ETH zürich

Model-based and Model-free DRL

Yu Hong
MSc D-INFK
30 Mar 2021, DNN Seminar

D INFK

Outline

1. Model based vs. model free DRL

2. Guarantee and usage for Model based method

3. Successor features

Outline

1. Model based vs. model free DRL

2. Guarantee and usage for Model based method

3. Successor features

How to get more ‘likes’?

Model-based

ETH zürich

Department of Humanities, Social and Political Sciences

News & Events | The Department | Research | Studies | Doctoral Studies | Continuing Education | Information & Documents

Studies overview

Science in Perspective
Teaching Committee Science in Perspective
SiP FAQs

BA in Public Policy
Teaching Commission BA Public Policy

MA in History and Philosophy of Knowledge
Teaching Committee MA HPK

MA in Comparative and International Studies
Teaching Committee MA CIS

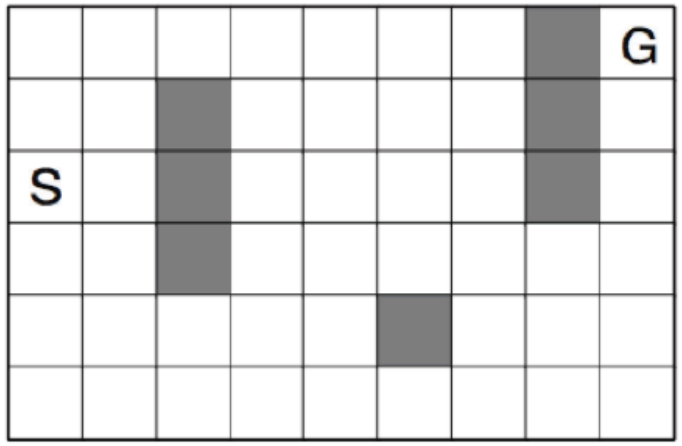
MA in Science Didactics

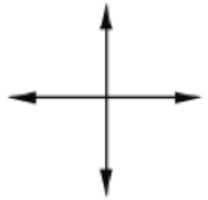
MSc in Science, Technology and Policy

Teacher Training

Model-free

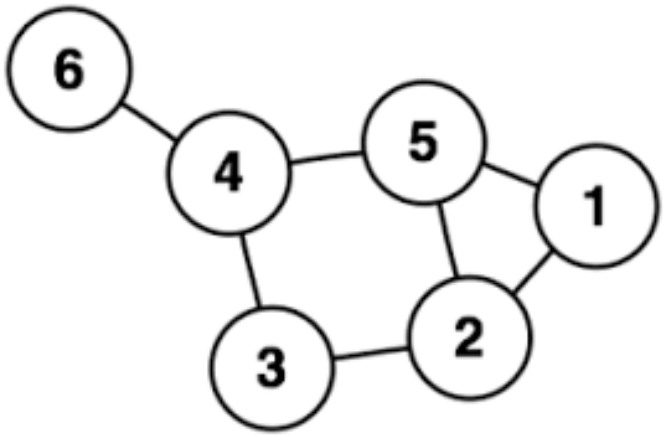
Simple maze





actions

Model-based



Model-free

	U	D	L	R
1[	0	0	0	0
2[	0	0	0	0
3[	0	0	0	0
4[	0	0	0	0

Start State - 1	State - 2
Snake State - 3	Treasure State - 4

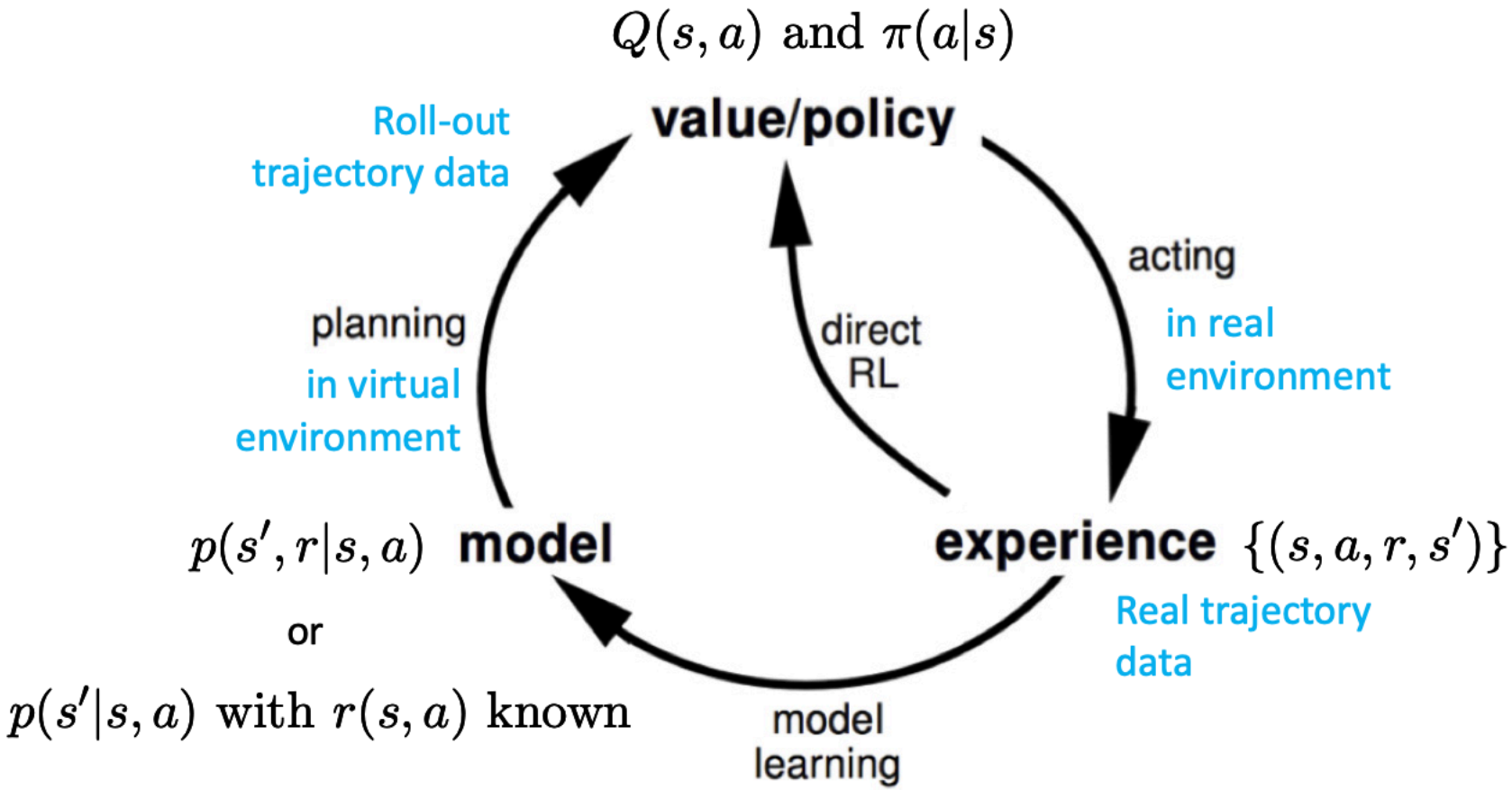
Model-based and Model-free DRL

	Model-based	Model-free
Model	supervised learning	No model
Environment	Less interaction	Interaction
Value function	Based on model	Based on samples

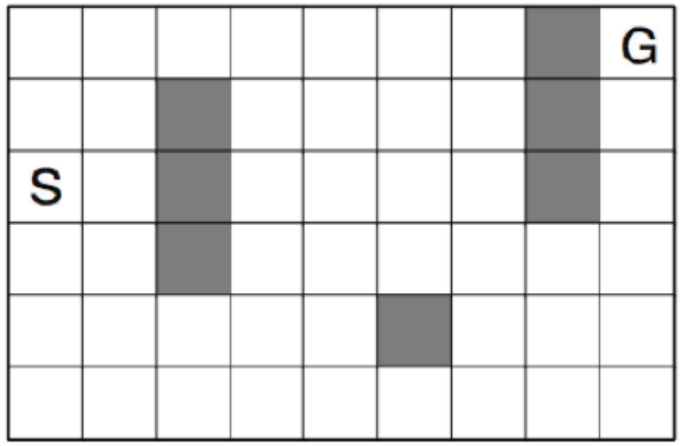
Model-based and Model-free DRL

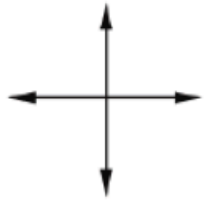
	Model-based	Model-free
Model	On policy	On-policy/Off-policy
Environment	Less interaction	Interaction
Value function	Compounded error	Best asymptotic performance

Dyna-Q


$Q(s,a)$ and $\pi(a|s)$
value/policy
Roll-out
trajectory data
acting
in real
environment
planning
in virtual
environment
direct
RL
$p(s',r|s,a)$ model
experience $\{(s,a,r,s')\}$
or
$p(s'|s,a)$ with $r(s,a)$ known
Real trajectory
data
model
learning

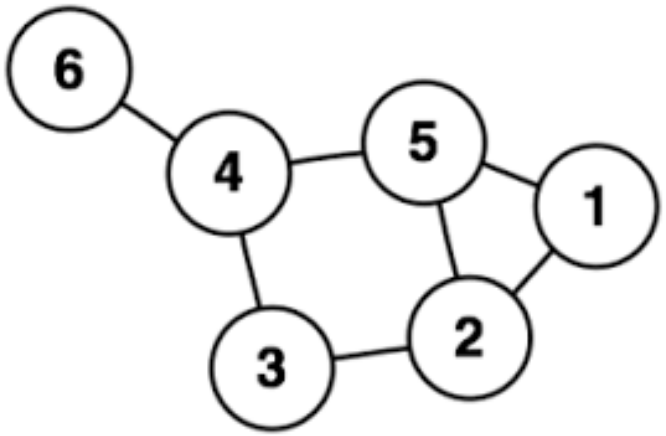
Simple maze



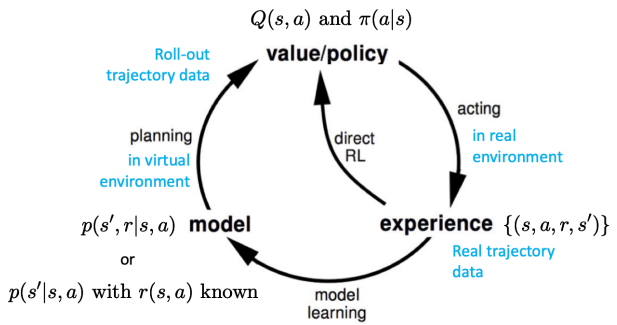


actions

Model-based



Combination

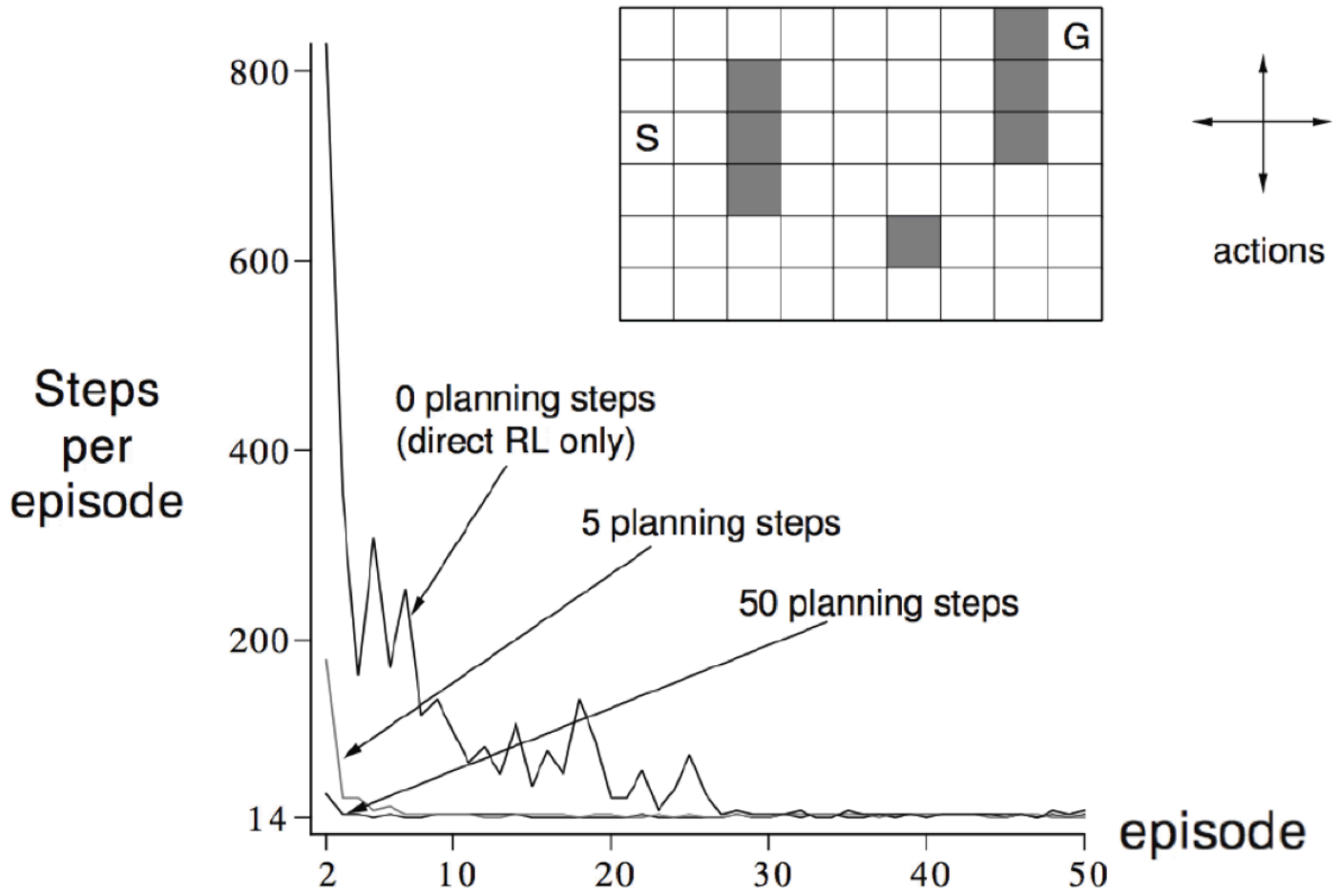


Model-free

	U	D	L	R
1[	0	0	0	0
2[	0	0	0	0
3[	0	0	0	0
4[	0	0	0	0

Start State - 1	State - 2
Snake State - 3	Treasure State - 4

Simple maze


G
S
actions
Steps per episode
800
600
400
200
14
0 planning steps
(direct RL only)
5 planning steps
50 planning steps
2
10
20
30
40
50
episode

Outline

1. Model based vs. model free DRL

2. Guarantee and usage for Model based method

3. Successor features

Monotonic improvement guarantee

$$V^{\pi, M^\star} \geq V^{\pi, \widehat{M}} - D(\widehat{M}, \pi)$$

assumption: (R_1) $\forall \pi$, $d(\pi, \pi_D) \leq \delta$

(R_2) $\hat{M} = M^* \Rightarrow D(\hat{M}, \pi) = 0$

(R_3) $D(\hat{M}, \pi)$ has form $\mathbb{E}_{\tau \sim \pi_D, M^*}[f(\hat{M}, \pi, \tau)]$

Monotonic improvement guarantee

Algorithm 1 Meta-Algorithm for Model-based RL

Inputs: Initial policy π_0. Discrepancy bound D and distance function d that satisfy equation (R1) and (R2).

For $k = 0$ to T:

$$\pi_{k+1}, M_{k+1} = \underset{\pi \in \Pi,\ M \in \mathcal{M}}{\text{argmax}} \quad V^{\pi, M} - D_{\pi_k, \delta}(M, \pi) \tag{3.3}$$

$$\text{s.t.} \quad d(\pi, \pi_k) \leq \delta \tag{3.4}$$

Monotonic improvement guarantee

Theorem 3.1. *Suppose that $M^\star \in \mathcal{M}$, that D and d satisfy equation (R1) and (R2), and the optimization problem in equation (3.3) is solvable at each iteration. Then, Algorithm 1 produces a sequence of policies $\pi_0, \ldots, \pi_T$ with monotonically increasing values:*

$$V^{\pi_0, M^\star} \leq V^{\pi_1, M^\star} \leq \cdots \leq V^{\pi_T, M^\star} \tag{3.5}$$

Moreover, as $k \to \infty$, the value $V^{\pi_k, M^\star}$ converges to some $V^{\bar{\pi}, M^\star}$, where $\bar{\pi}$ is a local maximum of $V^{\pi, M^\star}$ in domain Π.

assumption: (R1) $\forall \pi$, $d(\pi, \pi_D) \leq \delta$

(R2) $\hat{M} = M^* \Rightarrow D(\hat{M}, \pi) = 0$

(R3) $D(\hat{M}, \pi)$ has form $\mathbb{E}_{\tau \sim \pi_0, M^*}[f(\hat{M}, \pi, \tau)]$

$$\pi_{k+1}, M_{k+1} = \underset{\pi \in \Pi,\ M \in \mathcal{M}}{\operatorname{argmax}} \quad V^{\pi, M} - D_{\pi_k, \delta}(M, \pi)$$
$$\text{s.t. } d(\pi, \pi_k) \leq \delta$$

Proof: $\forall k$,

$V^{\pi_{k+1}, M^*} \overset{(R_1)}{\geq} V^{\pi_{k+1}, M_{k+1}} - D(M_{k+1}, \pi_{k+1})$

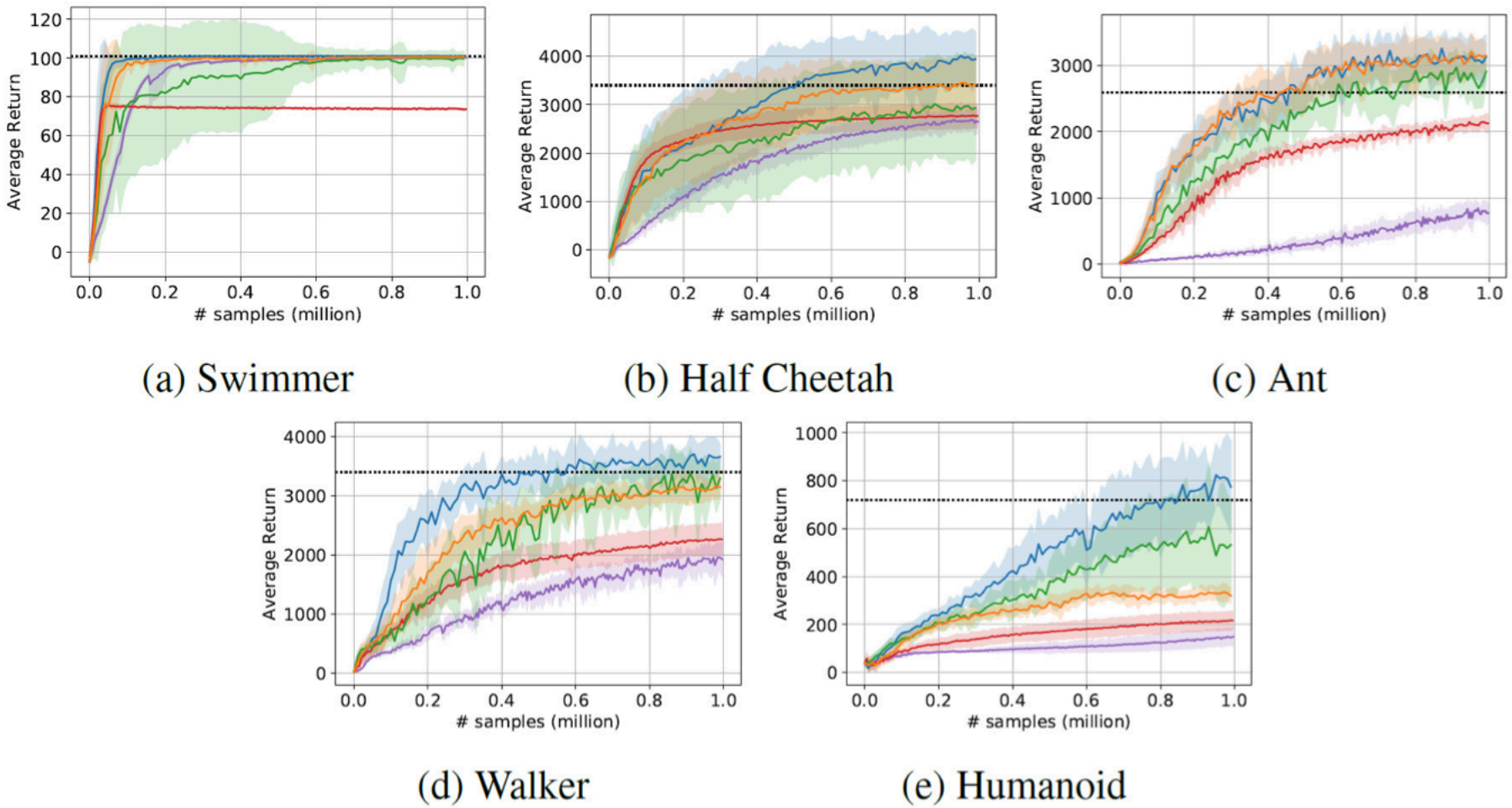
meta algorithm $\Rightarrow$ $\geq V^{\pi_k, M^*} - D(M^*, \pi_k) \overset{(R_2)}{=} 0$

$= V^{\pi_k, M^*}$

Algorithmic Framework for Model-based Deep Reinforcement Learning with Theoretical Guarantees

Meta algorithm-stochastic lower bound optimisation

All envs with maximum horizon of 500



How to perform rollout?

policy shift $\epsilon_\pi = \max_s D_{TV}(\pi \| \pi_D)$

model shift $\epsilon_m = \max_t \mathbb{E}_{s \sim \pi_{D,t}} \left[D_{TV}(p(s', r | s, a) \| p_\theta(s', r | s, a)) \right]$

$$V^{\pi, M^*} \geq V^{\pi, \hat{M}} - \left[\frac{2\gamma r_{max}(\epsilon_m + 2\epsilon_\pi)}{(1-\gamma)^2} + \frac{4 r_{max} \epsilon_\pi}{(1-\gamma)} \right]$$

When to Trust Your Model: Model-Based Policy Optimization

How to perform rollout?

Theorem 4.2. $V^{\pi_k, M^*} \geq V^{\pi_k, \hat{M}} - 2 r_{max} \left[\frac{\gamma^{k+1} \epsilon_\pi}{1-\gamma} + \frac{\gamma^k + 2}{1-\gamma} \epsilon_\pi + \frac{k}{1-\gamma} (\epsilon_m + 2\epsilon_\pi) \right]$

$\Rightarrow$ optimal $k = 0$!!!

Do not use the model !

When to Trust Your Model: Model-Based Policy Optimization

How to perform rollout?

model shift $\epsilon_m = \max_t \mathbb{E}_{s \sim \pi_{D,t}} \left[D_{TV}(p(s', r \mid s, a) \,\|\, p_\theta(s', r \mid s, a)) \right]$

a new model error (first-order) approximation

$$\epsilon_{m'} = \max_t \mathbb{E}_{s \sim \pi_t} \left[D_{TV}(p(s', r \mid s, a) \,\|\, p_\theta(s', r \mid s, a)) \right]$$

approximation: $\hat{\epsilon}_{m'}(\epsilon_\pi) \simeq \epsilon_m + \epsilon_\pi \frac{d\epsilon_{m'}}{d\epsilon_\pi}$

When to Trust Your Model: Model-Based Policy Optimization

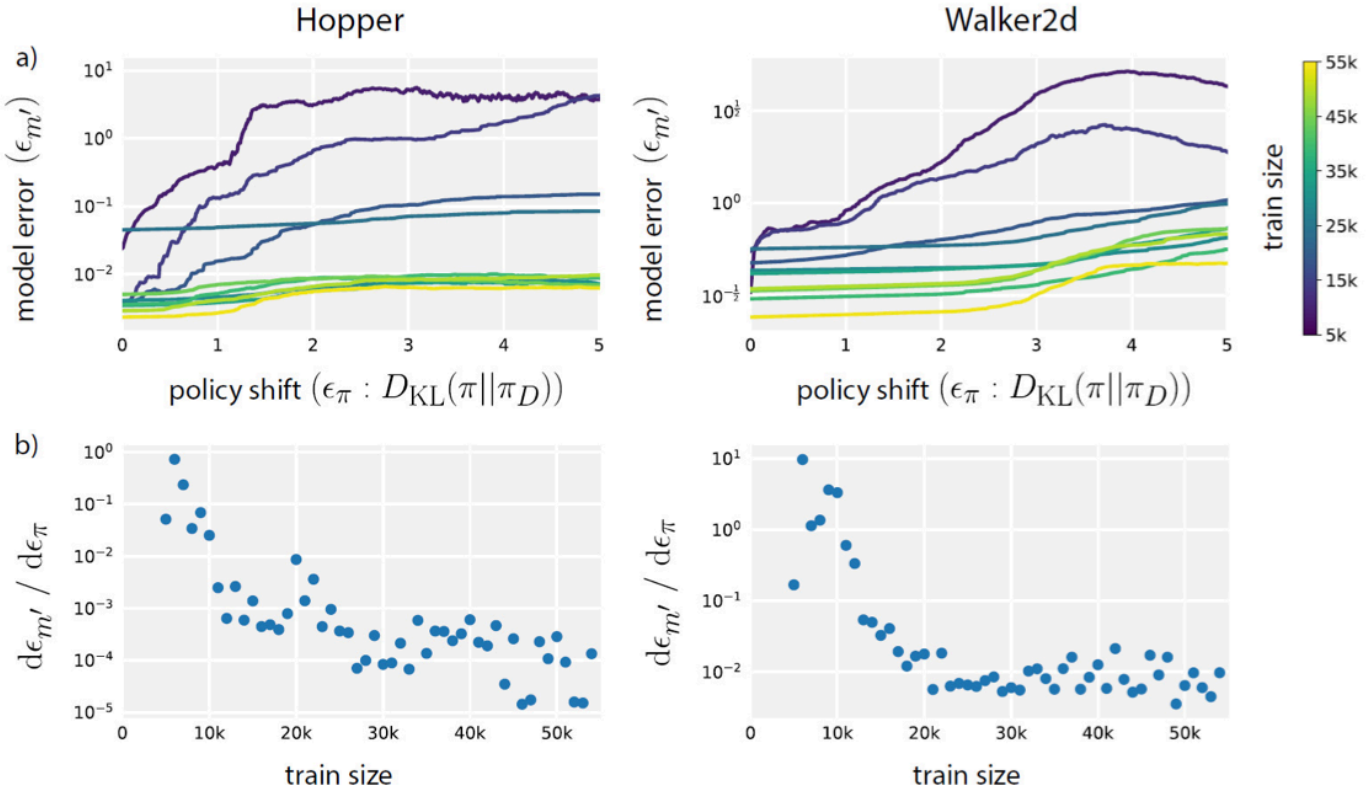
How to perform rollout?

The bound is rewritten as $V^{\pi_K, M^*} \geq V^{\pi_K, \hat{M}} - 2 r_{max} \left[\frac{\gamma^{k+1} \epsilon_\pi}{(1-\gamma)^2} + \frac{\gamma^k \epsilon_\pi}{1-\gamma} + \frac{k}{1-\gamma} \epsilon_m \right]$

$\Rightarrow$ if $\frac{d\epsilon_m}{d\epsilon_\pi}$ small enough, $k > 0$

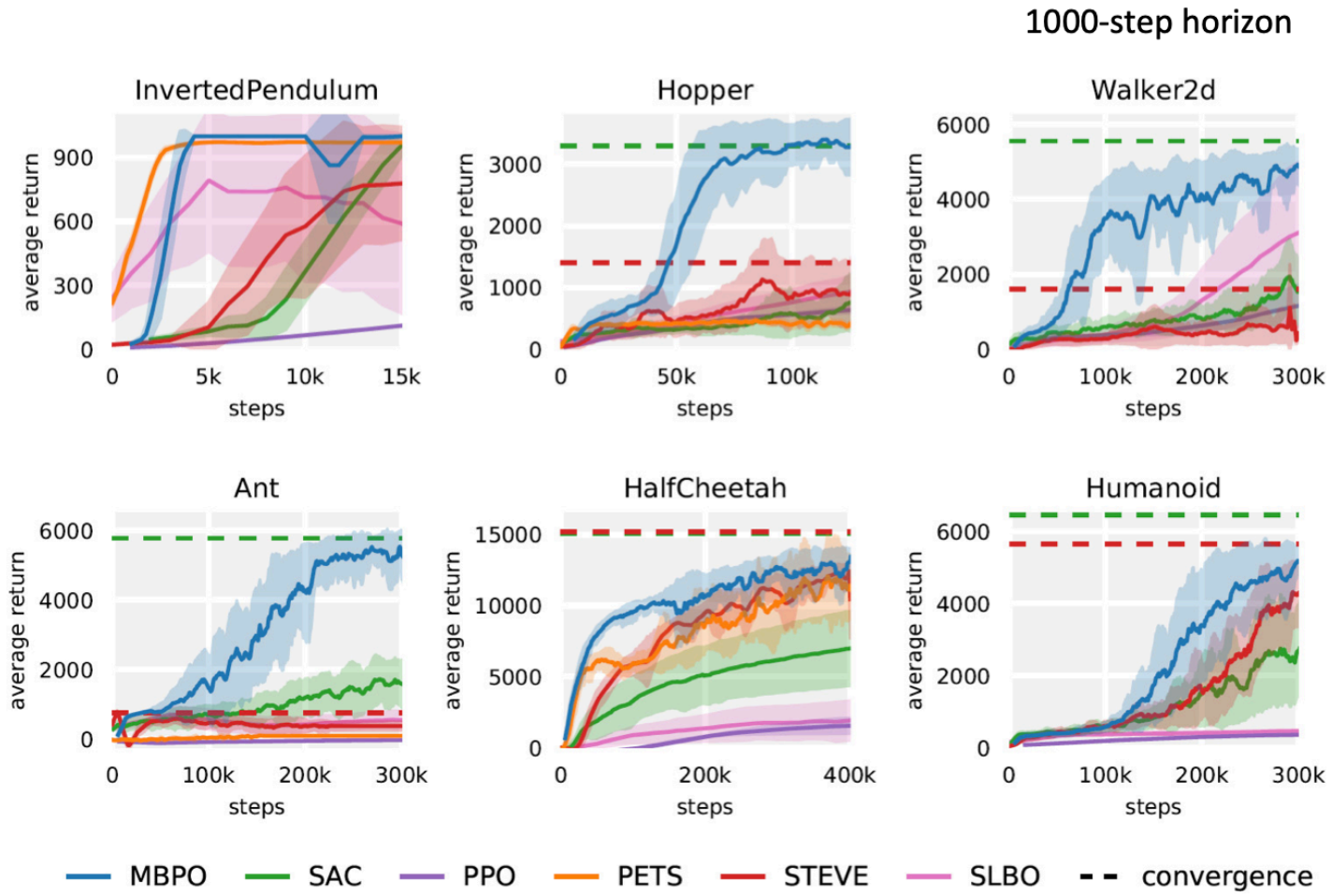
When to Trust Your Model: Model-Based Policy Optimization

Empirical analysis of $\frac{\mathrm{d}\epsilon_{m'}}{\mathrm{d}\epsilon_{\pi}}$



When to Trust Your Model: Model-Based Policy Optimization

Model-based policy optimisation



When to Trust Your Model: Model-Based Policy Optimization

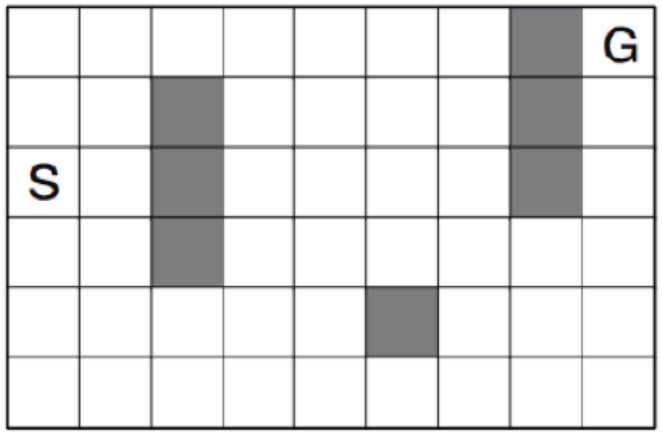
Outline

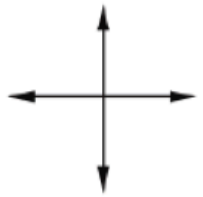
1. Model based vs. model free DRL

2. Guarantee and usage for Model based method

3. Successor features

Successor representation





actions

$$M^{\pi}(s_i, s_j) = \mathbb{E}_{\pi}[\sum_{k=0}^{\infty} \gamma^k \mathbb{I}[s_{t+k} = s_j] \mid s_t = s_i]$$

$$V^{\pi}(s_i) = \sum M^{\pi}(s_i, s_j) R(s_j)$$

assumption: $R(s_j) = \phi(s_j) \cdot w$

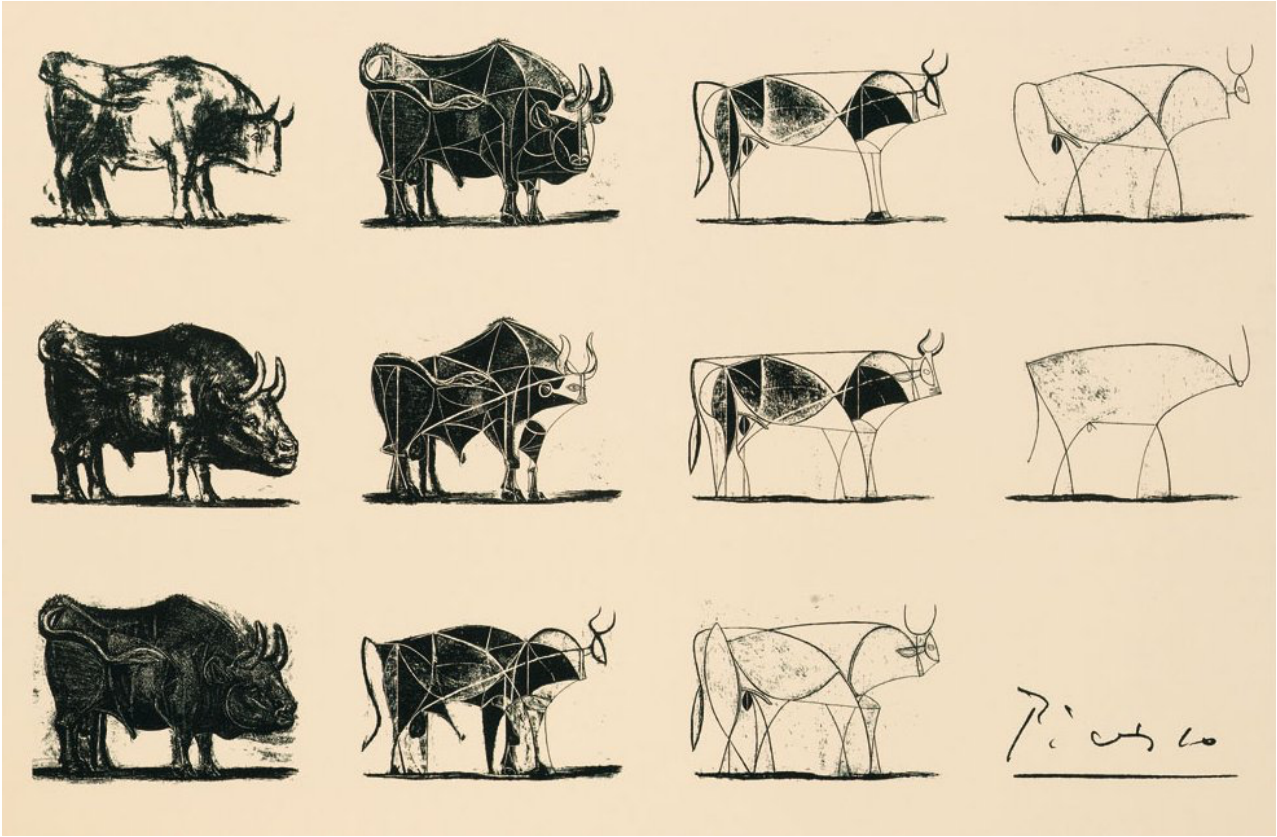
$$= \sum_j M^{\pi}(s_i, s_j)\, \phi(s_j) w = \left[\sum_j M^{\pi}(s_i, s_j)\, \phi(s_j)\right] \cdot w$$

$\psi(s_i)$

summary

Model based/ model free/ successor features

Justify model usage and how to use model



“I never made a painting as a work of art. It's all research.” – Pablo Picasso

ETH zürich

Thanks

Dayan, Peter. "Improving generalization for temporal difference learning: The successor representation." Neural Computation 5.4 (1993): 613-624.

Luo, Yuping, et al. "Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees." *arXiv preprint arXiv:1807.03858* (2018).

Barreto, André, et al. "Successor features for transfer in reinforcement learning." arXiv preprint arXiv:1606.05312 (2016).

Barreto, Andre, et al. "Transfer in deep reinforcement learning using successor features and generalised policy improvement." International Conference on Machine Learning. PMLR, 2018.

Janner, Michael, et al. "When to trust your model: Model-based policy optimization." arXiv preprint arXiv:1906.08253 (2019).