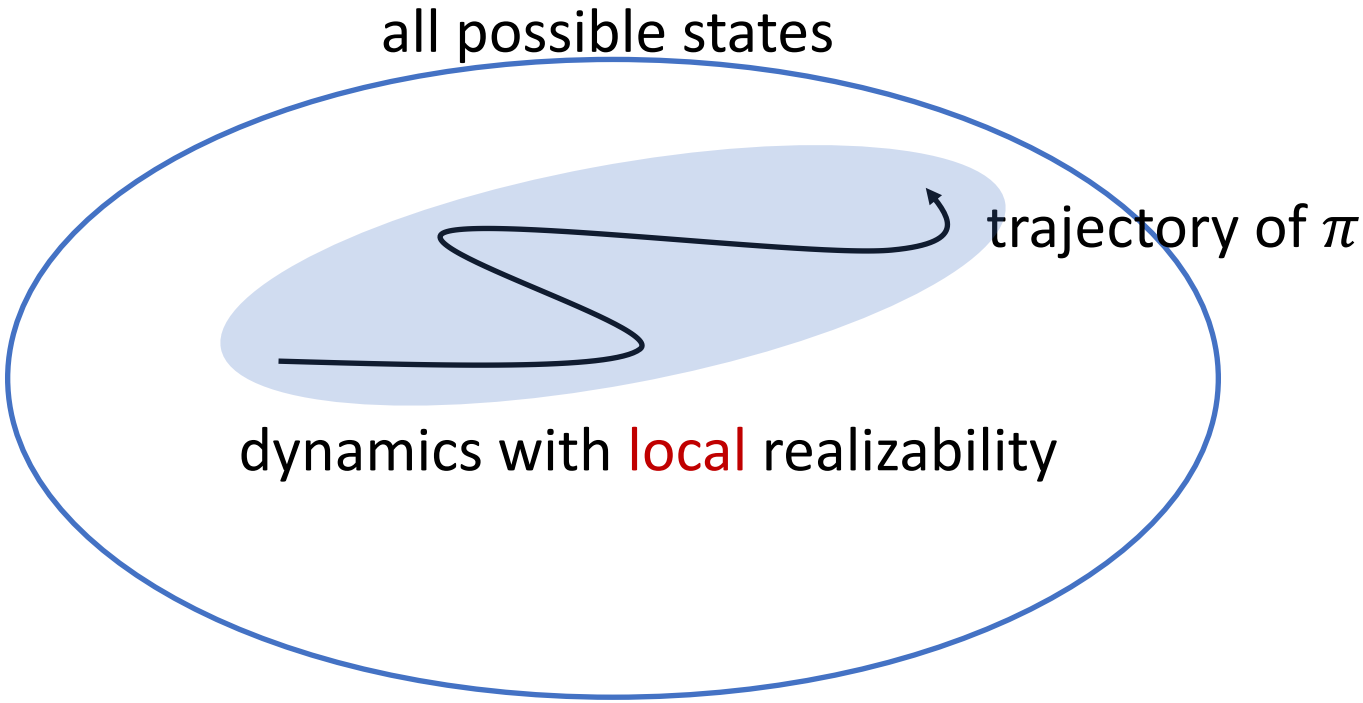
- Distribution mismatch: new policies generates different trajectories (data distribution), but we only have a fixed dataset
  - Especially with long horizons
- Bias-variance tradeoff: most off-policy estimators have to balance bias and variance
  - Especially when models are misspecified
  - More complicated for policy selection instead of policy evaluation
- We need theoretical guarantees in higher stakes domains (e.g., healthcare)

Existing works and limitations

- Importance-sampling estimators: minimal assumptions on the model, low data efficiency [Thomas et al., 2019]
- Offline RL: strong realizability assumptions, high data efficiency
  - E.g., an accurate model on *every* state [Xie & Jiang, 2021, Uehara & Sun, 2021, Voloshin et al., 2021]
- Can we make less assumptions and still have high data efficiency?
- This paper: leverage misspecified models (approximate parts of the state space)

Our insight: global realizability is not necessary

- For any fixed policy π, we only care about a subset of states
- Models with local realizability: exists an accurate model on states that may be *visited* by the policy π



Main results: offline policy optimization via locally realizable dynamics models

1. We design a model-based offline RL algorithm (MBLB) for policy optimization using locally realizable models.
   - Algorithmic insight: optimize the policy and dynamics together
   - Focus on the state-actions that visited by the current policy π
   - => Use different dynamics models for different policies
2. We prove a suboptimality upper bound for the learned policy in the finite sample regime.
3. We implement the MBLB algorithm on policy selection tasks in D4RL dataset and achieves competitive performance to existing results.

Offline policy learning: problem setup

- Infinite horizon MDPs with discount factor γ, ground-truth dynamics $T^\star$
- Dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{i=1,\ldots,n}$ drawn from behavior distribution μ
- Reward $r(s, a)$ is known
- Expected total reward

$$\eta(T^\star, \pi) = \mathbb{E}_{a_t \sim \pi(s_t), s_{t+1} \sim T^\star(s_t, a_t)}[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t)]$$

- Goal of the offline policy selection problem
  - Input: given a (small) policy class Π, the dataset $\mathcal{D}$
  - Output: the best policy in the policy class

$$\hat{\pi} = \mathrm{argmax}_{\pi \in \Pi} \, \eta(T^\star, \pi)$$

- Assumptions: locally realizable model class $\mathcal{T}$ and value class $\mathcal{G}$

Model-based pessimistic policy optimization

- Build a lower bound $\mathrm{lb}(T, \pi)$ for the expected return of π on the ground-truth dynamics

Simulation lemma: (Kakade and Langford, 2002)

$$\eta(T, \pi) - \eta(T^\star, \pi) = \frac{\gamma}{1-\gamma} \mathbb{E}_{(s,a)\sim\rho_T^\pi} \left[\mathbb{E}_{x\prime\sim T(s,a)} \left[V_{T^\star}^\pi(x') \right] - \mathbb{E}_{x\prime\sim T^\star(s,a)} \left[V_{T^\star}^\pi(x') \right] \right]$$

T: learned dynamics; $T^\star$: true dynamics

ρ_T^π : state-action distribution of running π on dynamics model T

$V_{T^\star}^\pi(x')$: expected total return of π starting from state x'

Model-based pessimistic policy optimization

- Build a lower bound $\mathrm{lb}(T, \pi)$ for the expected return of π on the ground-truth dynamics

Simulation lemma: (Kakade and Langford, 2002)

$$\eta(T, \pi) - \eta(T^\star, \pi) = \frac{\gamma}{1-\gamma} \mathbb{E}_{(s,a)\sim\rho_T^\pi} \left[\mathbb{E}_{x\prime\sim T(s,a)} \left[V_{T^\star}^\pi(x') \right] - \mathbb{E}_{x\prime\sim T^\star(s,a)} \left[V_{T^\star}^\pi(x') \right] \right]$$

T: learned dynamics; $T^\star$: true dynamics

ρ_T^π : state-action distribution of running π on dynamics model T

$V_{T^\star}^\pi(x')$: expected total return of π starting from state x'

Model-based pessimistic policy optimization

- Build a lower bound $\mathrm{lb}(T, \pi)$ for the expected return of π on the ground-truth dynamics

Simulation lemma: (Kakade and Langford, 2002)

$$\eta(T, \pi) - \eta(T^\star, \pi) = \frac{\gamma}{1-\gamma} \mathbb{E}_{(s,a)\sim\rho_T^\pi} \left[\mathbb{E}_{x\prime\sim T(s,a)} \left[V_{T^\star}^\pi(x') \right] - \mathbb{E}_{x\prime\sim T^\star(s,a)} \left[V_{T^\star}^\pi(x') \right] \right]$$

Model-based pessimistic policy optimization

- Build a lower bound $\mathrm{lb}(T, \pi)$ for the expected return of π on the ground-truth dynamics

Simulation lemma: (Kakade and Langford, 2002)

$$\eta(T,\pi) - \eta(T^\star,\pi) = \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{E}_{x\prime\sim T(s,a)}\left[V_{T^\star}^\pi(x')\right] - \mathbb{E}_{x\prime\sim T^\star(s,a)}\left[V_{T^\star}^\pi(x')\right]\right]$$

$$= \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a,s\prime)\sim\mathcal{D}}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)}\left(\mathbb{E}_{x\prime\sim T(s,a)}\left[V_{T^\star}^\pi(x')\right] - V_{T^\star}^\pi(s')\right)\right]$$

Model-based pessimistic policy optimization

- Build a lower bound $\mathrm{lb}(T, \pi)$ for the expected return of π on the ground-truth dynamics

Simulation lemma: (Kakade and Langford, 2002)

$$\eta(T,\pi) - \eta(T^\star,\pi) = \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{E}_{x\prime\sim T(s,a)}\left[V_{T^\star}^\pi(x')\right] - \mathbb{E}_{x\prime\sim T^\star(s,a)}\left[V_{T^\star}^\pi(x')\right]\right]$$

$$= \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a,s\prime)\sim\mathcal{D}}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)}\left(\mathbb{E}_{x\prime\sim T(s,a)}\left[V_{T^\star}^\pi(x')\right] - V_{T^\star}^\pi(s')\right)\right]$$

Model-based pessimistic policy optimization

- Build a lower bound $\mathrm{lb}(T, \pi)$ for the expected return of π on the ground-truth dynamics

Simulation lemma: (Kakade and Langford, 2002)

$$\eta(T,\pi) - \eta(T^\star,\pi) = \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{E}_{x\prime\sim T(s,a)}\left[V_{T^\star}^\pi(x')\right] - \mathbb{E}_{x\prime\sim T^\star(s,a)}\left[V_{T^\star}^\pi(x')\right]\right]$$

$$= \frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a,s\prime)\sim\mathcal{D}}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)}\left(\mathbb{E}_{x\prime\sim T(s,a)}\left[V_{T^\star}^\pi(x')\right] - V_{T^\star}^\pi(s')\right)\right]$$

$$\leq \frac{\gamma}{1-\gamma}\sup_{g\in\mathcal{G}}\left|\mathbb{E}_{(s,a,s\prime)\sim\mathcal{D}}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)}\left(\mathbb{E}_{x\prime\sim T(s,a)}[g(x')] - g(s')\right)\right]\right|$$

$\mathcal{G}$: value function class

Model-based pessimistic policy optimization

- Build a lower bound $\mathrm{lb}(T,\pi)$ for the expected return of π on the ground-truth dynamics

Simulation lemma: (Kakade and Langford, 2002)

$$\eta(T,\pi)-\eta(T^\star,\pi)=\frac{\gamma}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_T^\pi}\left[\mathbb{E}_{x\prime\sim T(s,a)}\left[V_{T^\star}^\pi(x')\right]-\mathbb{E}_{x\prime\sim T^\star(s,a)}\left[V_{T^\star}^\pi(x')\right]\right]$$

- Define the lower bound

$$\mathrm{lb}(T,\pi)\stackrel{\mathrm{def}}{=}\eta(T,\pi)-\frac{1}{1-\gamma}\sup_{g\in\mathcal{G}}\left|\mathbb{E}_{(s,a,s\prime)\sim\mathcal{D}}\left[\frac{\rho_T^\pi(s,a)}{\mu(s,a)}\left(\mathbb{E}_{x\prime\sim T(s,a)}[g(x')]-g(s')\right)\right]\right|$$

Algorithm (MBLB)

- Output $\hat{\pi}\leftarrow\mathrm{argmax}_{\pi\in\Pi}\max_{T\in\mathcal{T}}\mathrm{lb}(T,\pi)$

Safe policy improvement theorem

Theorem 4 (informal): The output of MBLB $\hat{\pi}$ satisfies

$$\eta(T^\star, \hat{\pi}) \geq \sup_{\pi \in \Pi} \eta(T^\star, \pi) - \frac{\epsilon_\rho(\pi)}{(1-\gamma)^3} - \tilde{O}(1/\sqrt{n})$$

- The local misspecification error $\epsilon_\rho(\pi) = 0$ if a dynamics model $T \in \mathcal{T}$ is accurate state-actions that may be *visited* by the policy π
- Prior works typically require $\inf_{T \in \mathcal{T}} \sup_{s,a} \text{dist}\big(T(s,a), T^\star(s,a)\big) \leq \epsilon$

Implementation of MBLB

Algorithm (MBLB)

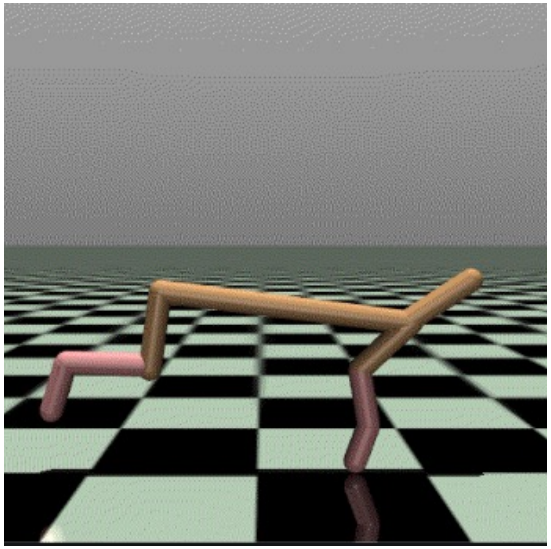
- Output $\hat{\pi} \leftarrow \operatorname{argmax}_{\pi \in \Pi} \max_{T \in \mathcal{T}} \mathrm{lb}(T, \pi)$

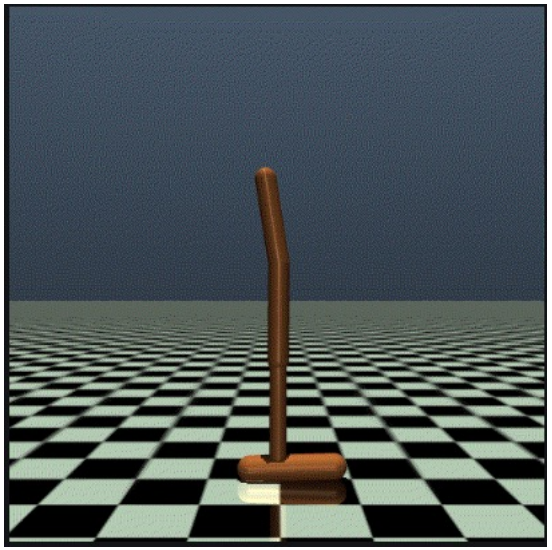
$$\mathrm{lb}(T,\pi) \stackrel{\text{def}}{=} \eta(T,\pi) - \frac{1}{1-\gamma} \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{(s,a,s') \sim \mathcal{D}} \left[\frac{\rho_T^{\pi}(s,a)}{\mu(s,a)} \left(\mathbb{E}_{x' \sim T(s,a)}[g(x')] - g(s') \right) \right] \right|$$

- Computing $\rho_T^{\pi}(s,a)$: state-action distribution of running π on learned dynamics model T
  - rollout policy π on dynamics model T (doesn't require online interactions!)
  - estimate the distribution by normalizing flow
- Estimate μ using the dataset $\mathcal{D}$

Empirical evaluation: policy-selection on D4RL

- Dataset $\mathcal{D}$: 5 offline datasets from D4RL on halfcheetah and hopper
- Policy class Π and dynamics class $\mathcal{T}$:
  - Run a model-based offline RL algorithm (MOPO) with 5 different hyperparameters
  - Collect the learned policies and dynamics models

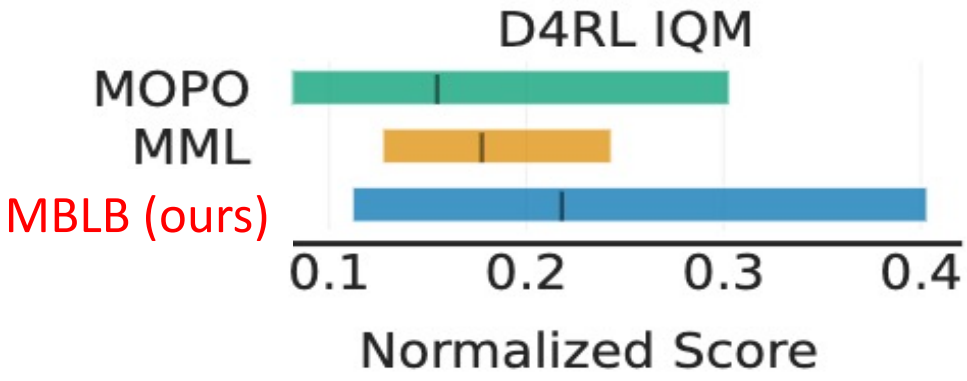




Figures from https://sites.google.com/view/d4rl-anonymous/

Empirical evaluation: policy-selection on D4RL

- We show the interquartile mean (IQM) of the learned policy [Agarwal et al., 2021]
  - i.e., discard the top & bottom 25% of the runs
- We compare with a model-based offline RL baseline (MOPO), and a prior work that requires global realizability (MML)



Summary

- Locally realizable models are enough for offline policy learning
- Algorithmic insight: optimize the policy and dynamics together
  - Focus on the state-actions visited by the current policy π

Future works

- Compute $\rho_T^{\pi}(s, a)$ more efficiently and differentiate through it
  - => Learn the policy/dynamics model from scratch

- Link to the full paper:

Thank you for your attention!

References

- Xie & Jiang, 2021, Batch Value-function Approximation with Only Realizability
- Xie et al., 2021, Bellman-consistent Pessimism for Offline Reinforcement Learning
- Uehara & Sun, 2021, Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage
- Voloshin et al., 2021, Minimax Model Learning
- Liu et al., 2020, Provably good batch off-policy reinforcement learning without great exploration
- Kakade and Langford, 2002, Approximately Optimal Approximate Reinforcement Learning
- Thomas et al., 2019, Preventing undesirable behavior of intelligent machines
- Agarwal et al., 2021, Deep reinforcement learning at the edge of the statistical precipice.