强化学习2022
第12节

涉及知识点：
参数化行动空间

参数化行动空间

张伟楠 – 上海交通大学

课程大纲

强化学习基础部分

1. 强化学习、探索与利用
2. MDP和动态规划
3. 值函数估计
4. 无模型控制方法
5. 规划与学习
6. 参数化的值函数和策略
7. 深度强化学习价值方法
8. 深度强化学习策略方法

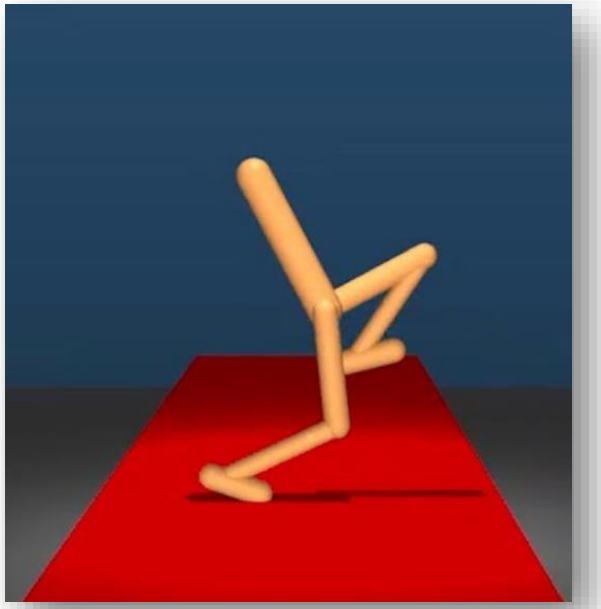
强化学习前沿部分

9. 基于模型的深度强化学习
10. 模仿学习
11. 离线强化学习
12. 参数化动作空间
13. 目标导向的强化学习
14. 多智能体强化学习
15. 强化学习大模型
16. 技术交流与回顾

马尔科夫决策过程（MDPs）中的动作空间

离散动作空间

$a \in \{a_1, a_2, \ldots, a_k\}$



连续动作空间

$a = (x_1, x_2, \ldots, x_d)$

Image credit: DeepMind & Nature

离散 vs. 连续

离散动作空间

$$a \in \{a_1, a_2, \ldots, a_k\}$$

- 智能体从有限动作集合中选择一个动作
- 分类问题
- 无法对已选择的动作进行微调

连续动作空间

$$a = (x_1, x_2, \ldots, x_d)$$

- 动作被表示为一个实值向量
- 回归问题
- 无法考虑所选动作之间的种类差异

参数化动作空间

- 每一个离散动作都可以被赋连续动作值（或者说参数，即参数化的动作）

{**冲刺** (方向, 速度), **转向** (方向), **踢球** (方向, 力度), **停止**}

参数化动作空间

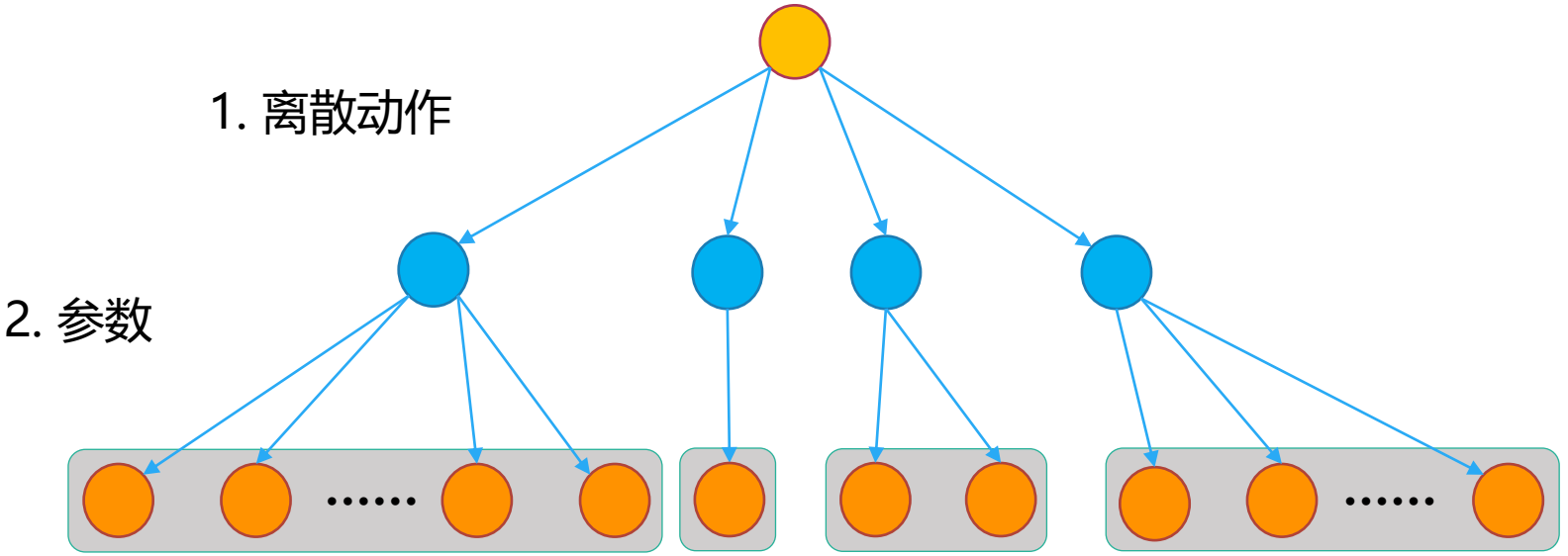
- 每一个离散动作都可以被赋连续动作值（或者说参数，即参数化的动作）

离散按键 + **连续**摇杆

参数化动作空间

- 一个参数化的动作空间是一个混合动作空间：包含离散的动作和连续的动作（参数）
- 智能体可以同时决定执行何种离散动作和使用什么样的连续参数来执行该动作
- 许多强化学习场景都有一个自然的参数化动作空间
  - MOBA/RTS游戏
  - 交易
  - 带时间参数的行为
- 参数化动作空间是结构化的，其结构往往非常关键

参数化动作空间的结构



- 动作选择是分层的
- 动作空间的结构是一种与马尔可夫决策过程相关的有用信息

参数化动作马尔可夫决策过程

- 一个参数化动作指的是一个用实值向量参数化的离散动作
- 通过这种将不同类型连续动作区别对待的方式对动作建模，将结构引入动作空间
- 在参数化动作马尔可夫决策过程（PAMDPs）模型的情况下
  - 有不同的离散动作需要合适的参数使其适应实际情景，或者有多个互斥的连续动作

$$\{a_1, a_2, (x_1, x_2, x_3), (y_1, y_2), (z_1, z_2, z_3)\}$$

Masson, Warwick, et al. Reinforcement Learning with Parameterized Actions. AAAI 2016.

使用参数化动作马尔可夫决策过程的强化学习算法

- **离散动作空间**：深度Q网络（DQN），双Q网络（Double-DQN），A3C，……
- **连续动作空间**：确定性策略梯度（DPG），深度确定性策略梯度（DDPG），……
- **参数化动作空间**：？

要求算法能够处理离散-连续混合动作空间。主要有三个途径：

- 离散化连续动作空间
- 将离散动作空间放宽到连续空间
- 分开处理离散动作和连续动作

离散化连续动作空间

- 离散-连续混合动作空间可以被表示为：

$$A = \{(k, x_k) | x_k \in X_k \text{ for all } k \in \{1 \dots K\}\}$$

- 我们可以用离散的子集来近似每一个X_k，然后使用深度Q网络（DQN）或A3C算法训练离散策略
- 但是，这种做法并不完美：
  - 可能会导致X_k自然结构的丢失
  - 当X_k是欧几里得空间内的区域时，构建一个良好的近似通常需要大量的离散动作

Xiong, Jiechao, et al. Parametrized Deep Q-Networks Learning: Reinforcement Learning with Discrete-Continuous Hybrid Action Space. 2017.

放宽离散动作空间

- 原始的离散-连续混合动作空间：

$$A = \{(k, x_k) | x_k \in X_k \text{ for all } k \in \{1 \dots K\}\}$$

通过把离散动作空间放宽到连续空间，策略输出变为：

$$(f_1, f_2, \dots, f_K, x_1, x_2, \dots, x_K)$$

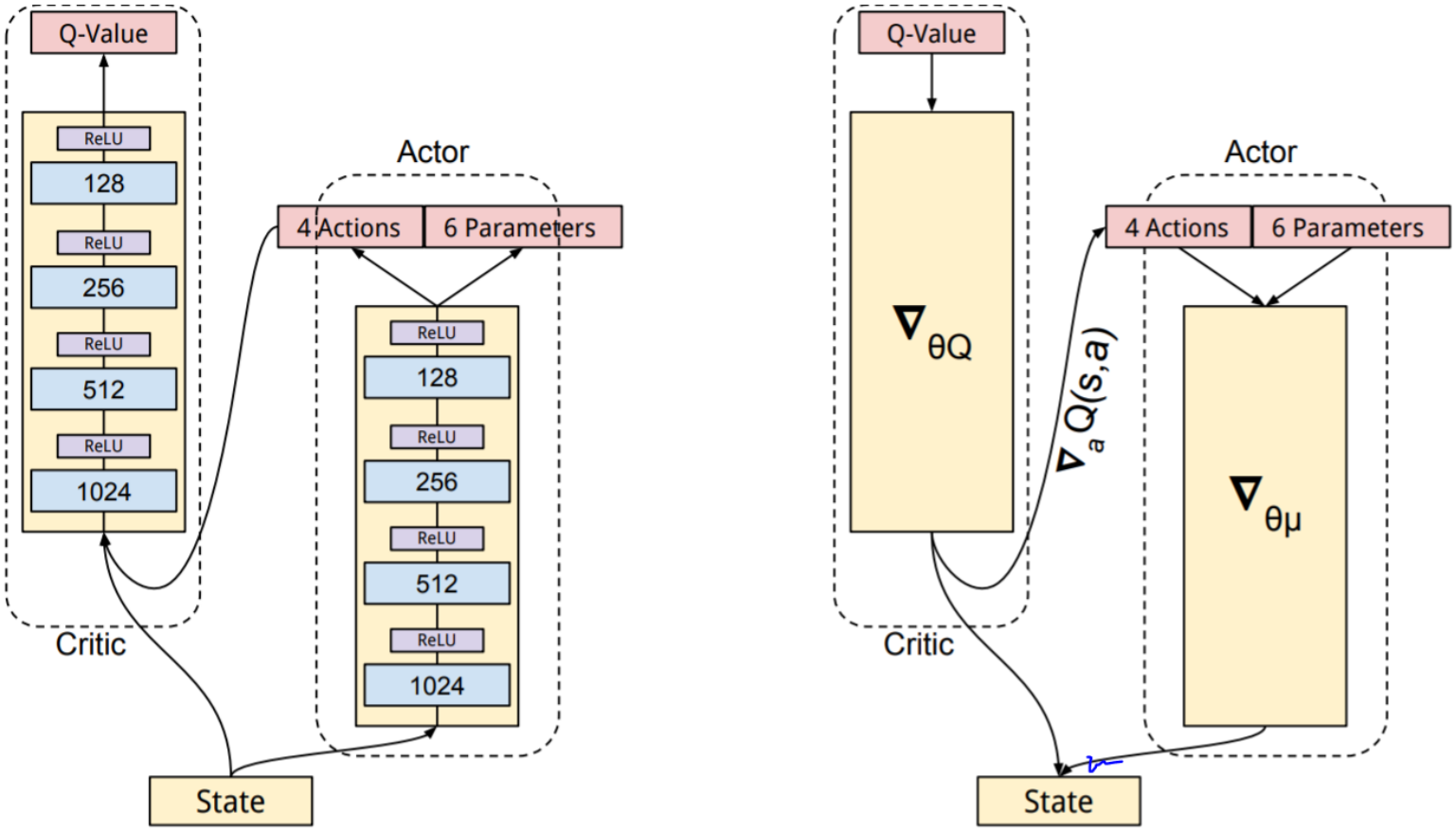
在这里$(f_1, f_2, \dots, f_K)$用于选择离散动作，选择可以是

- 确定性的（通过选取$\arg\max_i(f_i)$），或
- 随机性的（通过概率$\text{soft max}(f_i)$）

从而可以使用深度确定性策略梯度（DDPG）算法训练连续策略

Hausknecht and Stone. Deep reinforcement learning in parameterized action space. ICLR 2016.

深度确定性策略梯度（DDPG）模型



注：这里的关键部分是DDPG中的Q值网络（评论家），它平滑处理了离散的arg max 操作

Hausknecht and Stone. Deep reinforcement learning in parameterized action space. ICLR 2016.

Q-PAMDP算法

- Q-PAMDP算法交替学习动作选择和参数选择策略
- 离散动作策略记作$\pi^d(a|s)$。为了选择动作的参数，我们为每个动作a定义动作参数策略，记作$\pi^a(x|s)$。则输出策略由下式给出

$$\pi(a,x|s) = \pi^d(a|s)\pi^a(x|s)$$

  - 离散动作策略和动作参数策略的参数为：

$$\pi_w^d(a|s) \qquad w = [w_1, w_2, \dots, w_l]$$

$$\pi_\theta^a(x|s) \qquad \theta = [\theta_{a_1}, \theta_{a_2}, \dots, \theta_{a_k}]$$

Masson and Ranchod. Reinforcement Learning with Parameterized Actions. AAAI 2016.

Q-PAMDP算法

- 给定策略从某一初始状态开始的期望累计奖励作为优化的目标函数：

$$J(\theta, w) = \mathbb{E}_{s_0 \sim D}[V^{\pi_\Theta}(s_0)]$$

- 对于固定的θ，令

$$W(\theta) = \arg\max_w J(\theta, w) = w_\theta^*$$

- 我们可以为每个固定的θ使用Q-学习算法学习$W(\theta)$。最后，对于固定的w，我们定义，

$$J_w(\theta) = J(\theta, w)$$

$$H(\theta) = J(\theta, W(\theta))$$

$H(\theta)$是对于固定θ的最佳离散策略表现

Masson and Ranchod. Reinforcement Learning with Parameterized Actions. AAAI 2016.

Q-PAMDP算法

Algorithm 1 Q-PAMDP(k)

Input:
Initial parameters θ_0, ω_0
Parameter update method P-UPDATE
Q-learning algorithm Q-LEARN
Algorithm:
$\omega \leftarrow \text{Q-LEARN}^{(\infty)}(M_\theta, \omega_0)$
repeat
  $\theta \leftarrow \text{P-UPDATE}^{(k)}(J_\omega, \theta)$
  $\omega \leftarrow \text{Q-LEARN}^{(\infty)}(M_\theta, \omega)$
until θ converges

$\pi_w^d(a|s)$ $\quad w = [w_1, w_2, \dots, w_l]$

$\pi_\theta^a(x|s)$ $\quad \theta = [\theta_{a_1}, \theta_{a_2}, \dots, \theta_{a_k}]$

P-UPDATE(f, θ) 为一个针对目标函数f优化θ的策略搜索方法

Masson and Ranchod. Reinforcement Learning with Parameterized Actions. AAAI 2016.

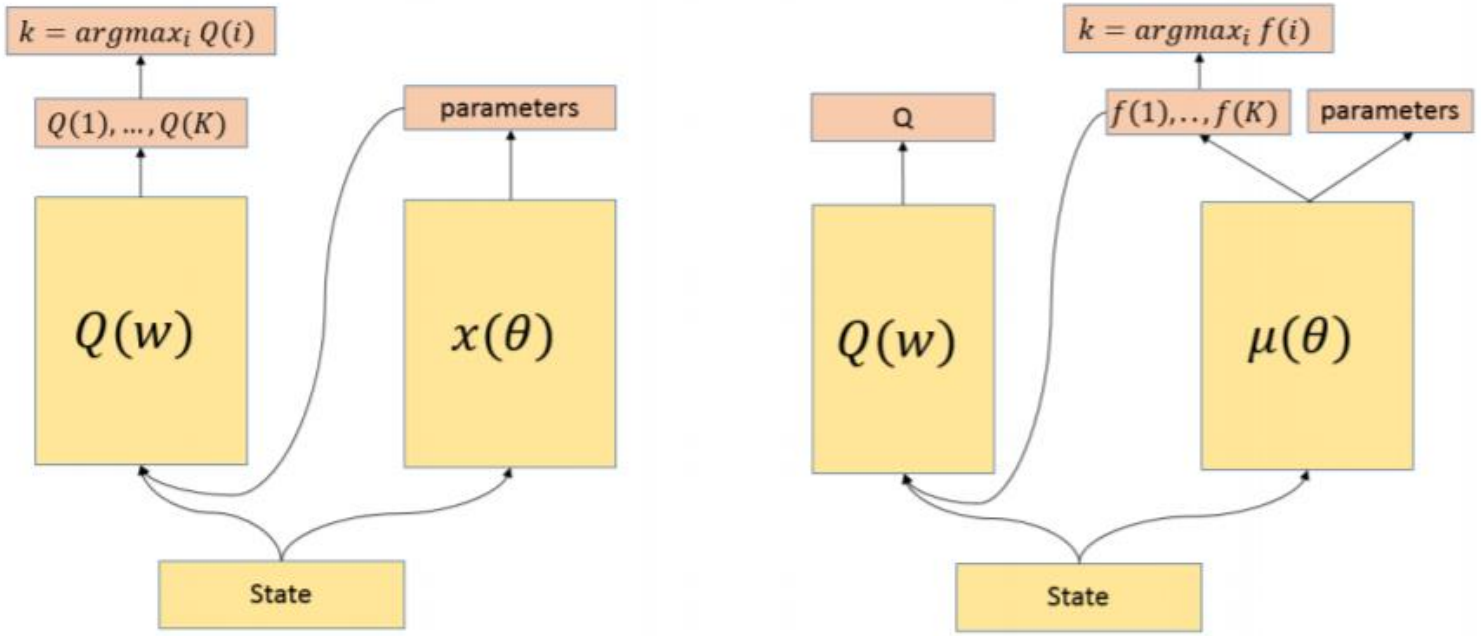
参数化深度Q网络（DQN）

- **离散动作空间：**深度Q网络（DQN）
- **连续动作空间：**深度确定性策略梯度（DDPG）
- **混合动作空间：**DQN + DDPG？

思路：使用深度Q网络选择离散动作，使用深度确定性策略梯度决定连续参数。

Xiong, Jiechao, et al. Parametrized Deep Q-Networks Learning: Reinforcement Learning with Discrete-Continuous Hybrid Action Space. 2017

参数化深度Q网络（DQN）

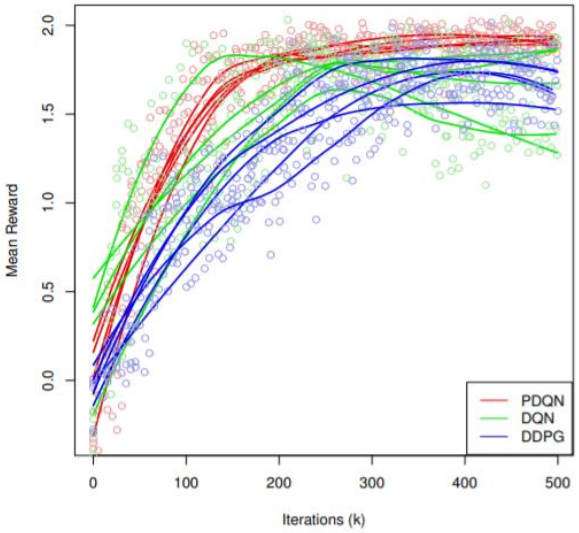


(a) Network of P-DQN　　(b) Network of DDPG

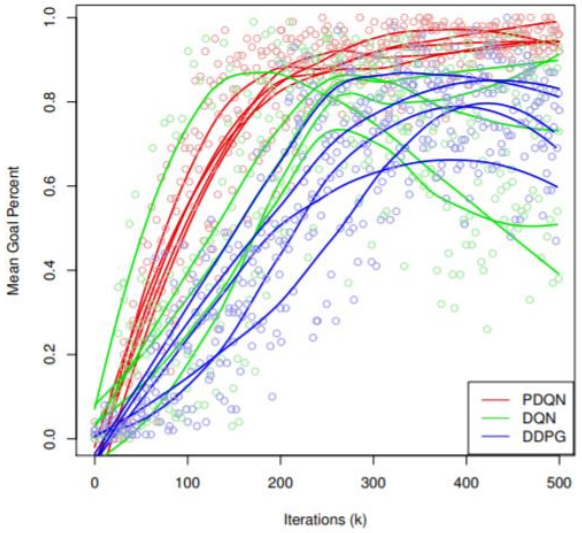
$$\ell_t^Q(w) = \frac{1}{2}\left[Q\left(s_t, k_t, x_{k_t}; w\right) - y_t\right]^2 \quad \text{且} \quad \ell_t^\Theta(\theta) = -\sum_{k=1}^{K} Q(s_t, k, x_k(s_t;\theta); w_t)$$

Xiong, Jiechao, et al. Parametrized Deep Q-Networks Learning: Reinforcement Learning with Discrete-Continuous Hybrid Action Space. 2017

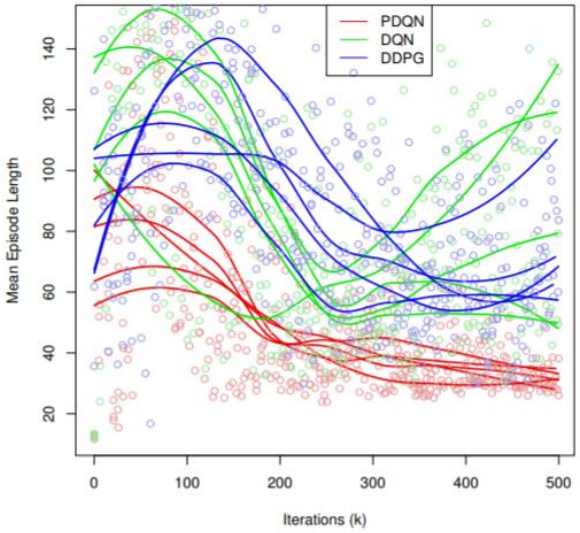
简单实例上的实验



平均奖励



平均目标百分比



平均轮长度

Xiong, Jiechao, et al. Parametrized Deep Q-Networks Learning: Reinforcement Learning with Discrete-Continuous Hybrid Action Space. 2017

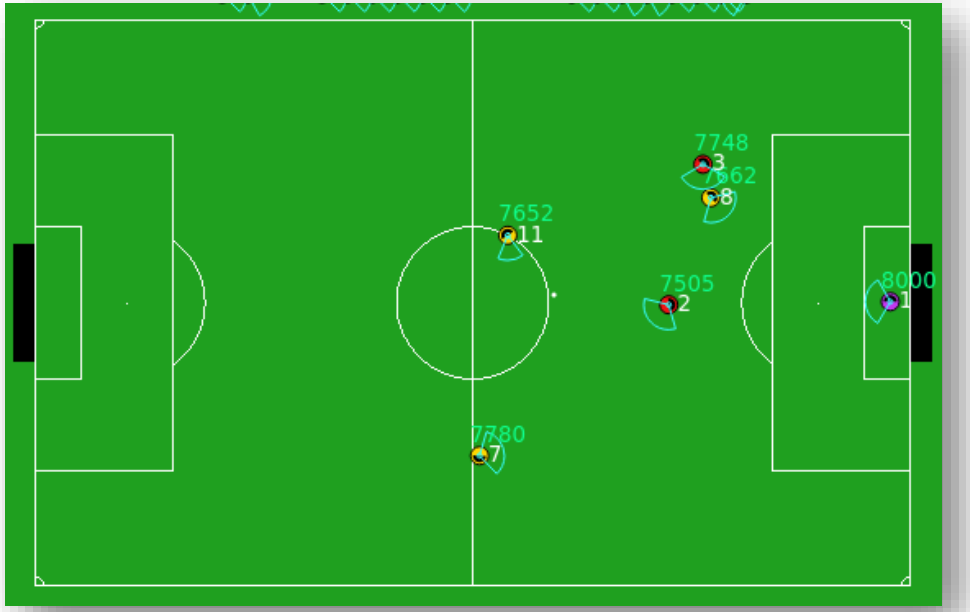
半场区域进攻

冲刺 (方向，速度)

转向 (方向)

滑铲 (方向)

踢球 (方向，力度)



- RoboCup 2D Half-Field-Offense（HFO）——一个探索单智能体学习，多智能体学习和adhoc团队合作的研究平台。
  - HFO具有低级连续状态空间和参数化连续动作空间

https://github.com/LARG/HFO

HFO实验结果

	Scoring Percent	AS to Goal		Scoring Percent	AS to Goal
Helios' Champion	.962	72.0	$PDQN_1$	.997	78.1
SARSA	.81	70.7	$PDQN_2$	.997	78.1
$DDPG_1$	1	108.0	$PDQN_3$	.996	78.1
$DDPG_2$	.99	107.1	$PDQN_4$	.994	81.5
$DDPG_3$	.98	104.8	$PDQN_5$	.992	78.7
$DDPG_4$	.96	112.3	$PDQN_6$	.991	79.9
$DDPG_5$	.94	119.1	$PDQN_7$	.985	82.2
$DDPG_6$	.84	113.2	$PDQN_8$	.984	87.9
$DDPG_7$	.80	118.2	$PDQN_9$	.979	78.5

Xiong, Jiechao, et al. Parametrized Deep Q-Networks Learning: Reinforcement Learning with Discrete-Continuous Hybrid Action Space. 2017

总结参数化动作空间

- 一个参数化的动作空间是一个混合动作空间：包含离散的动作和连续的动作（参数）
  - 智能体可以同时决定执行何种离散动作和使用什么样的连续参数来执行该动作

- 参数化动作空间：要求算法能够处理离散-连续混合动作空间。主要有三个途径：
  - 离散化连续动作空间
  - 将离散动作空间放宽到连续空间
  - 分开处理离散动作和连续动作

THANK YOU