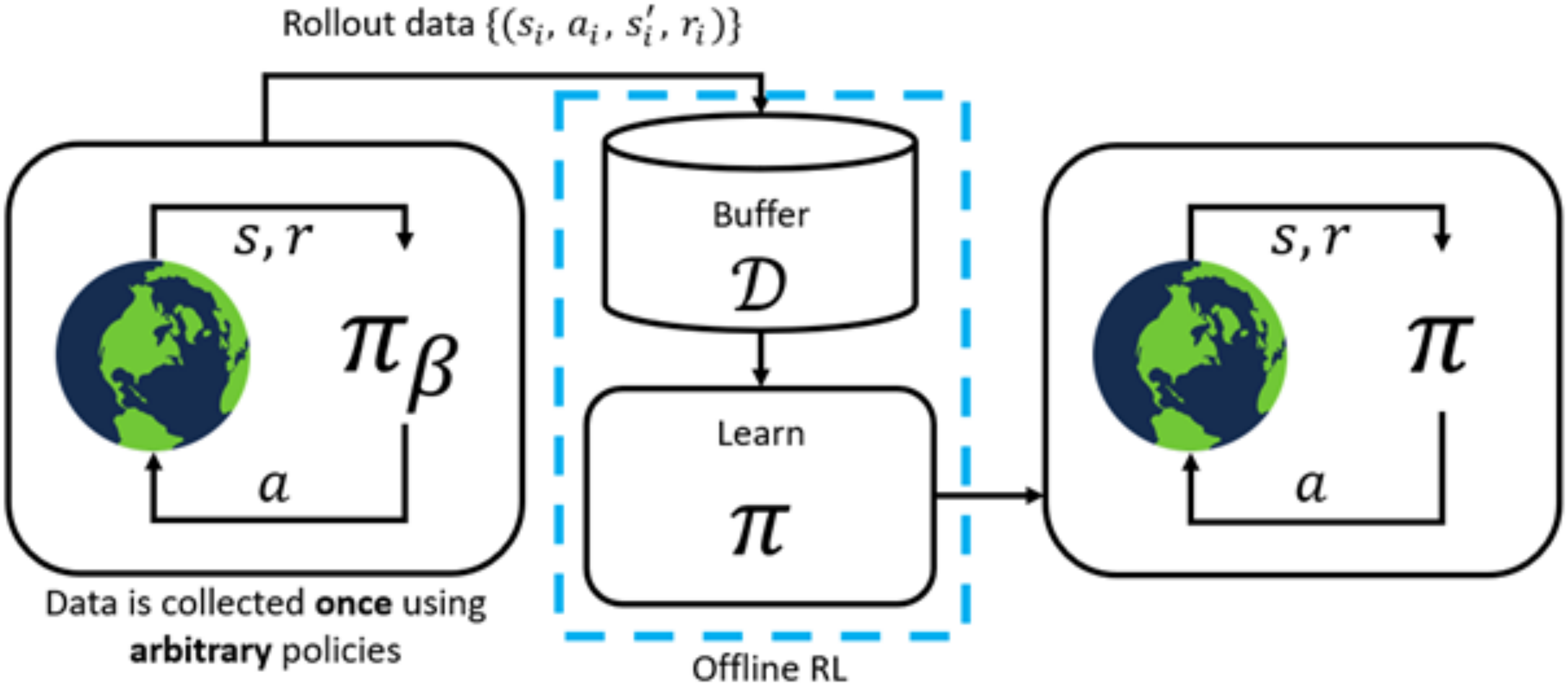
MOPO: Model-based Offline Policy Optimization

2022.06 Yang Chang

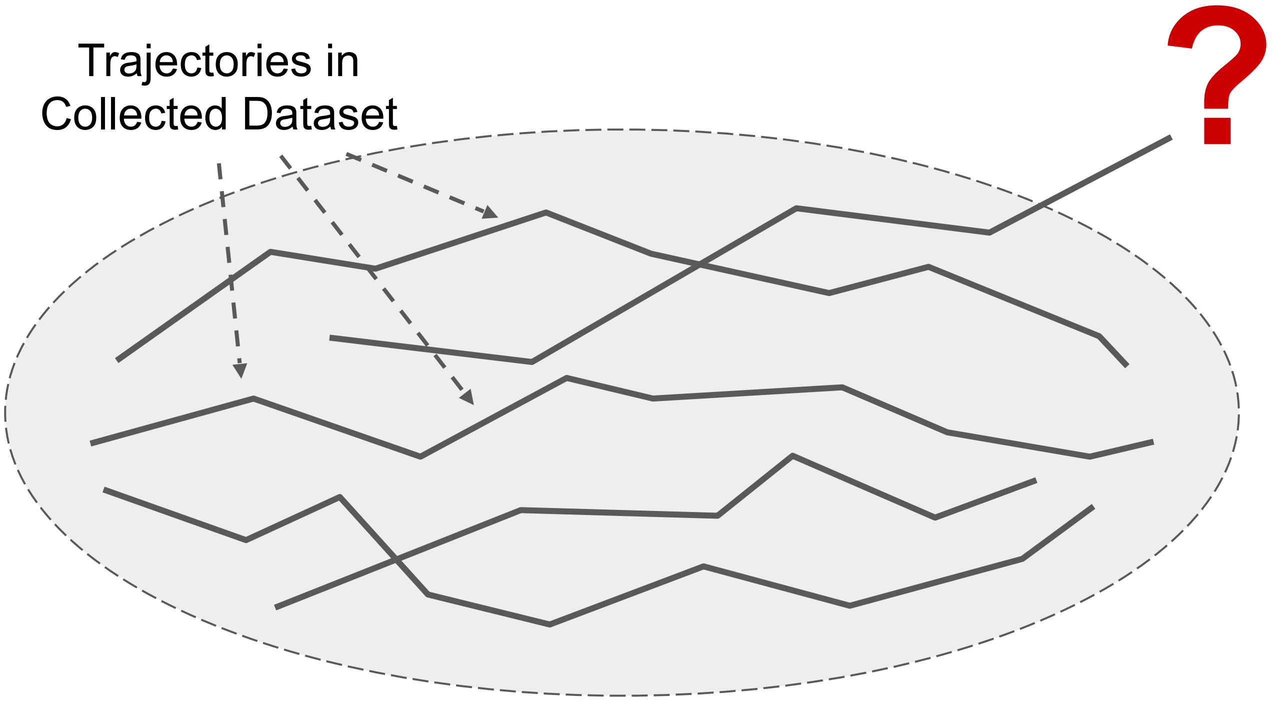
Outline

- Introduction
- Related Work
- MOPO
- Experiment
- Conclusion

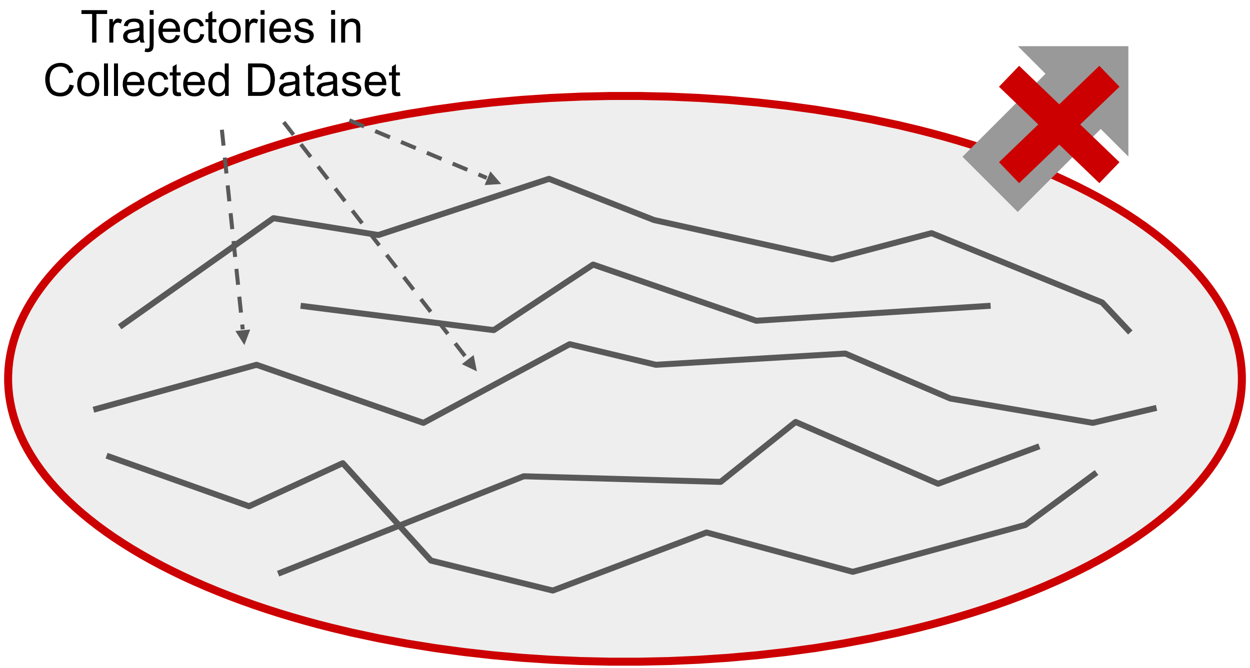
What is offline RL?



Challenging issues



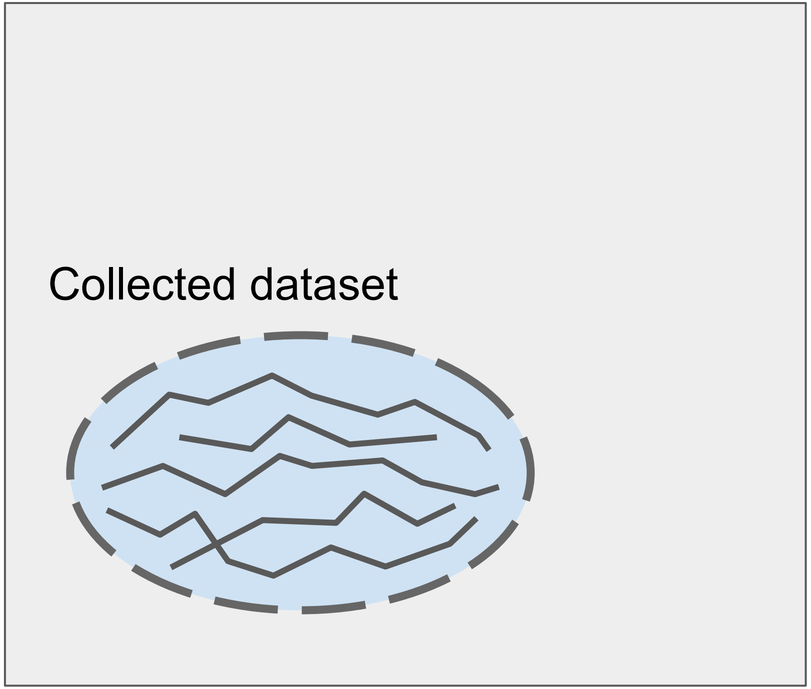
Previous offline RL methods



Can we go beyond the dataset?

- Idea: measure model uncertainty, and avoid (s, a) with high uncertainty.
- Uncertainty estimator $u(s, a)$

Real environment



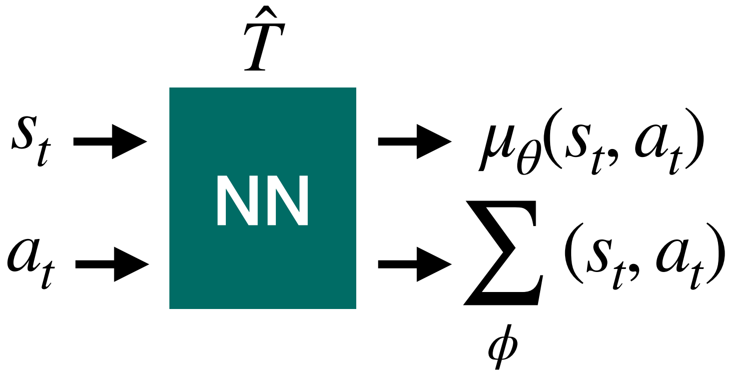
MOPO: Model-Based Offline Policy Optimization

- Real transition dynamic: $T(s, a)$
- Estimated transition dynamic from static dataset D_{env}: $\hat{T}(s, a)$
- Integral probability metric (IBM): d_F

- $u : S \times A$ is an admissible error estimator if $d_F(\hat{T}(s, a), T(s, a)) \leqq u(s, a)$

Practical Implementation

- Learn a model of **transition distribution** $\hat{T}(s' \mid s, a)$ from D_{env}
- Learn an **ensemble** of N dynamics models
- Take **maximum** standard deviation as uncertainty estimator
- $\tilde{r}(s, a) = \hat{r}(s, a) - \lambda \max_{i=1}^{N} \| \sum_{\phi}^{i} (s, a) \|_F$

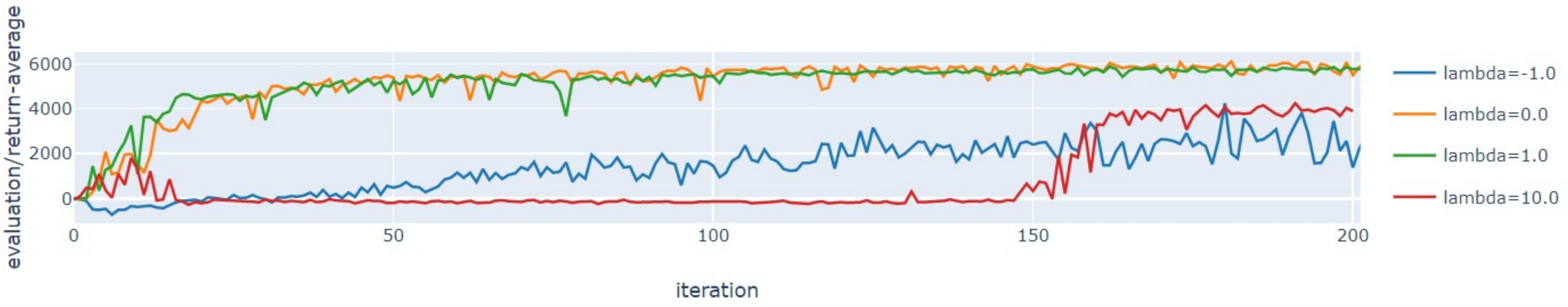


Ablation Study

- Different λ
- With & Without ensemble
- Pick different elements as uncertainty estimator
- Train the pre-trained agent with different algorithm

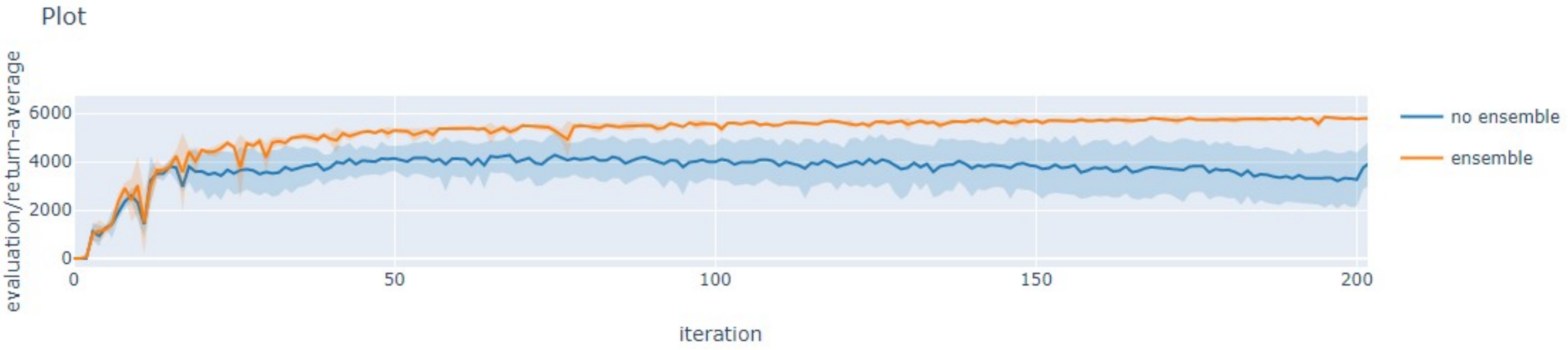
Ablation Study

- Different λ
  - $\tilde{r}(s,a) = \hat{r}(s,a) - \lambda \max_{i=1}^{N} \| \sum_{\phi}^{i} (s,a) \|_F$



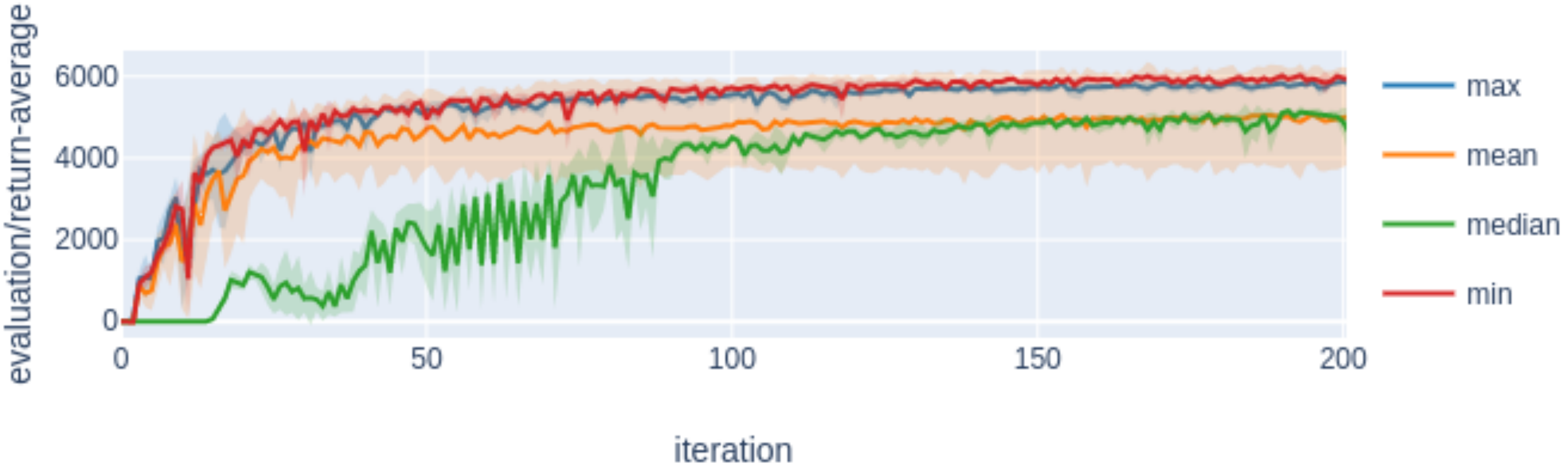
Ablation Study

- With & Without ensemble



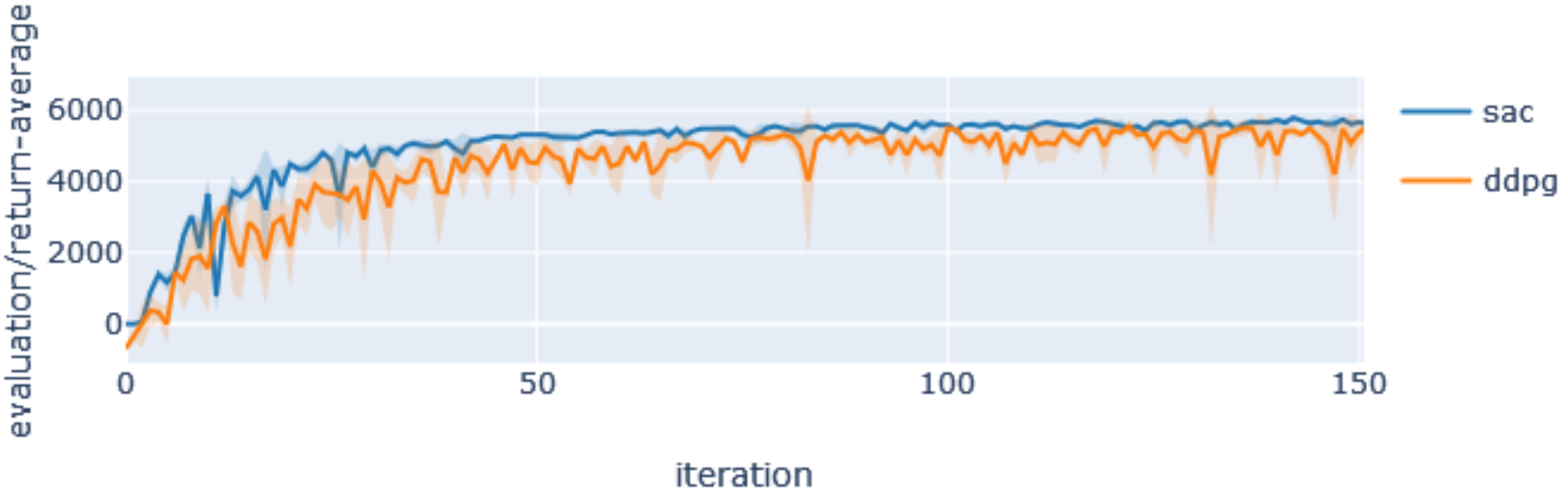
Ablation Study

- Pick different elements as uncertainty estimator



Ablation Study

- Train the pre-train agent with different algorithm



Conclusion

- $\lambda \in (0,1)$
- Ensemble is important
- Choose minimum std as uncertainty estimator is slightly better than maximum
- For the pre-trained agent, SAC is better than DDPG

Q & A