When to Trust Your Model:
Model-Based Policy Optimization

Michael Janner    Justin Fu    Marvin Zhang    Sergey Levine
UC Berkeley
The problem of model bias

- Model-based reinforcement learning methods are fast!

- … but often have poor asymptotic performance on high-dimensional tasks due to modeling inaccuracies

- How can we best leverage a predictive model for learning a policy?

PETS. Chua et al, 2018  
Algorithm 1 Model-Based Policy Optimization

1: Initialize data-collecting policy $\pi$
2: for $N$ epochs do
3: Collect data with $\pi$ in real environment
4: Train model $p_\theta$ on real data
5: Optimize policy $\pi$ under predictive model
Algorithm 1 Model-Based Policy Optimization

1: Initialize data-collecting policy $\pi$
2: for $N$ epochs do
3: Collect data with $\pi$ in real environment
4: Train model $p_\theta$ on real data
5: Optimize policy $\pi$ under predictive model

Little guidance on how to properly optimize a policy with a model
The problem with long rollouts

Environment rollout

Model rollout (x1000)

accurate, low variance

low accuracy, high variance

mean prediction +/- std dev
Bounding returns

- Use policy performance under model rollouts to derive lower bound in real environment

\[ \eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max} \left[ \frac{\gamma^{k+1}\epsilon_\pi}{(1 - \gamma)^2} + \frac{(\gamma^k + 2)\epsilon_\pi}{(1 - \gamma)} + \frac{k}{1 - \gamma}(\epsilon_m + 2\epsilon_\pi) \right] \]

- discrepancy due to **policy shift**
- discrepancy due to **model error**

See also Luo et al, ICLR 2019 and Sun et al, NeurIPS 2018.
Bounding returns

- Use policy performance under branched length-$k$ model rollouts to derive lower bound in real environment

$$\eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max} \left[ \frac{\gamma^{k+1} \epsilon_{\pi}}{(1 - \gamma)^2} + \frac{(\gamma^{k} + 2) \epsilon_{\pi}}{(1 - \gamma)} + \frac{k}{1 - \gamma} (\epsilon_m + 2 \epsilon_{\pi}) \right]$$

- discrepancy due to policy shift
- discrepancy due to model error

See also Luo et al, ICLR 2019 and Sun et al, NeurIPS 2018.
Bounding returns

- Use policy performance under branched length-\(k\) model rollouts to derive lower bound in real environment

\[ \eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max} \left[ \frac{\gamma^{k+1}\epsilon_{\pi}}{(1 - \gamma)^2} + \frac{(\gamma^{k} + 2)\epsilon_{\pi}}{(1 - \gamma)} + \frac{k}{1 - \gamma}(\epsilon_m + 2\epsilon_{\pi}) \right] \]

  discrepancy due to policy shift

  discrepancy due to model error

- Real off-policy data is always preferable to model-generated on-policy data!

  \( \rightarrow k = 0? \)

See also Luo et al, ICLR 2019 and Sun et al, NeurIPS 2018.
Model generalization in practice

Model error on data-collecting policy $\pi_D$

model error ($\epsilon_m$)

policy shift ($\text{KL}(\pi || \pi_D)$)
Model generalization in practice

Model error on different policy
Model generalization in practice
Model generalization in practice
Model generalization in practice
Model generalization in practice

![Graph showing model error (ε_m) against policy shift (KL(π||π_D)) and train size.](image)
Model generalization in practice
(Better) bounds on returns

- Use policy performance under model rollouts to derive lower bound in real environment

\[
\eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max}\left[\frac{\gamma^{k+1}\epsilon_{\pi}}{(1 - \gamma)^2} + \frac{(\gamma^k + 2)\epsilon_{\pi}}{(1 - \gamma)} + \frac{k}{1 - \gamma}(\epsilon_m + 2\epsilon_{\pi})\right]
\]
(Better) bounds on returns

- Use policy performance under model rollouts to derive lower bound in real environment:

\[
\eta[\pi] \geq \eta^{\text{branch}}[\pi] - 2r_{\max} \left[ \frac{\gamma^{k+1}\epsilon_\pi}{(1-\gamma)^2} + \frac{(\gamma^k + 2)\epsilon_\pi}{1-\gamma} + \frac{k}{1-\gamma}(\epsilon_m + 2\epsilon_\pi) \right]
\]

- Accounting for empirical model generalization motivates model usage

- If \( \epsilon_{\pi} \approx \epsilon_m \), \( k \) will be small*

*When \( k=1 \), this corresponds to Dyna [Sutton, 1991]
MBPO Results

- **InvertedPendulum**
- **Hopper**
- **Walker2d**
- **Ant**
- **HalfCheetah**
- **Humanoid**

Legend:
- **PETS**
- **STEVE**
- **SLBO**
- **--- convergence**
MBPO Results

InvertedPendulum

Hopper

Walker2d

Ant

HalfCheetah

Humanoid

average return

average return

average return

average return

average return

average return

steps

steps

steps

steps

steps

steps

0 5k 10k 15k

0 50k 100k

0 100k 200k 300k

0 100k 200k 300k

0 100k 200k 300k

0 100k 200k 300k

0 500 1000 1500 2000 2500

0 500 1000 1500 2000 2500

0 500 1000 1500 2000 2500

0 500 1000 1500 2000 2500

0 500 1000 1500 2000 2500

0 500 1000 1500 2000 2500

SAC  PPO  PETS  STEVE  SLBO  convergence
MBPO Results

- InvertedPendulum
- Hopper
- Walker2d
- Ant
- HalfCheetah
- Humanoid

**Graphs:**
- x-axis: steps (ranging from 0 to 300k)
- y-axis: average return

**Lines:**
- MBPO
- SAC
- PPO
- PETS
- STEVE
- SLBO
- convergence
MBPO Results

added benefit: no model exploitation
Summary

- Empirical model generalization motivates model usage
- Short model rollouts give large benefits to policy optimization
- MBPO avoids model exploitation and scales to long-horizon tasks

people.eecs.berkeley.edu/~janner/mbpo/