

多智能体系统与强化学习

主讲人：高阳、杨林、杨天培

<https://reinforcement-learning-2025.github.io/>

第三讲：函数估计

走向大规模问题

高 阳

大 纲

值函数估计

值函数估计的更新

线性值函数估计中的状态表征

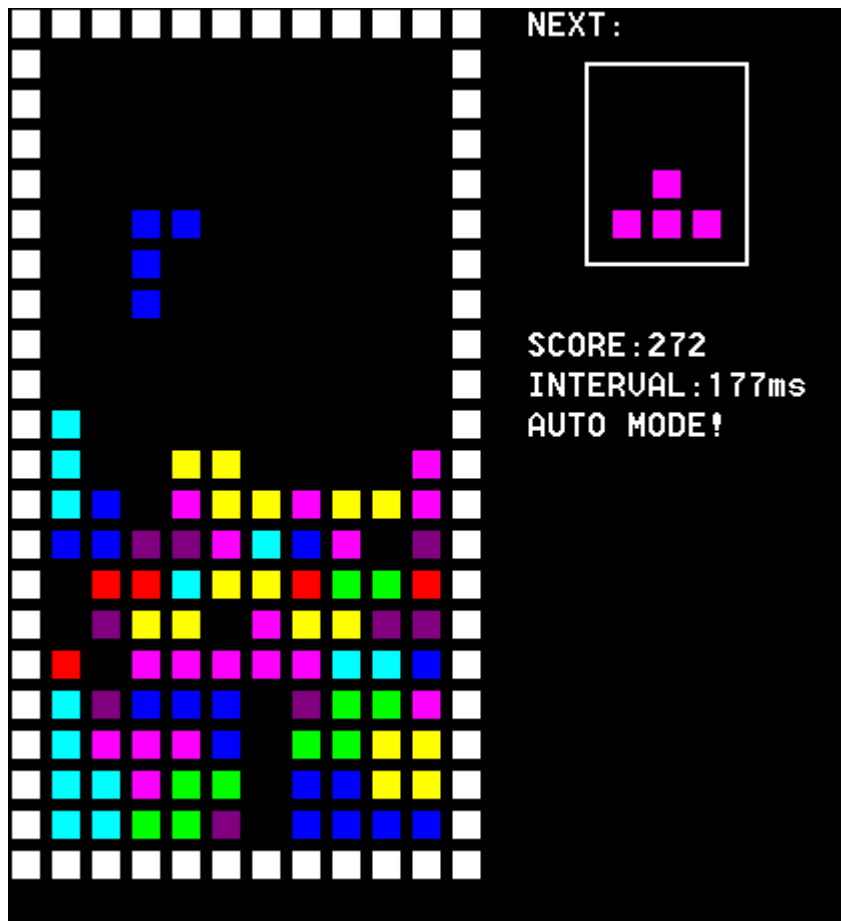
大 纲

值函数估计

值函数估计的更新

线性值函数估计中的状态表征

俄罗斯方块游戏



□ Tetris游戏

- ✓ 1984年，由俄罗斯程序员Alexey Pazhitnov发明
- ✓ 标准游戏桌面：行宽10，列高20
- ✓ 块的形状：S、Z、L、I、O、T

□ RL问题视角

- ✓ 状态空间 $2^{200 \times 6}$
- ✓ 动作：平移、旋转
- ✓ 即时奖赏：每消除一行得正分

用强化学习求解俄罗斯方块游戏

Playing Tetris with Deep Reinforcement Learning

Matt Stevens

mslf@stanford.edu

Sabeek Pradhan

sabeekp@stanford.edu

Abstract

We used deep reinforcement learning to train an AI to play tetris using an approach similar to [7]. We use a convolutional neural network to estimate a Q function that describes the best action to take at each game state. This approach failed to converge when directly applied to predicting individual actions with no help from heuristics. However, we implemented many features that improved convergence. By grouping actions together, we allowed the Q network to focus on learning good game configurations instead of how to move pieces. By using transfer learning from a heuristic model we were able to greatly improve performance, having already learned relevant features about Tetris and optimal piece configurations. And by using prioritized sweeping, we were able to reduce the some of the inherent instability in learning to estimate single actions. Although we did not surpass explicitly featurized agents in terms of performance, we demonstrated the basic ability of a neural network to learn to play Tetris, and with greater training time we might have been able to beat even those benchmarks.

1. Introduction

The primary motivation for this project was a fun application of neural networks. Tetris is a well-known game that is loved and hated by many. Recent results from the team at DeepMind have shown that neural networks can have remarkable performance at game playing, using a minimal amount of prior information about the game. However, there has not yet been a successful attempt to use a convolutional neural network to learn to play Tetris. We hope to fill this gap.

Furthermore, research into solving control problems with neural networks and minimal explicit featurization may have applications outside of just video games, such as in self-driving cars or robotics.

Tetris is a game that involves placing pieces on top of each other so that they fill in a rectangular grid. A common human approach to this problem involves looking for open spaces that match the shape of the current piece in play. This

is a ultimately a task in visual pattern detection and object recognition, so we believe that a convolutional neural network is a natural approach to solving this problem. Tetris also involves a fair degree of strategy, and recent advances in deep reinforcement learning have shown that convolutional neural networks can be trained to learn strategy.

2. Related Work

One of the seminal works in the field of deep reinforcement paper was DeepMind's 2013 paper, Playing Atari with Deep Reinforcement Learning [6]. This paper and their 2015 follow-up [7] served as our primary inspiration, both for the general idea applying computer vision to reinforcement learning for video games and for ideas on how to proceed with our model architecture. Their results are extremely impressive and state-of-the-art, with [7] achieving superhuman performance on over half the games tested and falling short primarily on games that require long-term strategic planning and have very sparse rewards, such as Montezuma's Revenge.

Several of the techniques that [6], [7], and our paper use have much longer histories. For example, experience replay, which stores the game-playing agent's experiences to be trained long after the original policy has been abandoned, was introduced in [4]. RMSProp, an optimization technique that naturally anneals learning rates down over time and focuses its learning rates on dimensions with more consistent and thus informative gradients, was first introduced in [9]. [6] and [7] recommended but did not implement prioritized sweeping as a way to select saved examples to train on; we implemented this in our paper. Prioritized sweeping, which trains the model on the examples with the largest error, was introduced in [8]; a more computationally efficient version, which we did not implement but would consider as an extension, was introduced in [1]. Finally, one optimization method we experimented with but that [6] and [7] did not mention was Adadelta. Adadelta, which is less hyperparameter-dependent than other optimization methods and anneals effective learning rates down based on how much a given dimension has already been updated, was introduced in [11].

Finally, it is worth noting that previous attempts to build



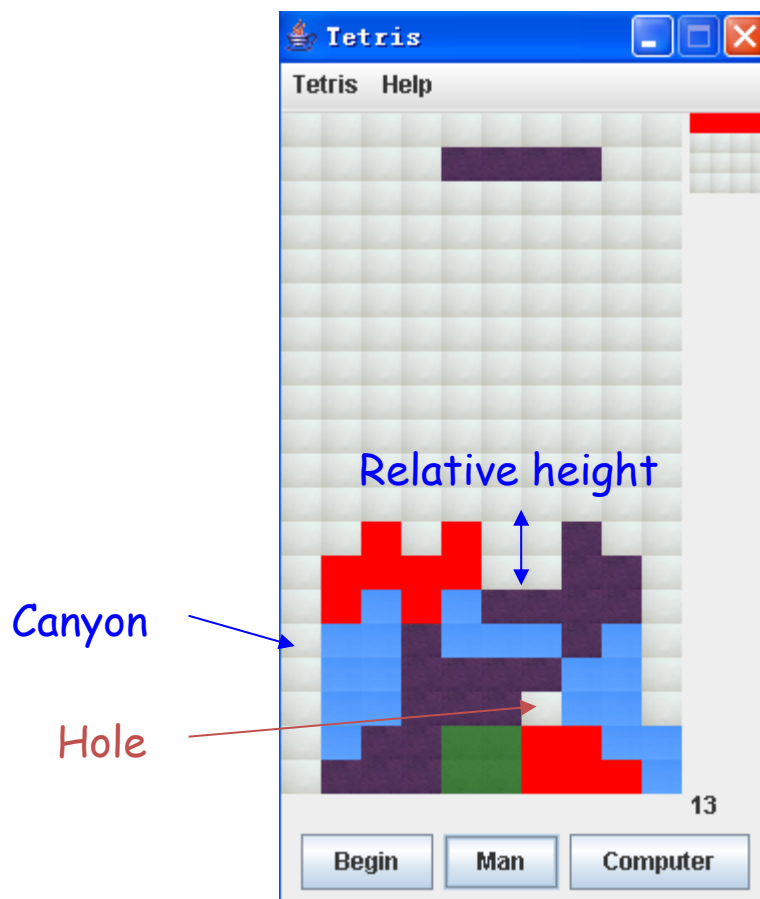
关系强化学习(Relational RL)

□ 属性

- Height of wall (max, avg, min)
- Number of Holes
- Height difference adjacent cols
- Canyon (width, height)
- ...

□ 宏动作

- Fits
- Increases height, ...
- Number of deleted lines
- ...



值函数估计

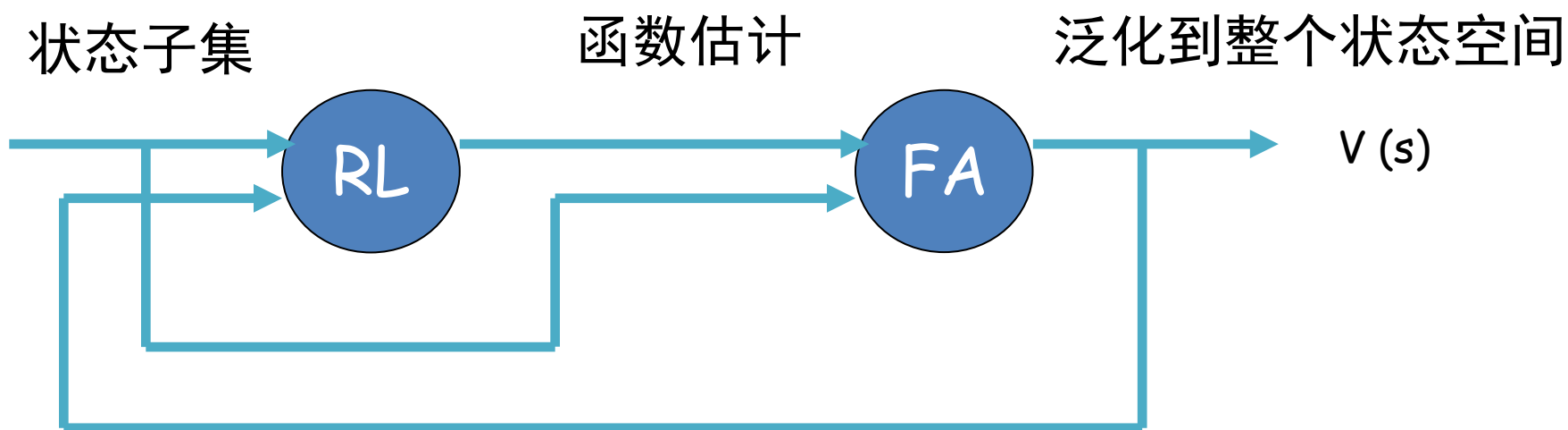
□ 简单RL问题

- ✓ 值函数存储以查找表(lookup table)方式
- ✓ 对应算法称之为值表型强化学习算法

□ 复杂RL问题

- ✓ 在时间步 t ，值函数 $V(s_t)$ ，依赖于参数向量 \mathbf{w}_t
- ✓ \mathbf{w} 可以是显式的，如关系强化学习
- ✓ \mathbf{w} 可以是隐式的，如神经网络强化学习

强化学习中的值函数估计



$$V_0, M(V_0), \Gamma(M(V_0)), M(\Gamma(M(V_0))), \Gamma(M(\Gamma(M(V_0)))) , \dots$$

注意：存在两个逼近过程

大 纲

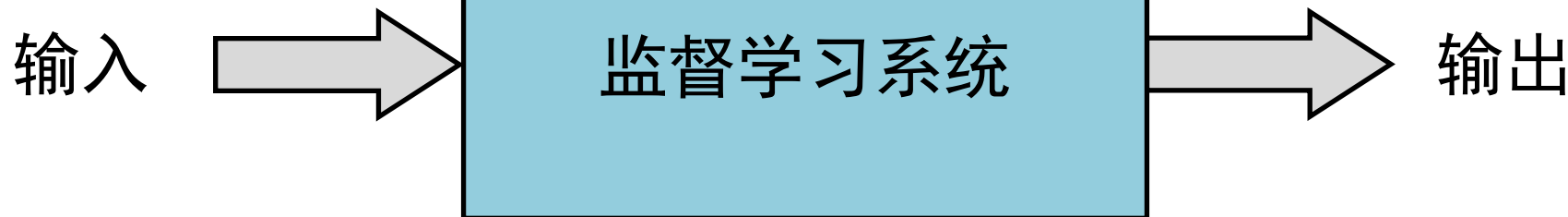
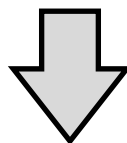
值函数估计

值函数估计的更新

线性值函数估计中的状态表征

监督学习视角

实际输出=期望（目标输出）



训练样本 = {输入, 目标输出}

误差 = (目标输出 - 实际输出)

值表强化学习

□ 值表型强化学习更新公式

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

监督学习训练视角下

{Description of s_t , $r_{t+1} + \gamma V(s_{t+1})$ }



输入



目标输出

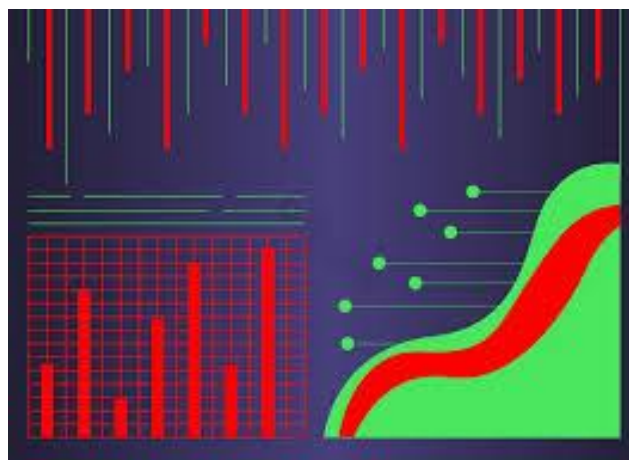
其他值函数估计方法

□ 值函数估计器

- ✓ 特征的线性组合
- ✓ 决策树
- ✓ 神经网络
- ✓ 核函数
- ✓

但强化学习算法仍然需要特殊考虑：

- ✓ 数据非独立同分布 (non-iid)
- ✓ 学习过程非稳定 (non-stationary)



梯度下降方法(前修课程：最优化、机器学习)

□ 存在一个带有参数 \mathbf{w} 的可微函数 $J(\mathbf{w})$

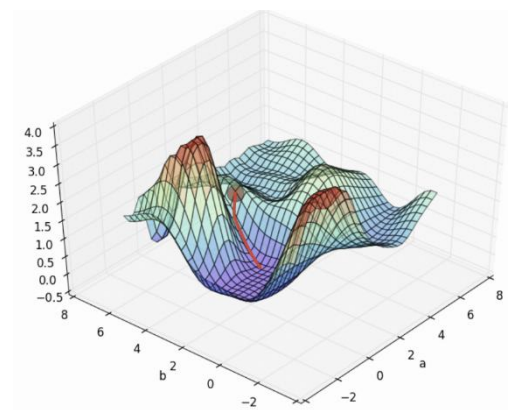
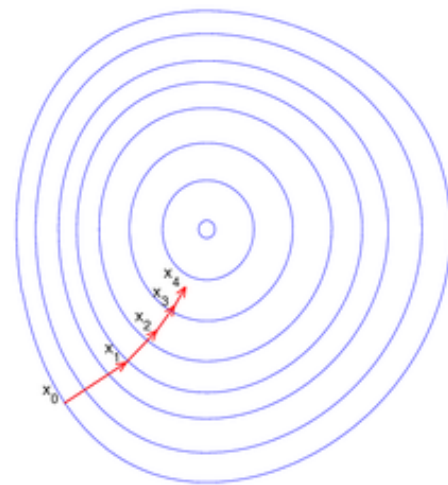
□ 定义 $J(\mathbf{w})$ 的梯度如下：

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = \begin{pmatrix} \frac{\partial J(\mathbf{w})}{\partial w_1} \\ \vdots \\ \frac{\partial J(\mathbf{w})}{\partial w_n} \end{pmatrix}$$

□ **目标**：找到 $J(\mathbf{w})$ 的一个局部最小值

□ 一个局部下降最快的方向——**负梯度**

$$\Delta \mathbf{w} = -\frac{1}{2} \alpha \nabla_{\mathbf{w}} J(\mathbf{w}) \quad \alpha \text{ 为一个步长参数}$$



基于随机梯度下降的值函数估计

□ **目标：** 寻找 \mathbf{w} 最小化值函数**估计值**与**真实值**之间的均方误差

$$J(\mathbf{w}) = \mathbb{E}_{\pi} \left[\left(\underset{\substack{\uparrow \\ \text{真实值}}}{V^{\pi}(s)} - \underset{\substack{\uparrow \\ \text{估计值}}}{\hat{V}(s, \mathbf{w})} \right)^2 \right]$$

□ 误差减少的方向

$$-\frac{1}{2} \nabla_{\mathbf{w}} J(\mathbf{w}) = \mathbb{E}_{\pi} \left[\left(V^{\pi}(s) - \hat{V}(s, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{V}(s, \mathbf{w}) \right]$$

□ 采样梯度下降

$$\Delta \mathbf{w} = -\frac{1}{2} \alpha \nabla_{\mathbf{w}} J(\mathbf{w}) = \alpha \left(V^{\pi}(s) - \hat{V}(s, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{V}(s, \mathbf{w})$$

线性值函数估计

□ 线性值函数可由特征与参数线性组合而成

$$\hat{V}(s, \mathbf{w}) = \mathbf{x}(s)^\top \mathbf{w} = \sum_{i=1}^n w_i x_i(s)$$

□ 目标函数可进一步表示为

$$J(\mathbf{w}) = \mathbb{E}_\pi[(V^\pi(s) - \mathbf{x}(s)^\top \mathbf{w})^2]$$

□ 鉴于随机梯度方法可以收敛到最优解，于是更新公式如下

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} - \frac{1}{2} \alpha \nabla_{\mathbf{w}} J(\mathbf{w}) \\ &= \mathbf{w} + \alpha \underbrace{\left(V^\pi(s) - \hat{V}(s, \mathbf{w}) \right)}_{\text{预测误差}} \underbrace{\mathbf{x}(s)}_{\text{特征值}} \end{aligned}$$

↑ ↑ ↑
步长 预测误差 特征值

线性值函数估计

□ 值表可看作线性值函数估计的一种特殊形式

□ 值表特征可以表示

$$\mathbf{x}^{\text{table}}(s) = \begin{bmatrix} 1(s = s_1) \\ \vdots \\ 1(s = s_n) \end{bmatrix} \rightarrow \text{指示函数}$$

□ 值表函数估计可以进一步表示

$$\hat{V}(s, \mathbf{w}) = \begin{bmatrix} 1(s = s_1) \\ \vdots \\ 1(s = s_n) \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

线性值函数估计

□ 随机梯度优化假设真实值函数 V^π 可知，但在强化学习中，仅有奖励信号。因此，需要找到目标值替代真实值函数

✓ 蒙特卡罗方法（MC），使用回报 G_t

$$\Delta \mathbf{w} = \alpha \left(G_t - \hat{V}(s_t, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{V}(s_t, \mathbf{w})$$

✓ 时差学习TD(0)，使用目标值 $R_{t+1} + \gamma \hat{V}(s_{t+1}, \mathbf{w})$

$$\Delta \mathbf{w} = \alpha \left(R_{t+1} + \gamma \hat{V}(s_{t+1}, \mathbf{w}) - \hat{V}(s_t, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{V}(s_t, \mathbf{w})$$

✓ 时差学习TD(λ)，使用 λ -回报 G_t^λ

$$\Delta \mathbf{w} = \alpha \left(G_t^\lambda - \hat{V}(s_t, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{V}(s_t, \mathbf{w})$$

蒙特卡罗值函数估计

□ 回报 G_t 是关于真实值函数的带噪音无偏采样(unbiased sampling)

□ 在在线收集的训练数据上运用监督学习对价值函数进行预测

$$\langle s_1, G_1 \rangle, \langle s_2, G_2 \rangle, \dots, \langle s_T, G_T \rangle$$

□ 对于每个数据样本

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} + \alpha \left(G_t - \hat{V}(s_t, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{V}(s_t, \mathbf{w}) \\ &= \mathbf{w} + \alpha \left(G_t - \hat{V}(s_t, \mathbf{w}) \right) \mathbf{x}(s_t) \end{aligned}$$

□ 蒙特卡罗估计至少能收敛到一个局部最优解

□ 在值函数为线性的情况下可以收敛到全局最优

时差TD(0)值函数估计

□ 目标值 $R_{t+1} + \gamma \hat{V}(s_{t+1}, \mathbf{w})$ 是关于真实值函数的有偏采样

□ 在“训练数据”上运用监督学习对价值函数进行预测

$$\langle s_1, R_2 + \gamma \hat{V}(s_2, \mathbf{w}) \rangle, \langle s_2, R_3 + \gamma \hat{V}(s_3, \mathbf{w}) \rangle, \dots, \langle s_{T-1}, r_T \rangle$$

□ 对于每个数据样本

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left(R_{t+1} + \gamma \hat{V}(s_{t+1}, \mathbf{w}) - \hat{V}(s_t, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{V}(s_t, \mathbf{w})$$

$$\approx \mathbf{w} + \alpha \left(R_{t+1} + \gamma \hat{V}(s_{t+1}, \mathbf{w}) - \hat{V}(s_t, \mathbf{w}) \right) \mathbf{x}(s_t)$$

半梯度

□ 线性情况下时差学习(接近)收敛到全局最优解

时差TD(λ)值函数估计

□ λ - 回报 G_t^λ 同样是真实值函数的有偏采样

□ 在“训练数据”上运用监督学习对价值函数进行预测

$$\langle s_1, G_1^\lambda \rangle, \langle s_2, G_2^\lambda \rangle, \dots, \langle s_{T-1}, G_{T-1}^\lambda \rangle$$

□ 前向视角线性TD(λ)

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left(G_t^\lambda - \hat{V}(s_t, \mathbf{w}) \right) \mathbf{x}(s_t)$$

□ 后向视角线性TD(λ)

$$\delta_t = R_{t+1} + \gamma \hat{V}(s_{t+1}, \mathbf{w}) - \hat{V}(s_t, \mathbf{w})$$

$$E_t = \gamma \lambda E_{t-1} + \mathbf{x}(s_t)$$

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \delta_t E_t$$

大 纲

值函数估计

值函数估计的更新

线性值函数估计中的状态表征

回顾：线性值函数估计

□ 回顾：线性值函数可由特征与参数线性组合而成

$$\hat{V}(s, \mathbf{w}) = \mathbf{x}(s)^\top \mathbf{w} = \sum_{i=1}^n w_i x_i(s)$$

上一小节重点：参数更新

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} - \frac{1}{2} \alpha \nabla_{\mathbf{w}} J(\mathbf{w}) \\ &= \mathbf{w} + \alpha \left(V^\pi(s) - \hat{V}(s, \mathbf{w}) \right) \mathbf{x}(s) \end{aligned}$$

步长

预测误差

特征值

状态表征同样重要

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) \\ \vdots \\ x_k(s) \end{bmatrix}$$

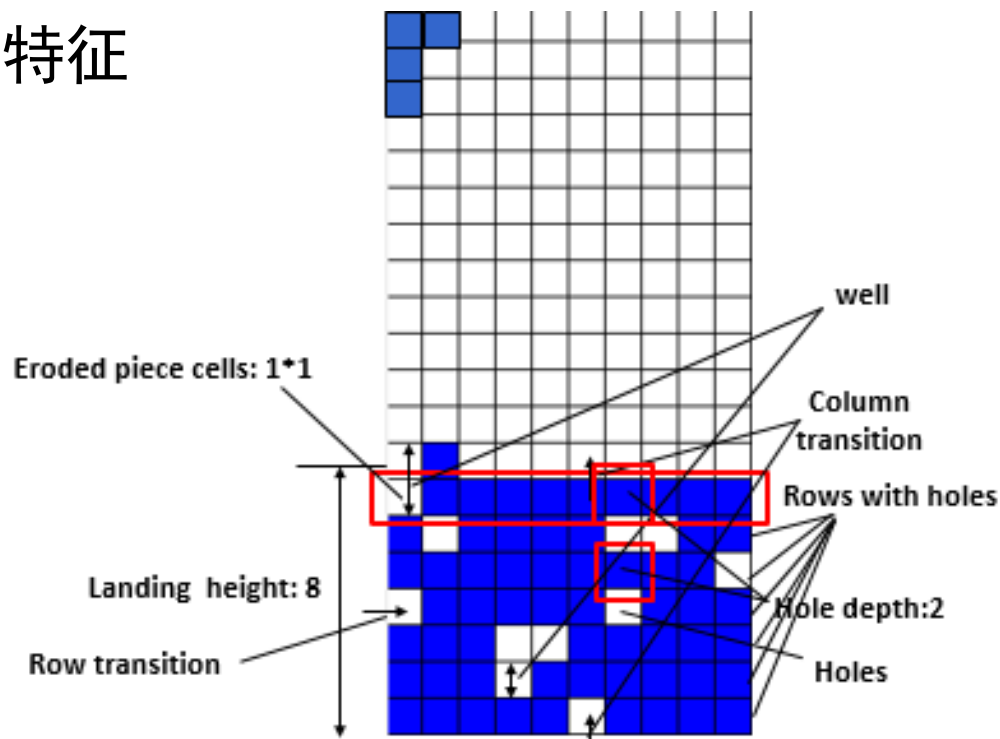
线性值函数估计的状态表征

□ 状态特征表示一般形式

$$\mathbf{x}(s) = \begin{bmatrix} x_1(s) \\ \vdots \\ x_k(s) \end{bmatrix}$$

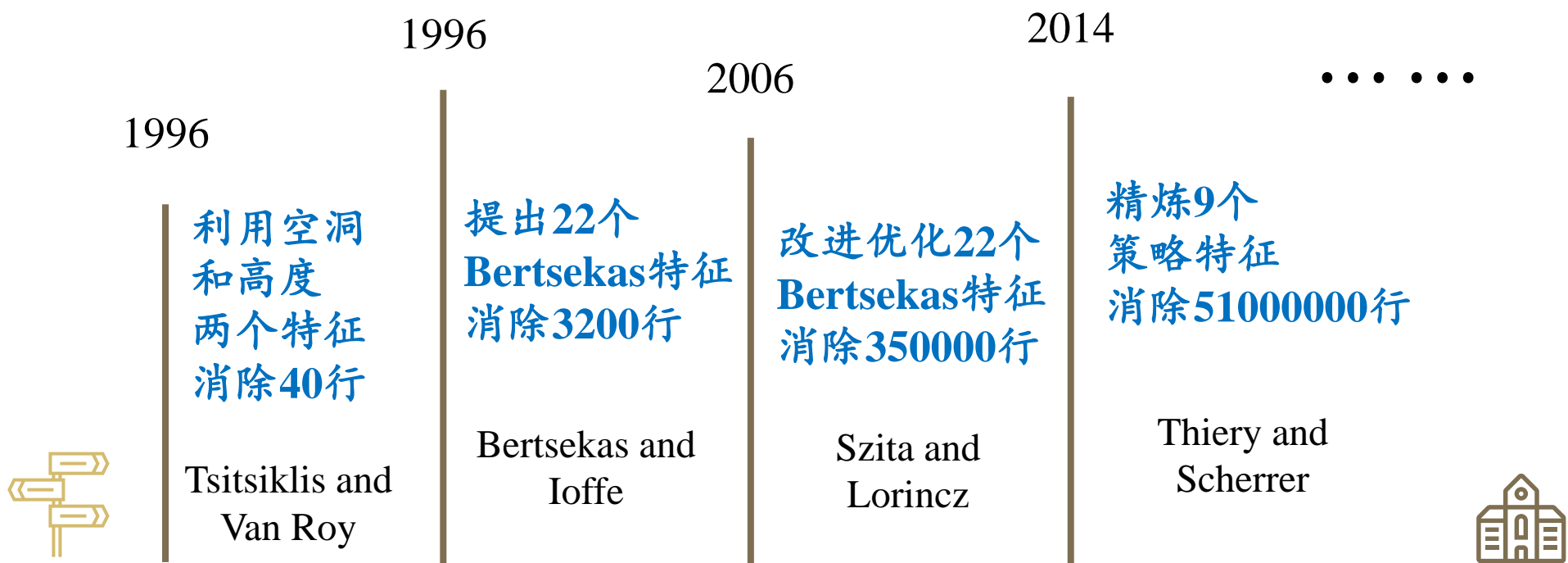
□ 专家定义俄罗斯方块(Tetris)特征

- ✓ 空洞数
- ✓ 下落高度
- ✓ 井深度
- ✓ 行转移数
-



线性值函数估计的状态表征

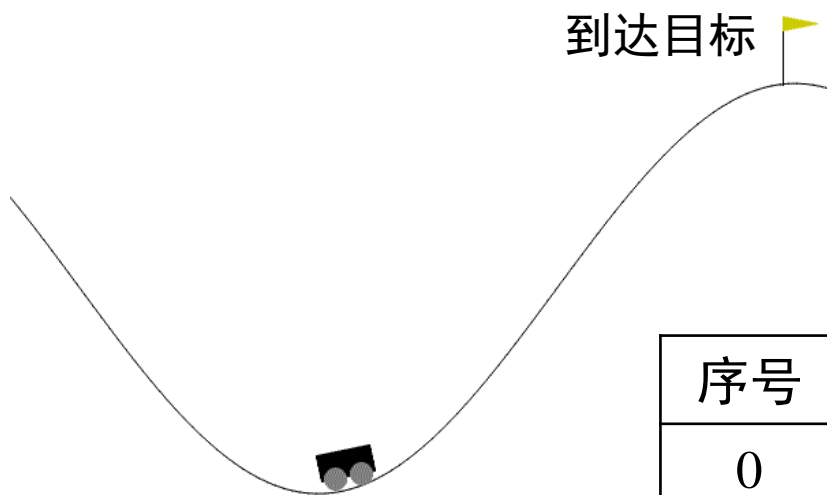
□ 俄罗斯方块(Tetris)特征专家定义历史



人工定义**依赖大量的专业知识**，挑战性较大且难以扩展

线性值函数估计的状态表征

□ 小车爬山为例



序号	动作空间
0	向左加速
1	不加速
2	向右加速

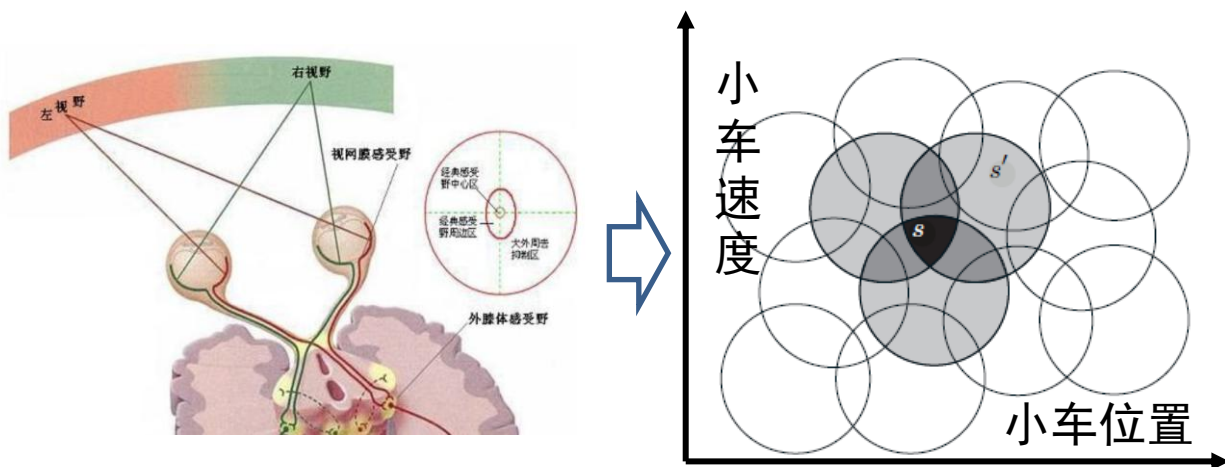
序号	原始观测空间	最大值	最小值
0	小车x轴的位置	0.6	-1.2
1	小车速度	0.07	-0.07

在二维连续状态空间中，人工定义特征困难

线性值函数估计的状态表征

□ 粗编码(Coarse Coding)

- ✓ 通过感受野(圆圈)对状态聚合



在同一感受野内
状态相似

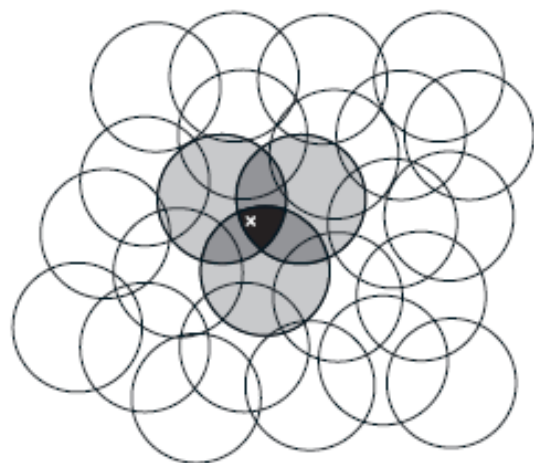
通过感受野对
状态空间划分

描述	编码
在感受野内	1
不在感受野内	0

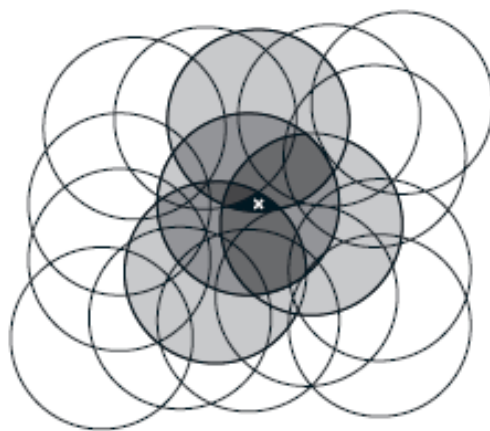
线性值函数估计的状态表征

□ 粗编码(Coarse Coding)

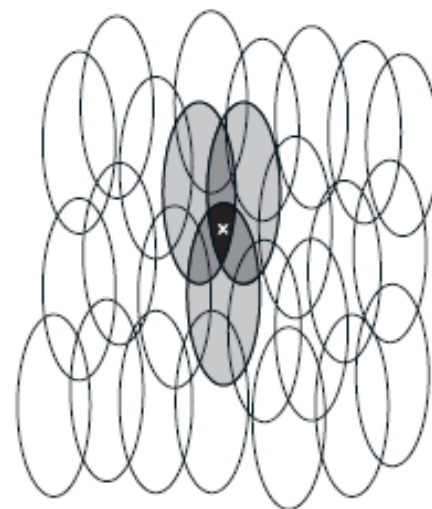
- ✓ 泛化能力(Generalization) 影响 **学习效率**：训练一个状态点会更新所有包含该状态点的圆对应的权重，进而影响圆中的状态
- ✓ 区分能力(Discriminations) 影响 **求解质量**：衡量两个状态点所对应圆的差异性，进而影响逼近质量



窄泛化
(弱泛化、强区分)



宽泛化
(强泛化、弱区分)

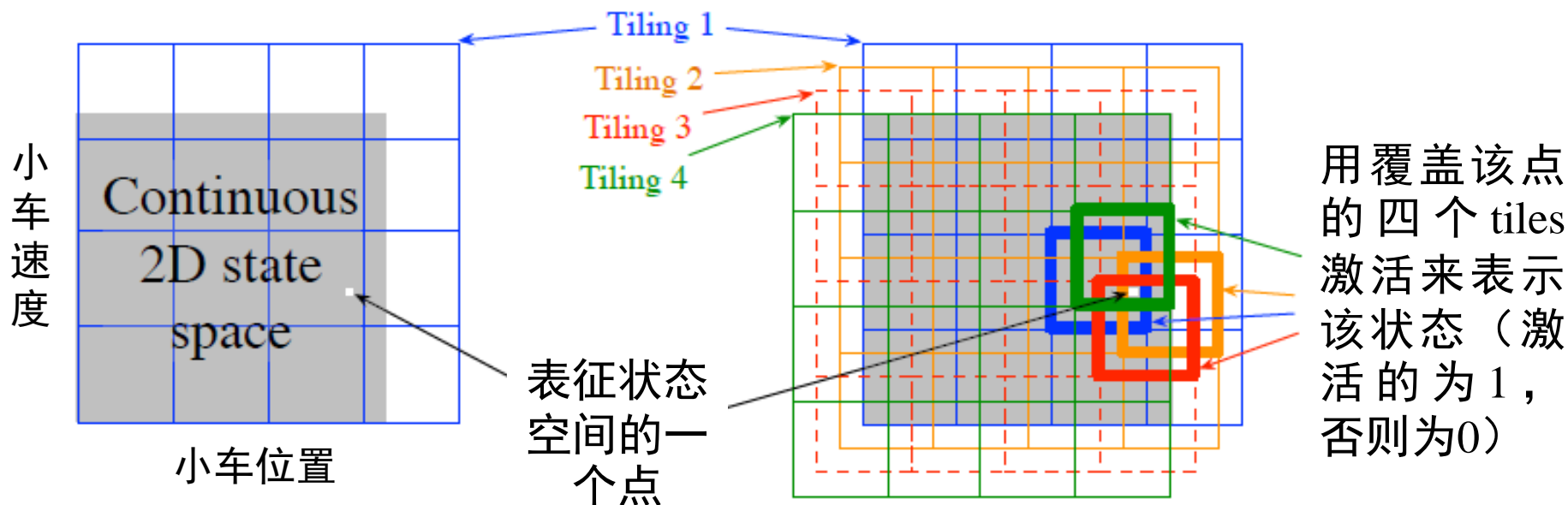


纵横泛化
(自适应泛化、区分)

线性值函数估计的状态表征

□ 瓦片编码(Tile Coding)

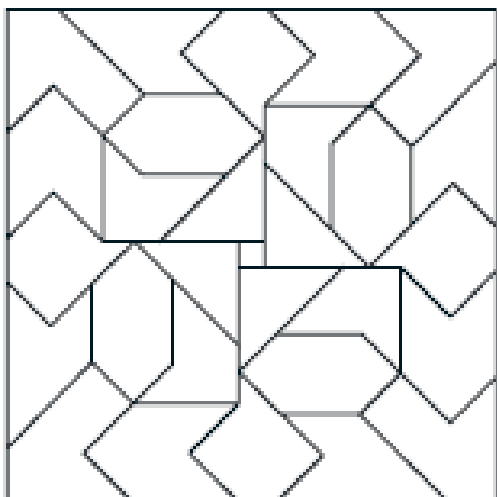
- ✓ 更好的区分能力：单个Tiling覆盖所有状态，Tiling内部相互独立。
- ✓ 更好的泛化能力：能保证每个Tiling中有一个tile激活



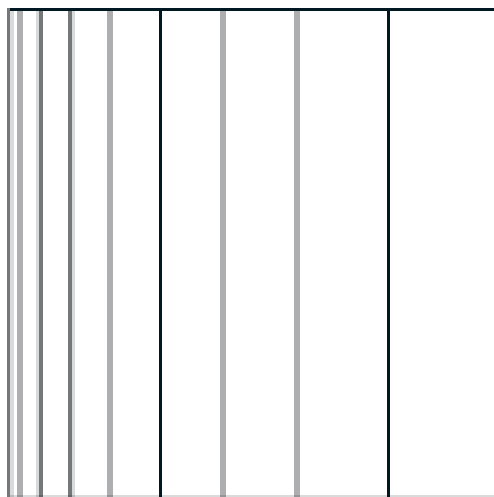
线性值函数估计的状态表征

□ 瓦片编码(Tile Coding)

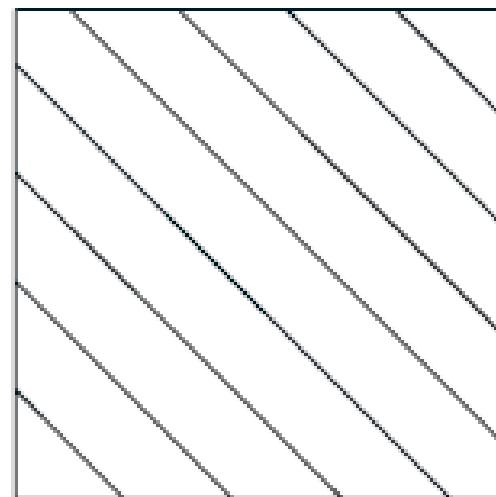
- ✓ 为了避免状态过渡聚合影响区分效果，根据不同的状态分布，可以适当调整分块方式，有下几种：



不规则型
(人工)



对数型
(某一维度log分布)



对角型
(状态分布对角均匀)

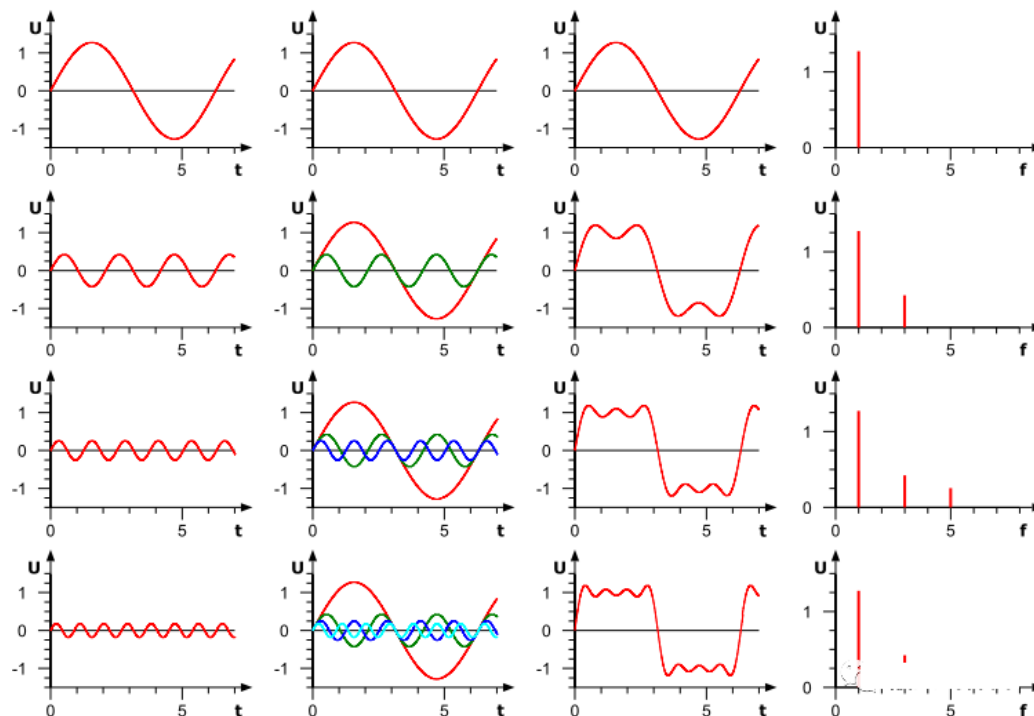
线性值函数估计的状态表征

□ 傅立叶基

- ✓ 无线逼近定理：有足够的基函数，就能足够精确地近似任意函数
- ✓ 基于傅立叶基的状态表征：周期函数表示为正弦和余弦函数的加权和

$$x_i(s) = \cos(i\pi s), s \in [0,1]$$

特征表示

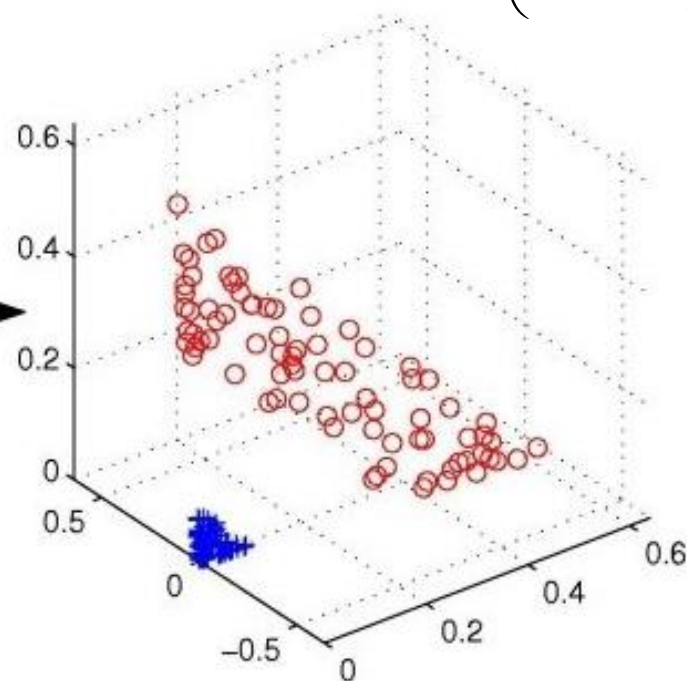
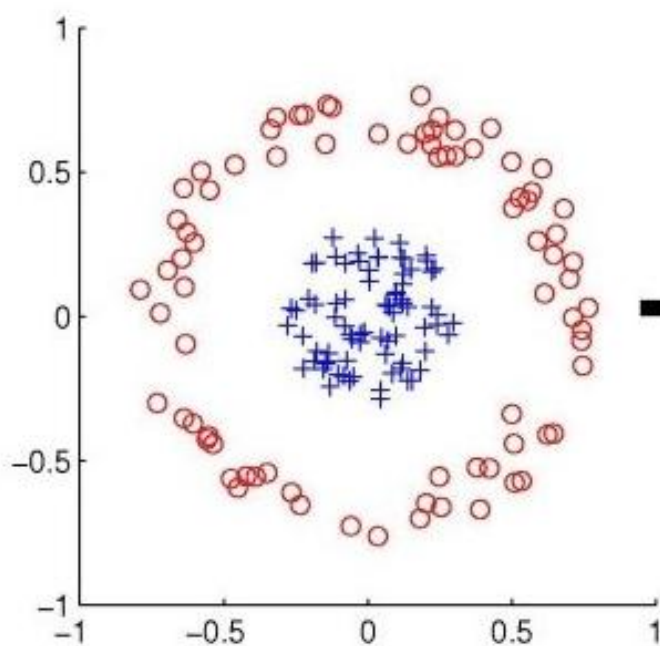


线性值函数估计的状态表征

□ 核函数

✓ 从状态空间的低维不可分到高维可分

$$\phi_s(i) = \exp\left(-\frac{\|s - c_i\|^2}{2\sigma_i^2}\right)$$



线性值函数估计的状态表征

□ 核函数

✓ 从选择核函数到关注核函数以期提升值函数估计能力

1944

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 24, NO. 12, DECEMBER 2013

Online Selective Kernel-Based Temporal Difference Learning

Xingguo Chen, Yang Gao, Member, IEEE, and Ruili Wang

Abstract—In this paper, an online selective kernel-based temporal difference (OSKTD) learning algorithm is proposed to deal with large scale and/or continuous reinforcement learning problems. OSKTD includes two online procedures: online sparsification and parameter updating for the selective kernel-based value function. A new sparsification method (i.e., a kernel distance-based online sparsification method) is proposed based on selective ensemble learning, which is computationally less complex compared with other sparsification methods. With the proposed sparsification method, the sparsified dictionary of samples is constructed online by checking if a sample needs to be added to the sparsified dictionary. In addition, based on local validity, a selective kernel-based value function is proposed to select the best samples from the sample dictionary for the selective kernel-based value function approximator. The parameters of the selective kernel-based value function are iteratively updated by using the temporal difference (TD) learning algorithm combined with the gradient descent technique. The complexity of the online sparsification procedure in the OSKTD algorithm is $O(n)$. In addition, two typical experiments (Maze and Mountain Car) are used to compare with both traditional and up-to-date $O(n)$ algorithms (GTD, GTD2, and TDC using the kernel-based value function), and the results demonstrate the effectiveness of our proposed algorithm. In the Maze problem, OSKTD converges to an optimal policy and converges faster than both traditional and up-to-date algorithms. In the Mountain Car problem, OSKTD converges, requires less computation time compared with other sparsification methods, gets a better local optima than the traditional algorithms, and converges much faster than the up-to-date algorithms. In addition, OSKTD can reach a competitive ultimate optima compared with the up-to-date algorithms.

Index Terms—Function approximation, online sparsification, reinforcement learning (RL), selective ensemble learning, selective kernel-based value function.

scale and/or continuous space RL problems suffer from the curse of dimensionality, which means that the complexity exponentially increases with the dimensionality of a state space (i.e., the number of state variables) [1]. In other words, traditional tabular RL becomes impractical. A possible approach toward coping with this curse is approximate RL or approximate dynamic programming including value function approximation [2], policy search [3], actor-critic approaches [4], hierarchical RL [5], and transfer learning [6]. In this paper, we focus on value function approximation which is the most popular method.

There are several value function approximation techniques, such as linear function approximation [1], [2], regression tree methods [8], neural networks [9]–[13], and kernel methods [14]–[17]. Among them, we focus on the kernel methods in this paper. According to the Representer Theorem [18], a value function in RL can be formed as $V(s) = \sum_{i=1}^N \theta_i k(s, s_i)$, where s is a state, θ is the parameter vector, $k(\cdot, \cdot)$ is the kernel function, and s_i is the sample.

A kernel approach for RL needs to consider the following issues.

- 1) The kernel-based reinforcement learning (KBRL) needs to be able to construct a sample dictionary online, since RL is an online learning method.
- 2) The online sparsification process in constructing the sample dictionary needs to consider the tradeoff between accuracy and the complexity of the value function.
- 3) It is difficult to select a proper kernel function for a problem, since different kernel functions may have different performances for different problems.

Knowledge-Based Systems 278 (2023) 110902



Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Online attentive kernel-based temporal difference learning

Xingguo Chen^{a,b,*}, Guang Yang^{a,c,1}, Shangdong Yang^{a,c,f}, Huihui Wang^{d,e}, Shaokang Dong^c, Yang Gao^c

^a Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, No. 9, Wenyuan Road, 210023, Nanjing, PR China

^b National Engineering Laboratory for Agri-product Quality Traceability, Beijing Technology and Business University, No. 11, Fucheng Road, 100048, Beijing, PR China

^c State Key Laboratory for Novel Software Technology, Nanjing University, No. 163, Xianlin Road, 210046, Nanjing, PR China

^d Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, No. 200, Xiao Lingwei Street, 210094, Nanjing, PR China

^e Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, No. 200, Xiao Lingwei Street, 210094, Nanjing, PR China

^f Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin 541004, China

ARTICLE INFO

Article history:

Received 9 September 2022

Received in revised form 19 July 2023

Accepted 9 August 2023

Available online 18 August 2023

Keywords:

Online kernel-based reinforcement learning

Catastrophic forgetting or interference

Sparse representation

Attentive function

Two-timescale optimization

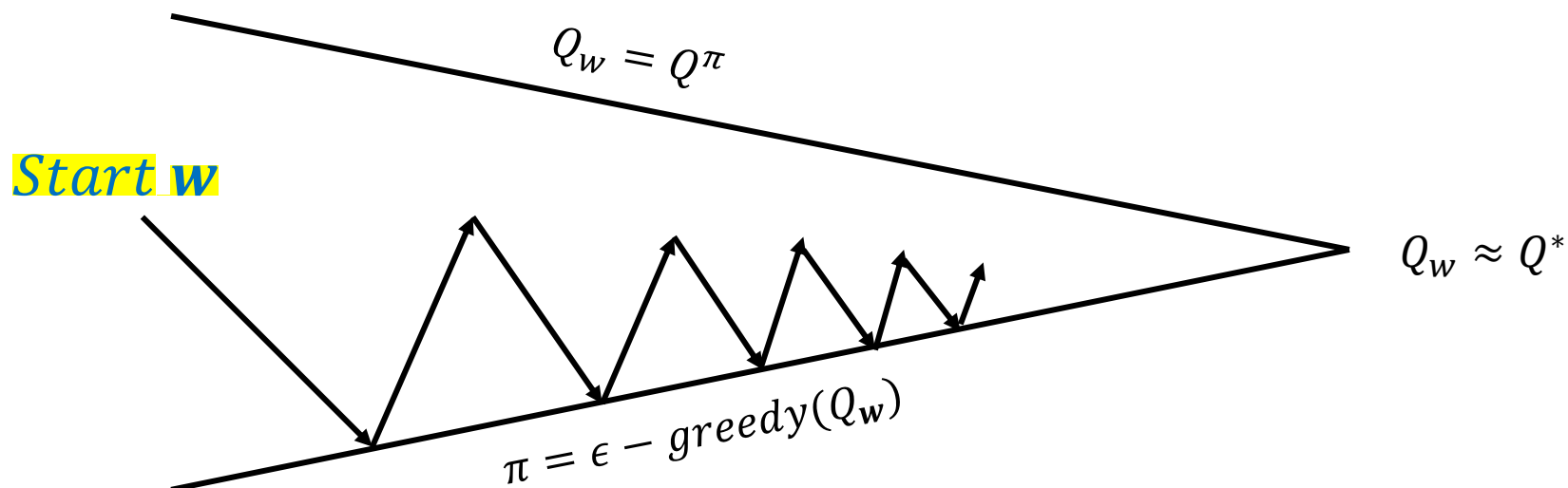
Stability analysis

ABSTRACT

Kernel-based reinforcement learning has received increasing attention because it requires less prior knowledge linear approximation and neural networks. Online kernel-based updating, however, is hindered by the challenge of catastrophic forgetting or interference. Sparse representation is a key method to address this issue, but existing methods fail to satisfy four criteria: learnability, nonprior, nontruncation, and explicitness. In this paper, we present an attentive kernel-based value function approximation as a learnable, nonprior, nontruncated, and explicit sparse representation. We propose the online attentive kernel-based temporal difference (OAKTD) algorithm, which employs two-timescale optimization, and provide a convergence analysis for our proposed algorithm. Experimental results show that OAKTD outperforms online kernel-based TD learning algorithms, and the TD learning algorithm with Tile Coding on classical tasks, i.e., Mountain Car, Acrobot, CartPole and Puddle World.

© 2023 Elsevier B.V. All rights reserved.

值函数估计用作控制



- 策略评估：估计状态-动作值函数， $Q(\cdot, \cdot, w) \approx Q^\pi$
- 策略改进： $\epsilon - \text{greedy}$ 策略探索改进

状态-动作值函数估计

- 同样，对动作-状态值函数进行估计

$$\hat{Q}(s, a, w) \approx Q^\pi(s, a)$$

- 最小均方误差

$$J(\mathbf{w}) = \mathbb{E}_\pi \left[\left(Q^\pi(s, a) - \hat{Q}(s, a, \mathbf{w}) \right)^2 \right]$$

- 在单个样本上进行随机梯度下降

$$\Delta \mathbf{w} = -\frac{1}{2} \alpha \nabla_{\mathbf{w}} J(\mathbf{w}) = \alpha \left(Q^\pi(s, a) - \hat{Q}(s, a, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{Q}(s, a, \mathbf{w})$$

线性状态-动作值函数估计

- 通过特征向量表示状态-动作

$$\mathbf{x}(s, a) = \begin{bmatrix} x_1(s, a) \\ \vdots \\ x_k(s, a) \end{bmatrix}$$

- 线性状态-动作值函数可以表示如下

$$\hat{Q}(s, a, \mathbf{w}) = \mathbf{x}(s, a)^\top \mathbf{w} = \sum_{i=1}^n w_i x_i(s, a)$$

- 基于随机梯度的更新如下

$$\Delta \mathbf{w} = \alpha \left(Q^\pi(s, a) - \hat{Q}(s, a, \mathbf{w}) \right) \mathbf{x}(s, a)$$

线性状态-动作值函数估计

□ 与预测算法一致，需要使用目标值替换未知真实值 $Q^\pi(s, a)$

✓ **MC**，使用回报 G_t

$$\Delta \mathbf{w} = \alpha \left(G_t - \hat{Q}(s_t, a_t, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{Q}(s_t, a_t, \mathbf{w})$$

✓ **TD(0)**，使用目标值 $r_{t+1} + \gamma \hat{Q}(s_{t+1}, a_{t+1}, \mathbf{w})$

$$\Delta \mathbf{w} = \alpha \left(r_{t+1} + \gamma \hat{Q}(s_{t+1}, a_{t+1}, \mathbf{w}) - \hat{Q}(s_t, a_t, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{Q}(s_t, a_t, \mathbf{w})$$

✓ **前向视角TD(λ)**，使用 λ -回报 G_t^λ

$$\Delta \mathbf{w} = \alpha \left(G_t^\lambda - \hat{Q}(s_t, a_t, \mathbf{w}) \right) \nabla_{\mathbf{w}} \hat{Q}(s_t, a_t, \mathbf{w})$$

✓ **后向视角TD(λ)**，类似

$$\delta_t = r_{t+1} + \gamma \hat{Q}(s_{t+1}, a_{t+1}, \mathbf{w}) - \hat{Q}(s_t, a_t, \mathbf{w}), E_t = \gamma \lambda E_{t-1} + \delta_t$$
$$\Delta \mathbf{w} = \alpha \delta_t E_t$$

算法收敛性总结

On/Off-Policy	算法	值表	线性	非线性
On-Policy	MC	✓	✓	✓
	TD(0)	✓	✓	✗
	TD(λ)	✓	✓	✗
Off-Policy	MC	✓	✓	✓
	TD(0)	✓	✗	✗
	TD(λ)	✓	✗	✗

是否有**稳定性更好**的方法？

——下一章节：策略梯度

思考和讨论

1. 特征与值函数估计器
2. 强化学习中的值函数估计更新与监督学习中有什么不同
3. 写出利用线性值函数的强化学习算法

谢谢！