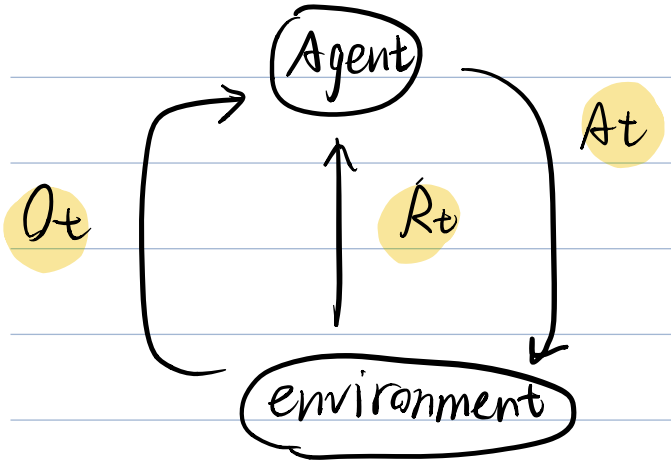


1. Basic Concept

$H_t = O_1, R_1, A_1 \cdots \cdots O_t, R_t, A_t$

$S_t = f(H_t)$ & : used to determine what happen next

- S_t^e : invisible to agent always
- S_t^a 类比：信息筛子选出下一步应该作用于 → action.

$S_t^a = f(H_t)$

information state : Markov State

Markov 性质：

$$P(S_{t+1} | S_t) = P(S_{t+1} | S_1 \cdots S_t)$$

i.e. S_t sufficient statistics

for future

$O_t = S_t^a = S_t^e =$

1 full observable environment

Markov decision process

2. Partial observable environment

$O_t = S_t^a \neq S_t^e$

partially observable POMDP
Markov decision process

S_t^a = 1# H_t complete history 对 S_t^e 的

= 2# ($P[S_t^e = s^1]$, ... ⇒ 概率分布

Beliefs of S_t^e $P(S_t^e = s^n)$)

observation.

= 3# $S_t^a = G(S_{t-1}^a W_s + O_t W_o)$

recurrent neural network

linear combination 旧信息 → 新

2.

RL Agent

- Policy : behavior f.
- Value function :
- Model : subjective representation of environment

;

policy 1# $a = \pi(s)$

2# $\pi(a|s) = P[A_t = a \mid S_t = s]$

value function: prediction of future reward
用于 evaluate how state
用于 select action

对于 state 不同 action. ⇒ future reward

$$V_\pi(s) = E_\pi(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots | S_t = s)$$

slope: γ^n 决定 discount

Model : prediction of environment

Transition P : next state

Rewards R : immediate reward

$$P_{ss'}^a = P(S_{t+1} = t' | A_t = a, S_t = s)$$

$$R_{ss'}^a = P(R_{t+1} | S_t = s, A_t = a)$$

3. 分类

- Value-based
  - value f → policy
- Policy-based
  - policy → action
- Actor-Critic
  - value + policy

Policy-based

policy

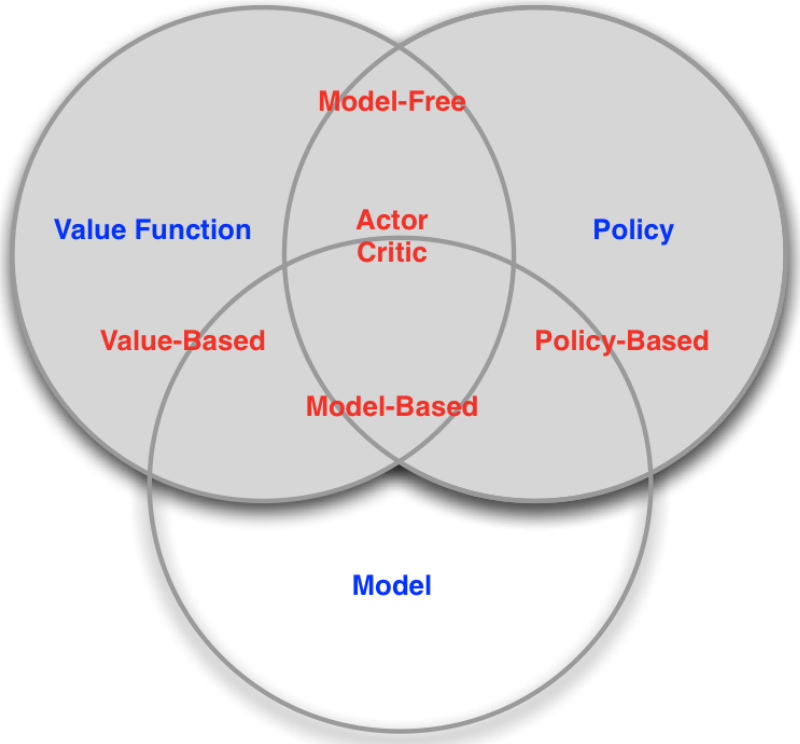
actor critic 均有

{ model free

无model 只根据 Policy or value function 行为 action 选择

model based

有 model



4.

Problems

1. Sequential decision making

① RL problem
- environment 开始 unknown
- 与 environment
- improve policy

② planning
- know environment
- 与 model (internal model: perfect interaction)
- improve policy

2. 对应比
Exploration vs Exploitation

try more info

use known info → reward

3. 对action

predict vs control

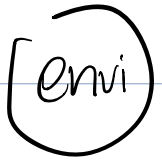
policy → future

predict value f

寻找 best policy

Model-Free

不理解环境



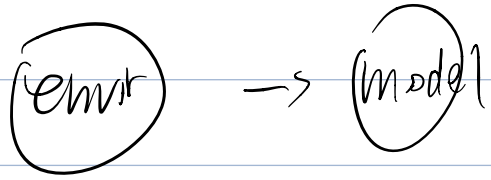
Q learning

Sarsa

Policy gradient

Model-Based

理解环境



建模 →

Q learning

Sarsa

Policy gradient

+ Imagination.

Policy Based

概率

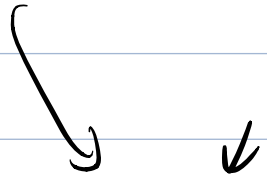
Policy gradient

Value-Based

价值

Q learning

Sarsa



actor critic

Monte-Carlo update

Temporal-

回合更新

单步更新

On-Policy

自己学自己实操

Sarsa

Sarsa(λ)

Off-Policy

自己

+ 观察

其他人

Q

Deep Q