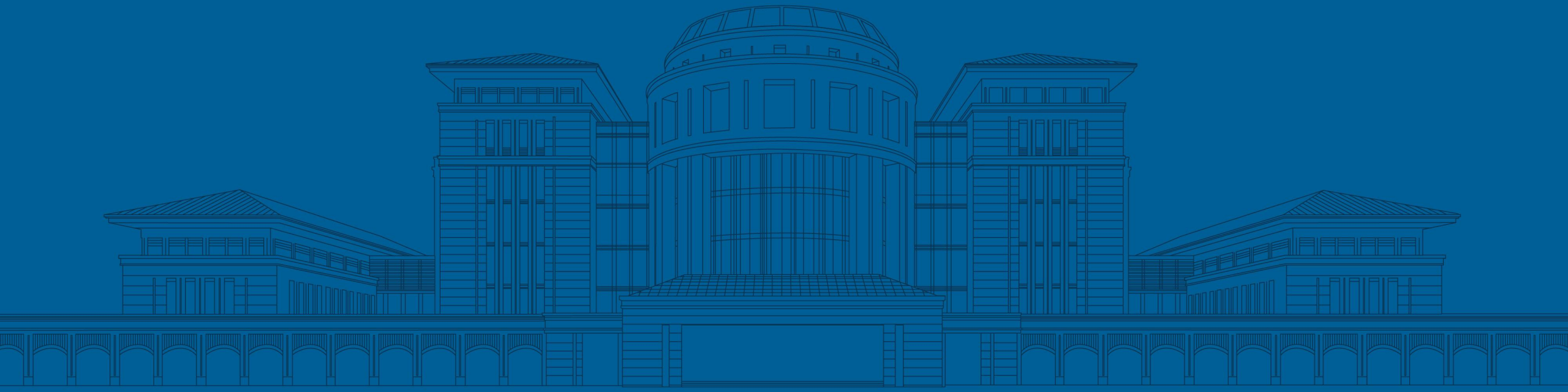
澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Group Report

speaker: Xiongyi Li

Dec 18th 2023

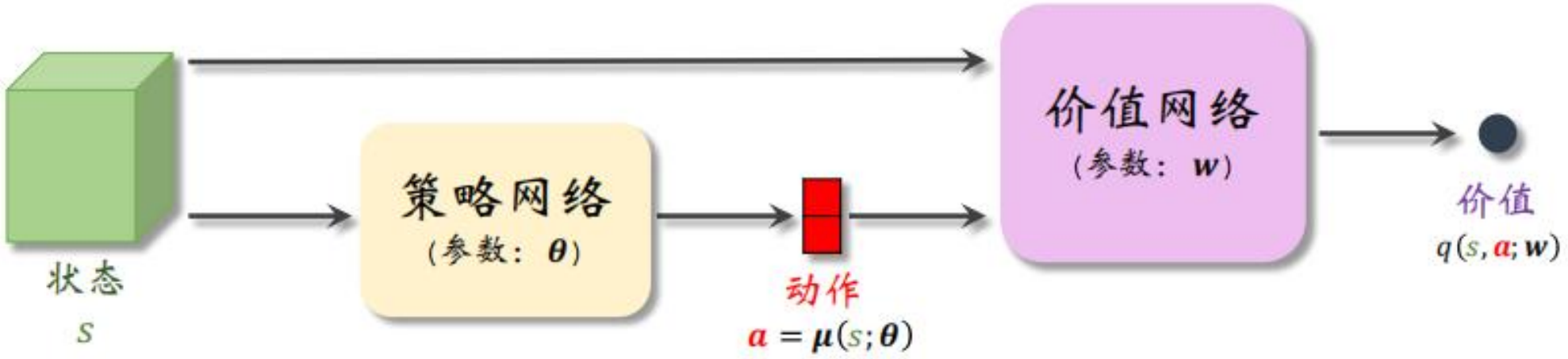


Deep Reinforcement Learning

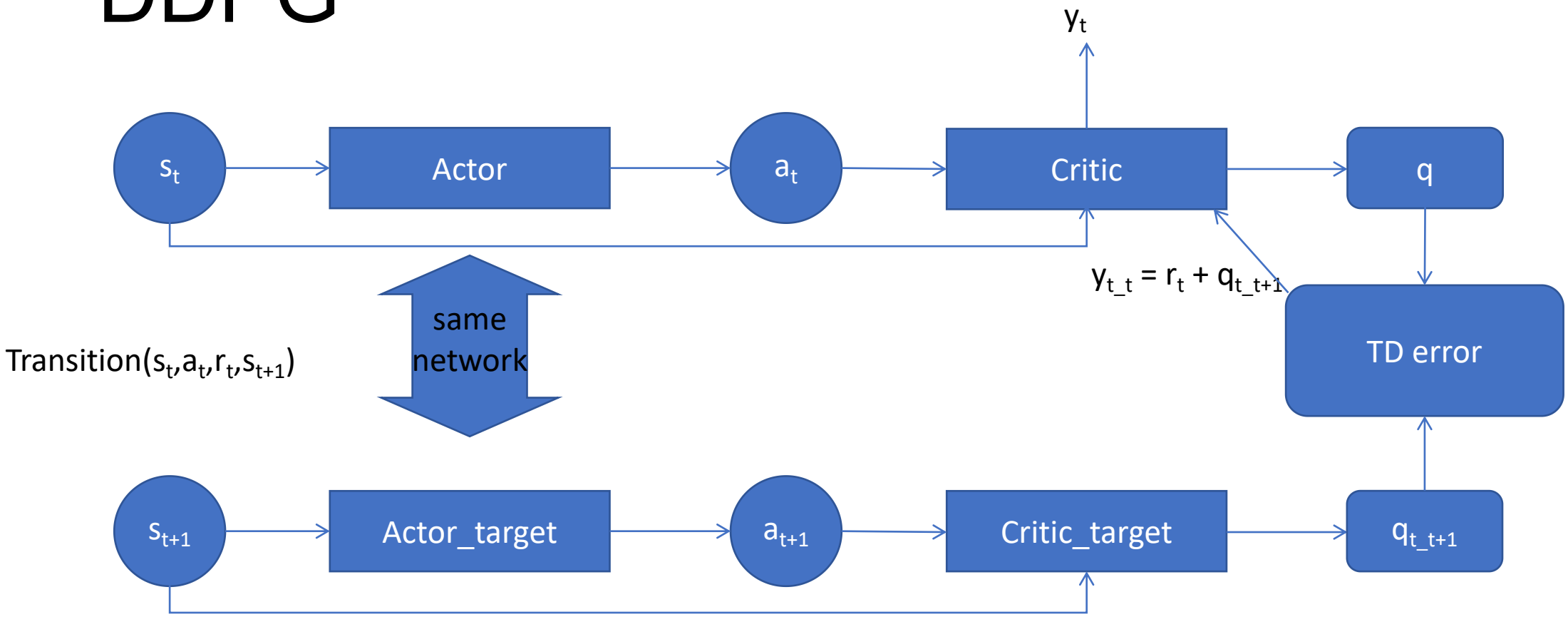
- DDPG
- PPO
- Gail

Deep Deterministic Policy Gradient(确定策略梯度)

- Use a deterministic policy network(actor):a = π(s; θ)
- Use a value network(critic): q = q(s, a ;w)
- q_{t+1} = q(s_{t+1},a'$_{t+1}$;w), a'$_{t+1}$ = $\pi(st_{+1;}\theta)$



DDPG

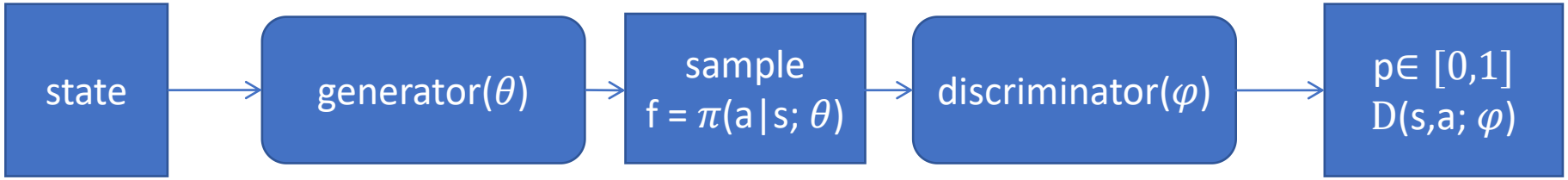

y_t
s_t
Actor
a_t
Critic
q
$y_{t\_t} = r_t + q_{t\_t+1}$
same network
Transition(s_t,a_t,r_t,s_{t+1})
TD error
s_{t+1}
Actor_target
a_{t+1}
Critic_target
q_{t_t+1}

Generative adversarial imitation learning

- Generator(生成器): produce fake sample to cheat discriminator
  - $\pi(a|s;\theta)$
  - input: state; output: $f = \pi(\cdot|s;\theta)$

Discriminator(判别器): determine real or generated

  - $D(s,a;\varphi)$
  - input: state; output: $p = D(s,\cdot|;\varphi)$



Gail

- Training:
- From training data , get τ^{real} = [s_1^{real},a_1^{real};s_2^{real},a_2^{real}...] length = m
- Use $\pi(a|s;\theta_{now})$, get τ^{fake} = [s_1^{fake},a_1^{fake};s_2^{fake},a_2^{fake}...] length = n
- Take u_t = lnD(s_t^{fake},a_t^{fake};φ), the bigger u_t is, the realer (s_t,a_t) will be
- Target: $L(\theta|\theta_{now}) = \frac{1}{n}\sum_{t=1}^{n}\frac{\pi(a_t|s_t;\theta)}{\pi(a_t|s_t;\theta_{now})} u_t$
- Update: θ_{new} = argmax(L)
- Loss: $F(\tau real, \tau fake;\varphi) = \frac{1}{m}\sum_{t=1}^{m} ln[1 - D(\tau^{real}, \tau^{fake};\varphi)] + \frac{1}{n}\sum_{t=1}^{n} ln[1 - D(\tau^{real}, \tau^{fake};\varphi)]$
- Update: $\varphi = \varphi - \beta\frac{\partial F}{\partial \varphi}$

Thank You!

Avenida da Universidade, Taipa, Macau, China
Tel : (853) 8822 8833 Fax : (853) 8822 8822
Email : mc35289@um.edu.mo Website : www.um.edu.mo