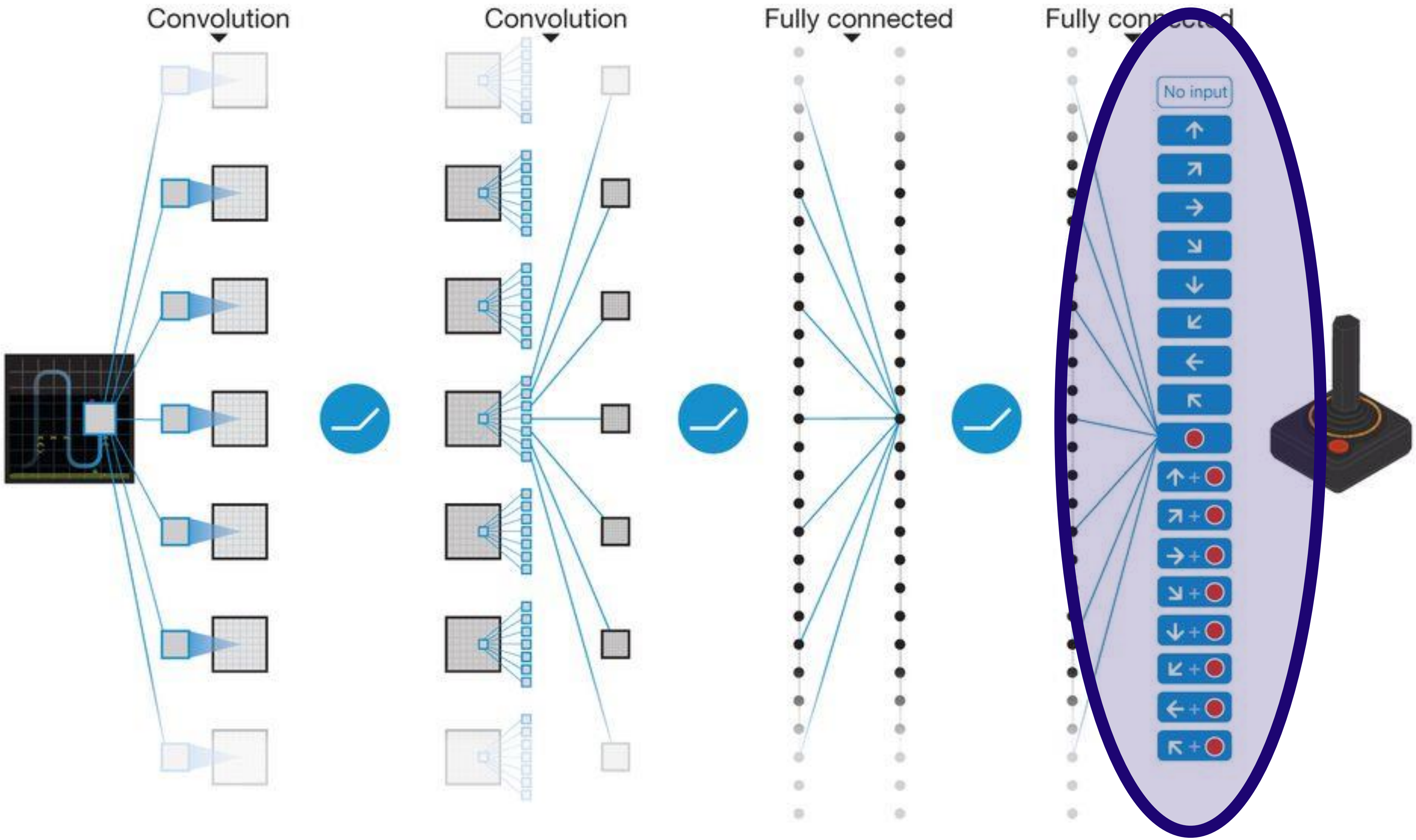
Continuous Control (Part 1)

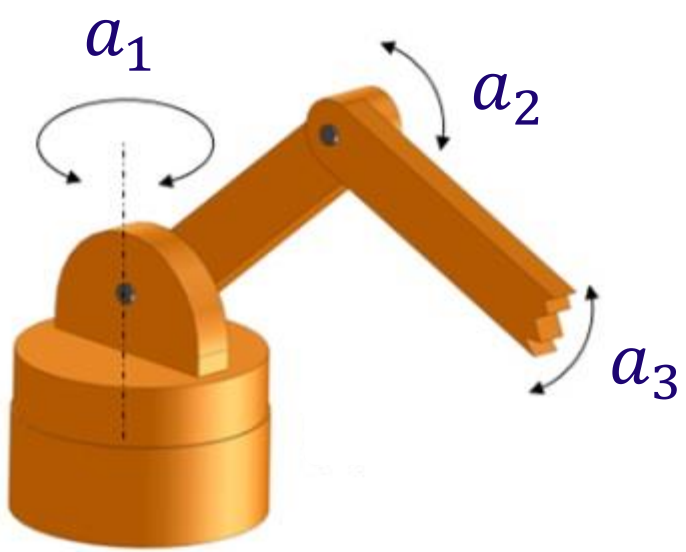
Continuous control with deep reinforcement learning
Timothy P. Lillicrap et. al. ICLR 2016.

Trust Region Policy Optimization
John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, Pieter Abbeel. ICML 2015.

Presented by
Mark Arnold



https://media.springernature.com/full/nature-static/assets/v1/image-assets/nature14236-f1.jpg

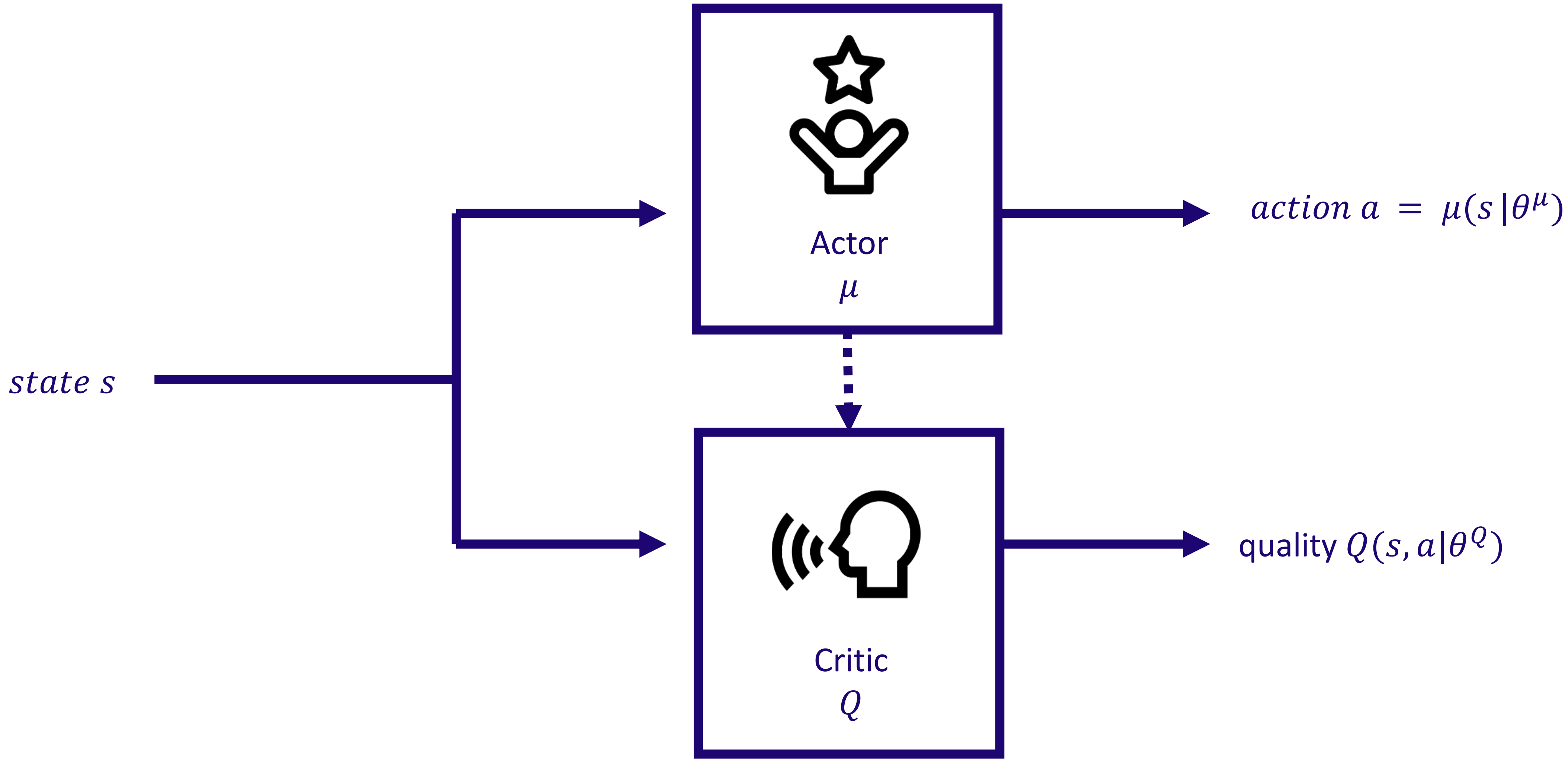


$a_i \in \{-k_2, -k_1, 0, k_1, k_2\}$ ➡ $5^3 = 125$

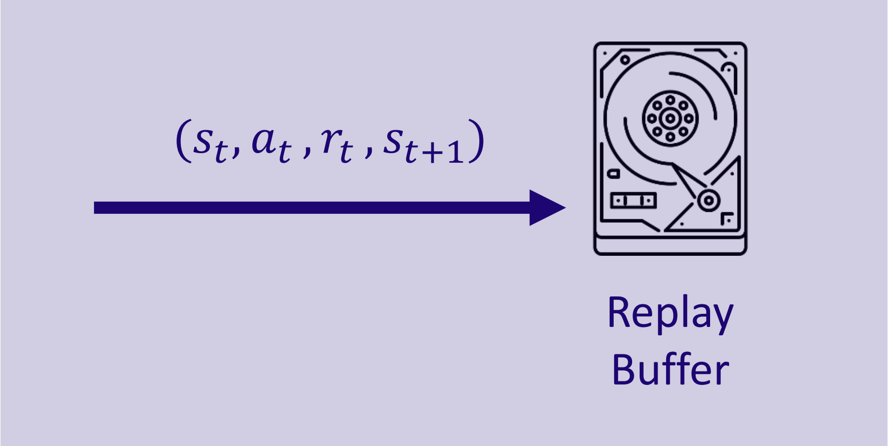
https://i.stack.imgur.com/lxlrc.jpg


BOOM..!

Deep Deterministic Policy Gradient (DDPG)

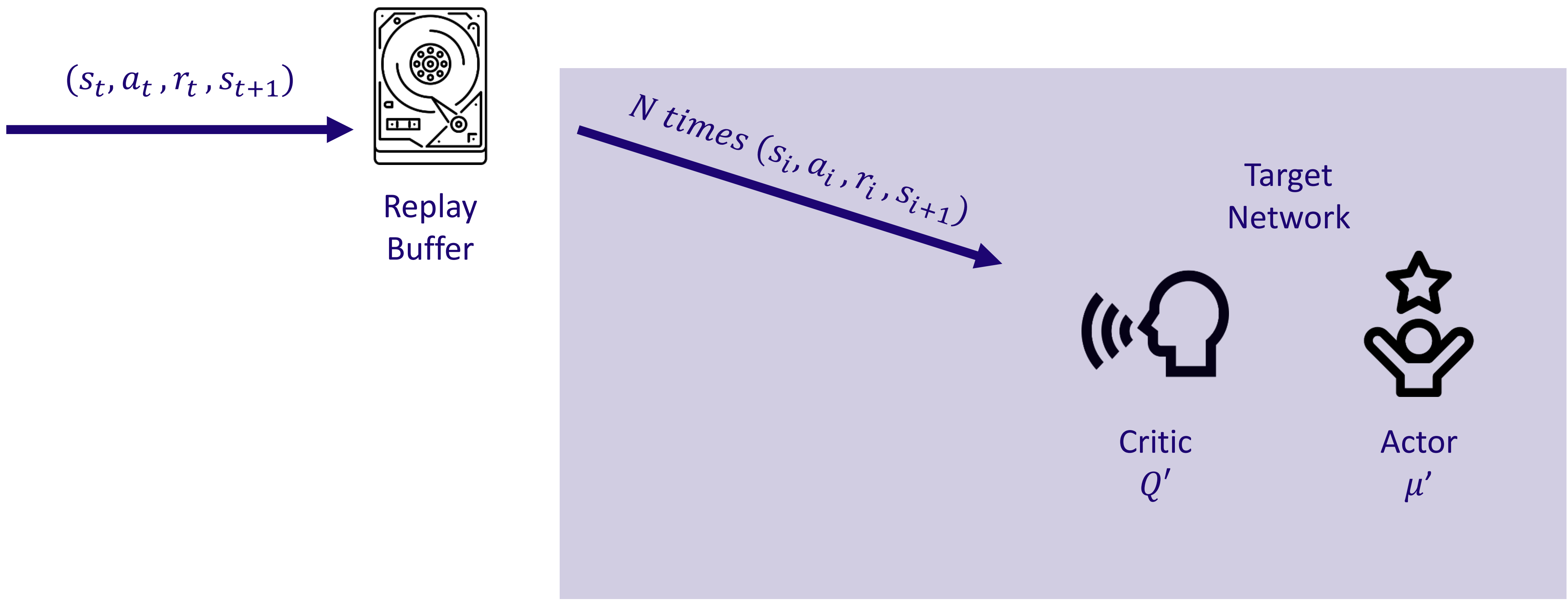


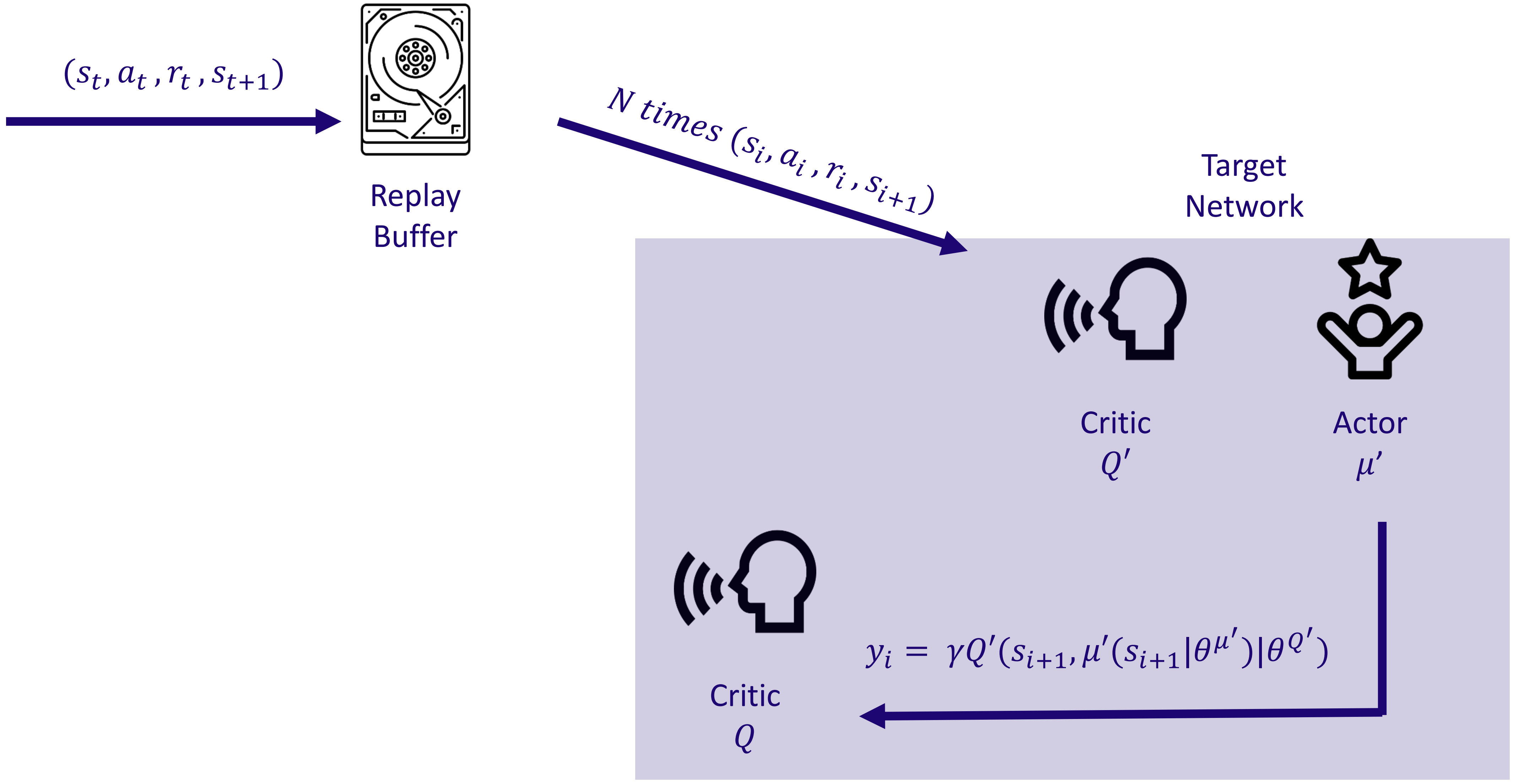
Icon 2 made by "https://www.freepik.com/" Icon3 made by "https://www.flaticon.com/authors/nikita-golubev"



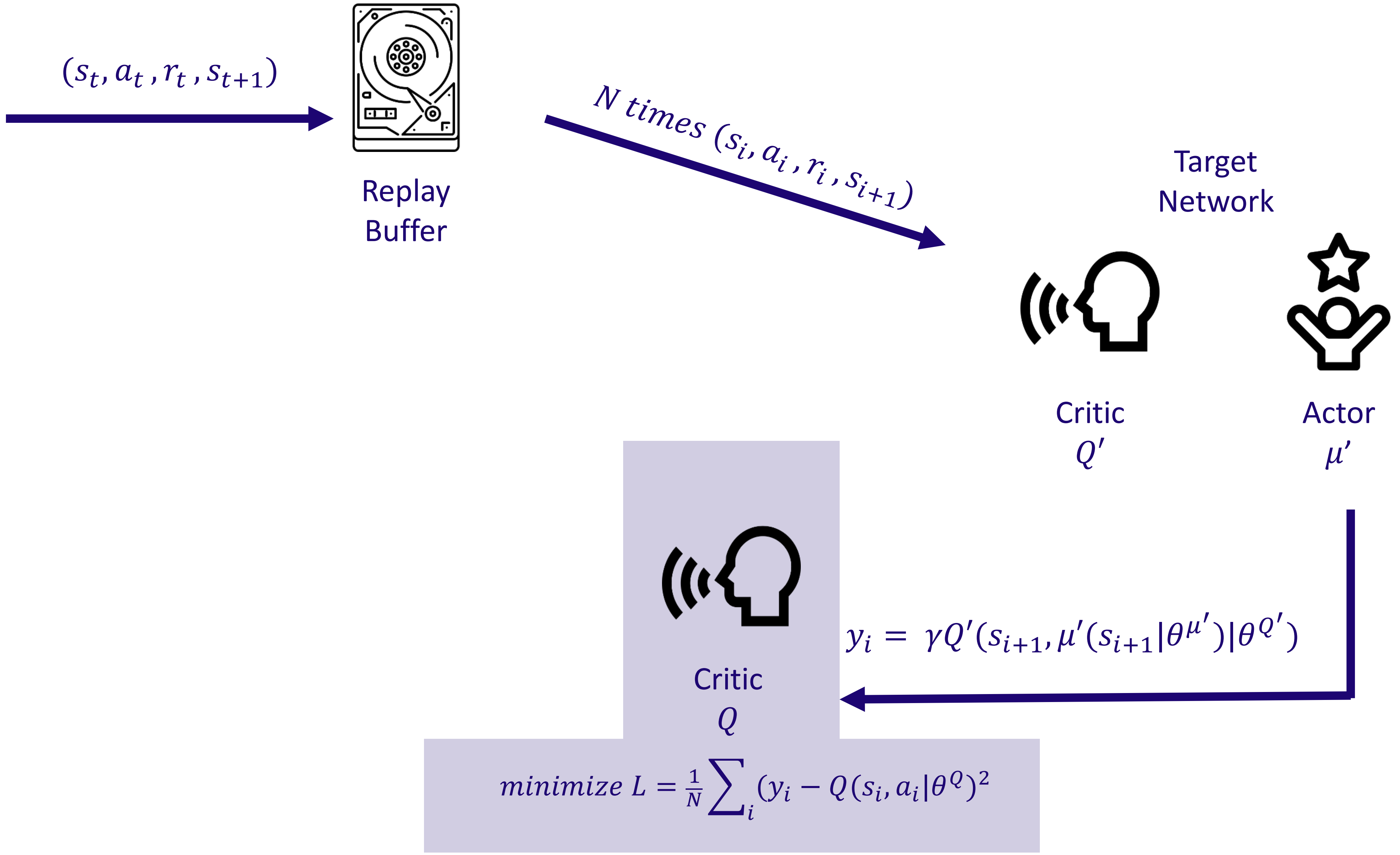
Icon 1 made by "https://www.flaticon.com/authors/chanut"

Icon 2 made by "https://www.freepik.com/" Icon3 made by "https://www.flaticon.com/authors/nikita-golubev"


(s_t, a_t, r_t, s_{t+1})
Replay
Buffer
N times (s_i, a_i, r_i, s_{i+1})
Target
Network
Critic
Q'
Actor
μ'

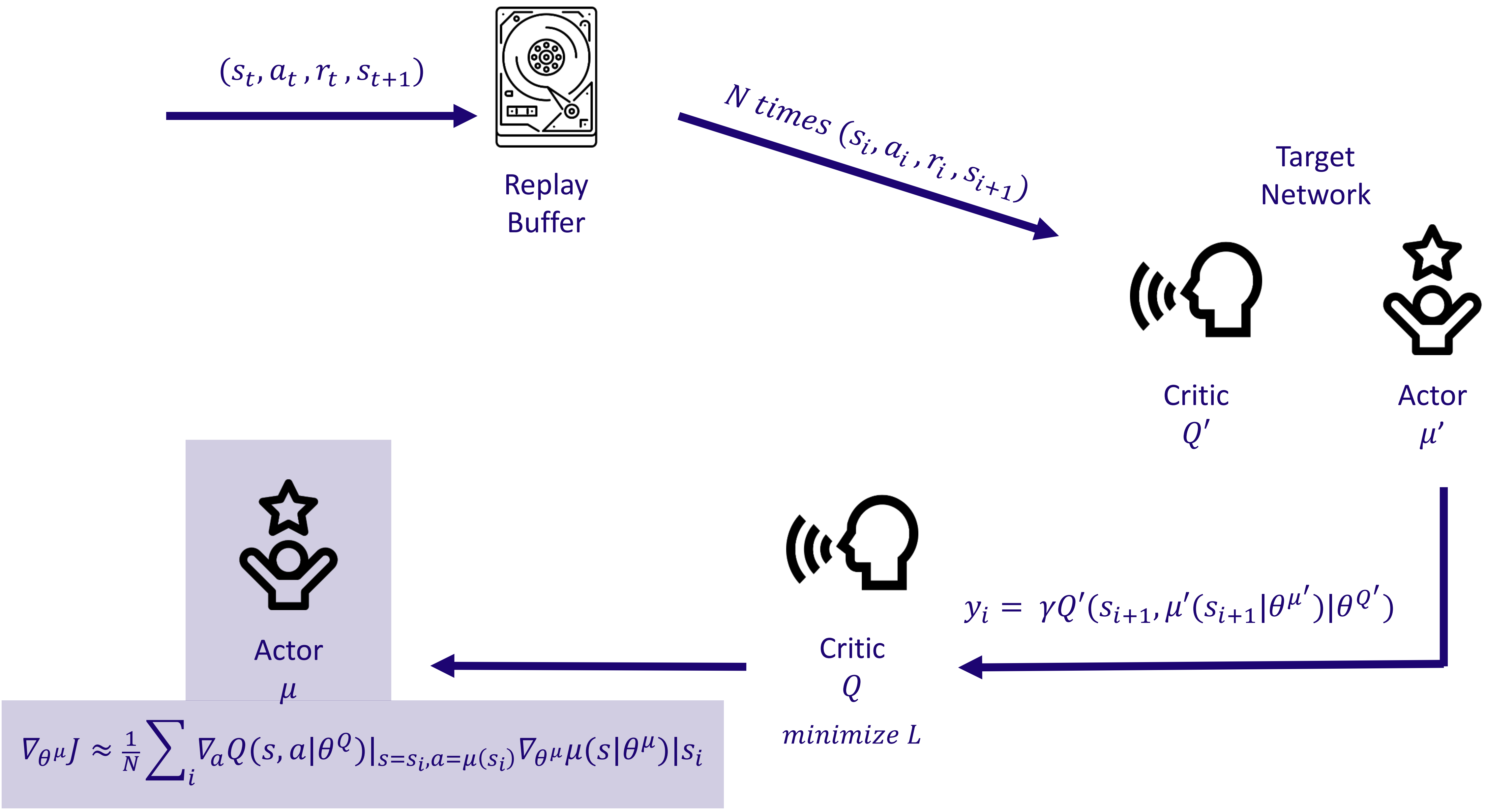


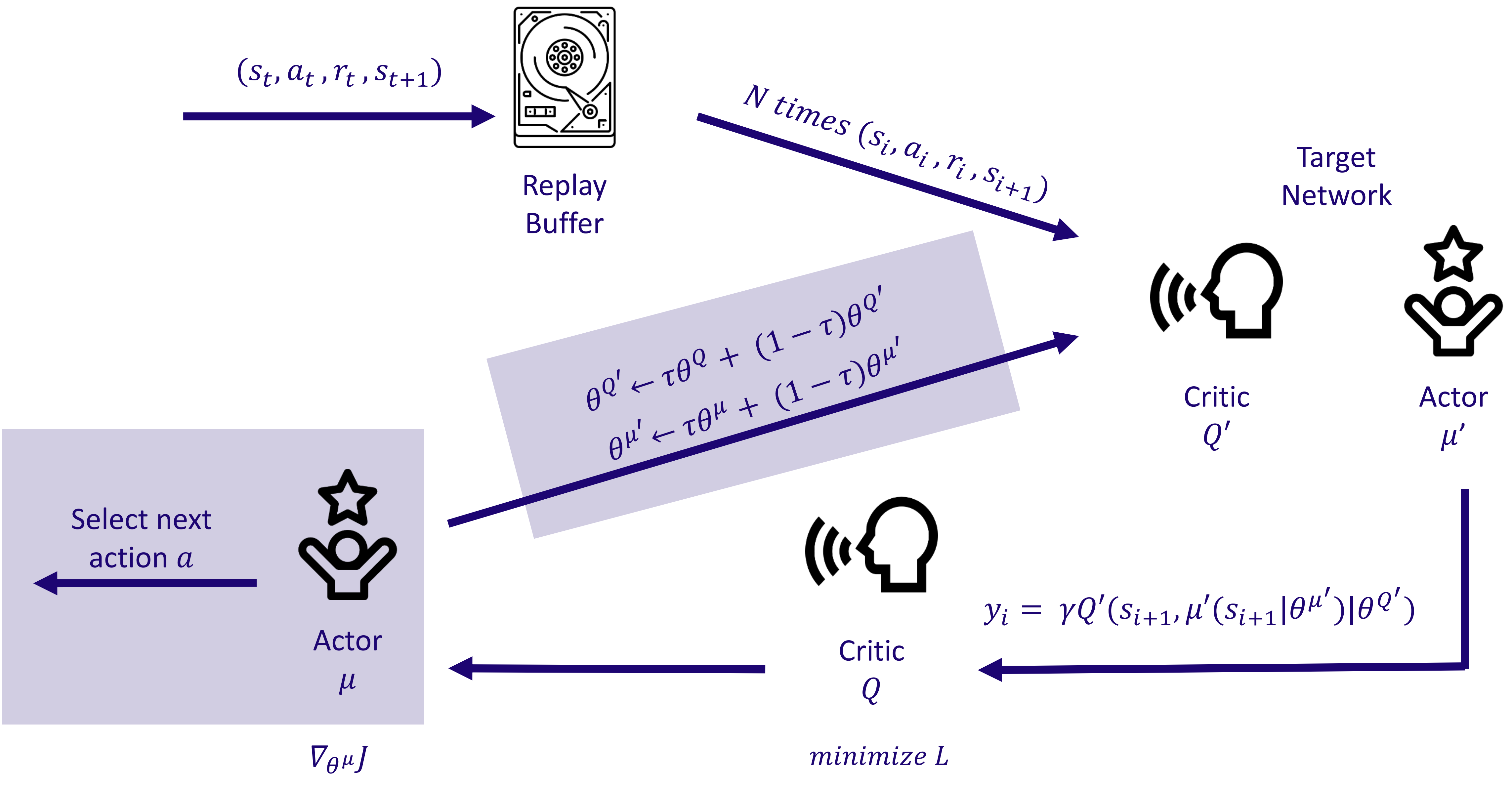
(s_t, a_t, r_t, s_{t+1})
Replay
Buffer
N times (s_i, a_i, r_i, s_{i+1})
Target
Network
Critic
Q'
Actor
μ'
Critic
Q
$y_i = \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$



Icon 1 made by "https://www.flaticon.com/authors/chanut"

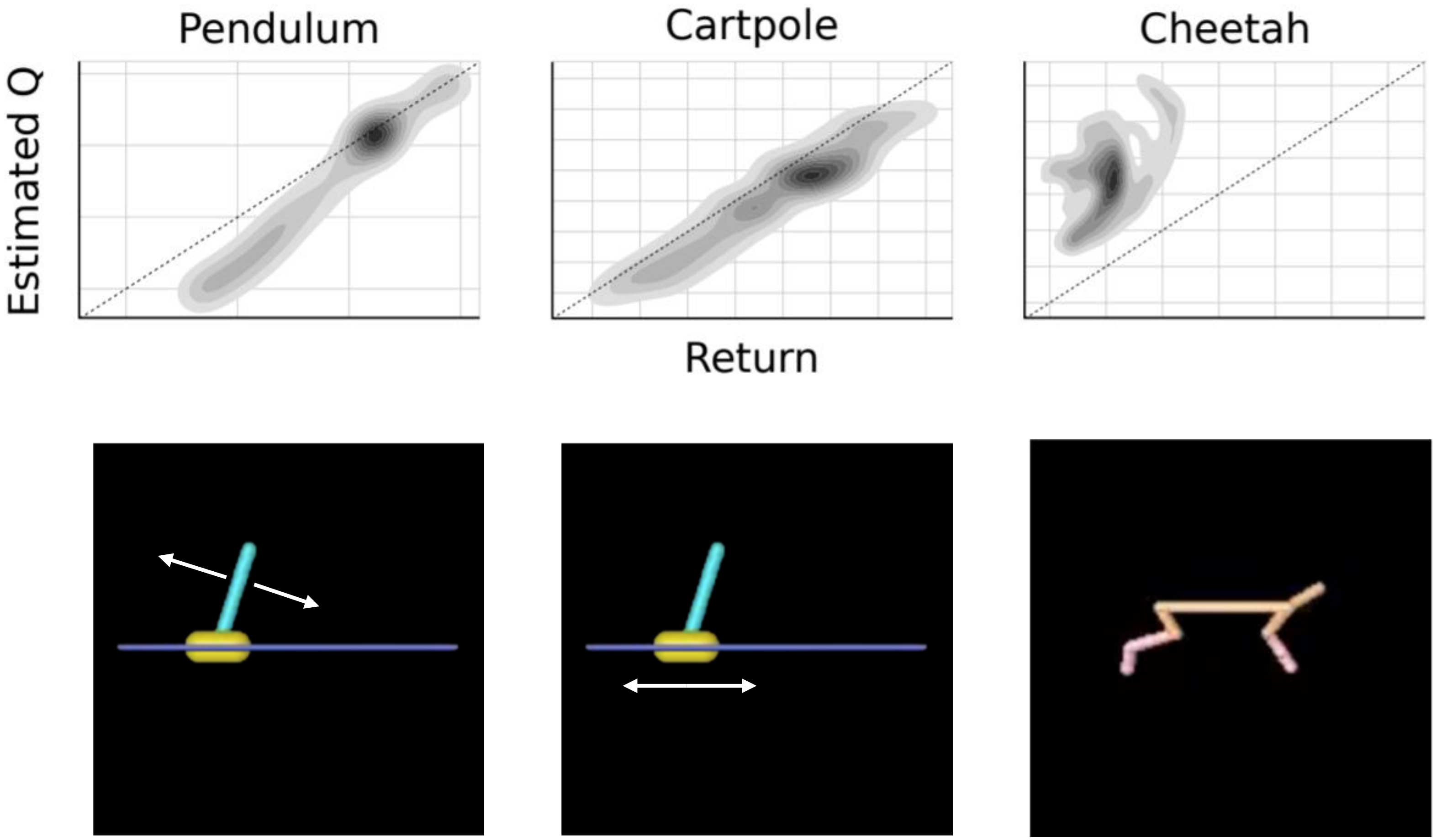
Icon 2 made by "https://www.freepik.com/" Icon3 made by "https://www.flaticon.com/authors/nikita-golubev"

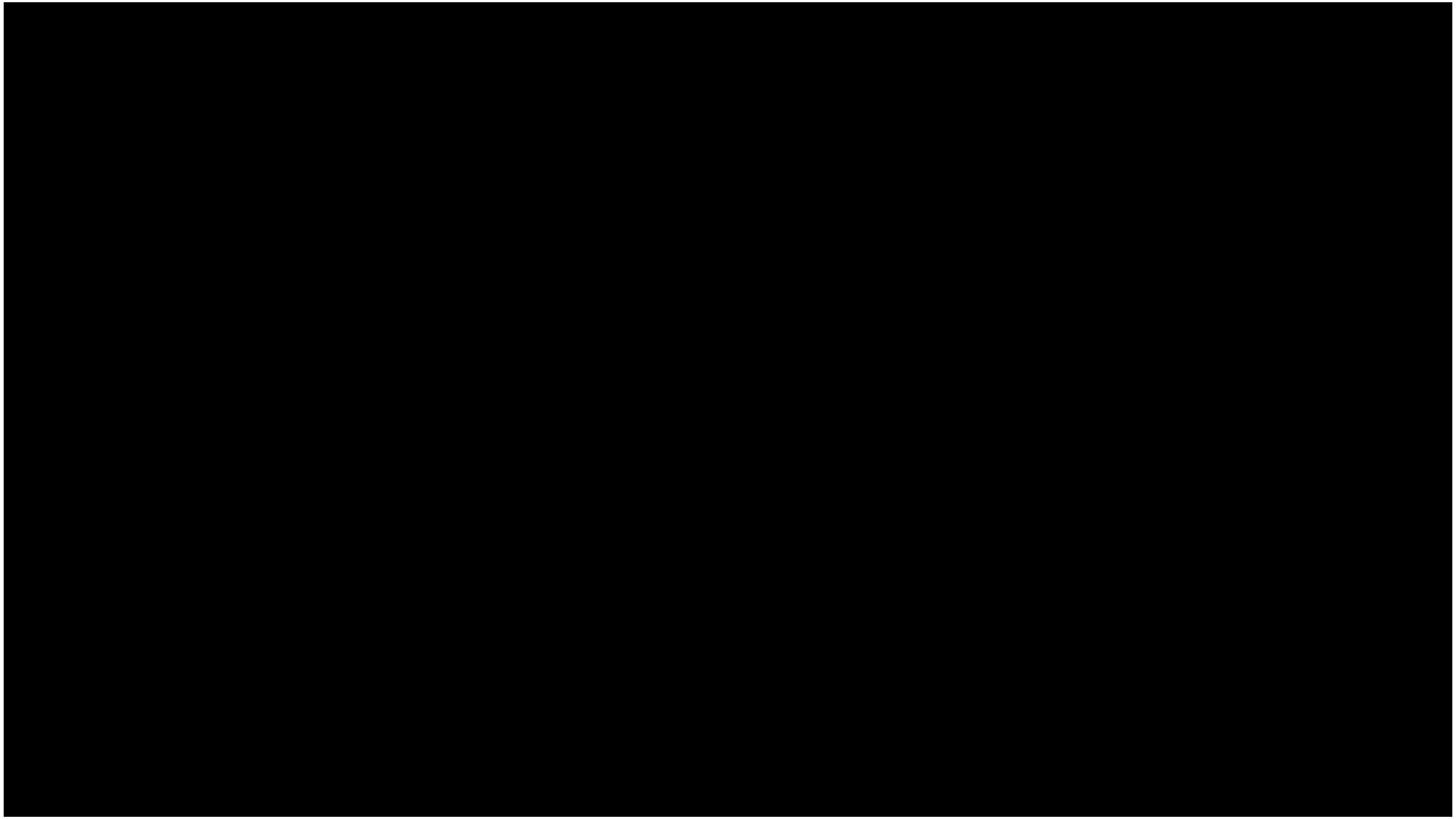

(s_t, a_t, r_t, s_{t+1})
Replay
Buffer
N times (s_i, a_i, r_i, s_{i+1})
Target
Network
Critic
Q'
Actor
μ'
$y_i = \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
Critic
Q
minimize L
Actor
μ
$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$



Icon 1 made by "https://www.flaticon.com/authors/chanut"

Icon 2 made by "https://www.freepik.com/" Icon3 made by "https://www.flaticon.com/authors/nikita-golubev"


Pendulum
Cartpole
Cheetah
Estimated Q
Return



Trust Region Policy Optimization (TRPO)

$$A_{\pi}(s,a) = Q_{\pi}(s,a) - V_{\pi}(a)$$

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(a)$$

$$\eta(\pi) = \mathrm{E}_{s_0,a_0,\ldots}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$$

$$where\ s_0 \sim \rho_0(s_0), a_t \sim \pi(a_t|s_t),$$
$$s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(a)$$

$$\eta(\pi) = \mathrm{E}_{s_0,a_0,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right]$$

$$where\ s_0 \sim \rho_0(s_0), a_t \sim \pi(a_t|s_t),$$
$$s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathrm{E}_{s_0,a_0,\cdot\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right]$$

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(a)$$

$$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \ldots$$

$$\eta(\pi) = \mathrm{E}_{s_0,a_0,\ldots}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$$

$$where\ s_0 \sim \rho_0(s_0), a_t \sim \pi(a_t|s_t),$$
$$s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathrm{E}_{s_0,a_0,\cdot\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t,a_t)\right]$$

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(a)$$

$$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_0 = s) + \ldots$$

$$\eta(\pi) = \mathrm{E}_{s_0, a_0, \ldots}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$$

$$where\ s_0 \sim \rho_0(s_0), a_t \sim \pi(a_t|s_t),$$
$$s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathrm{E}_{s_0, a_0, \cdot \sim \tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right]$$

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(a)$$

$$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_0 = s) + \ldots$$

$$\eta(\pi) = \mathrm{E}_{s_0, a_0, \ldots} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) \right]$$

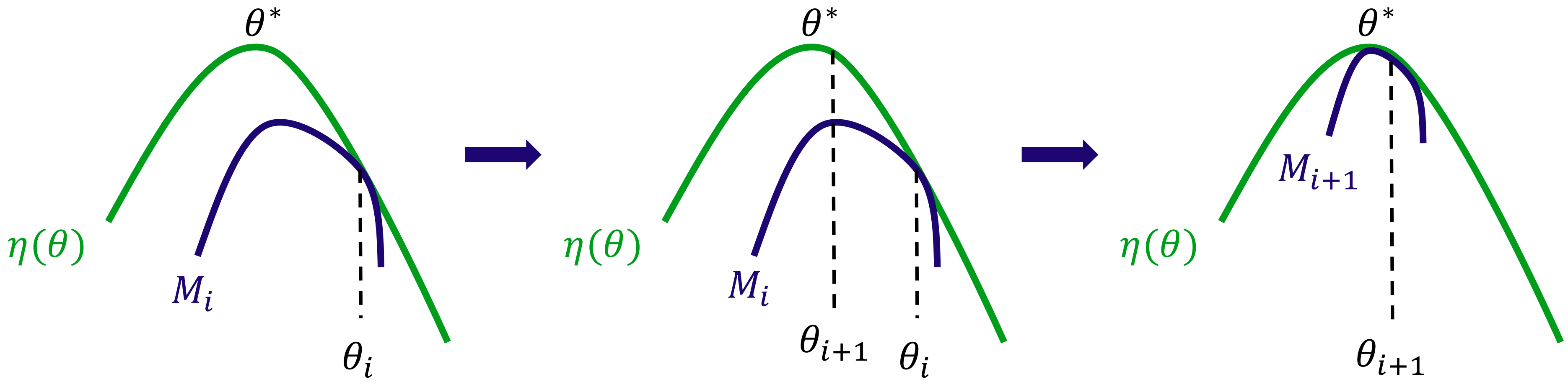
$$where\ s_0 \sim \rho_0(s_0), a_t \sim \pi(a_t|s_t), \\ s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

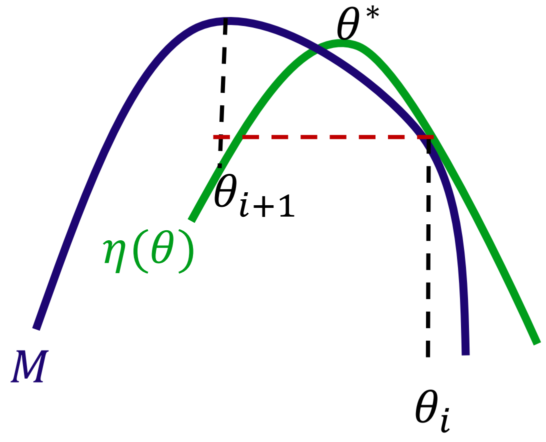
$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathrm{E}_{s_0, a_0, \cdot \sim \tilde{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t) \right]$$

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

Minorize-Maximization

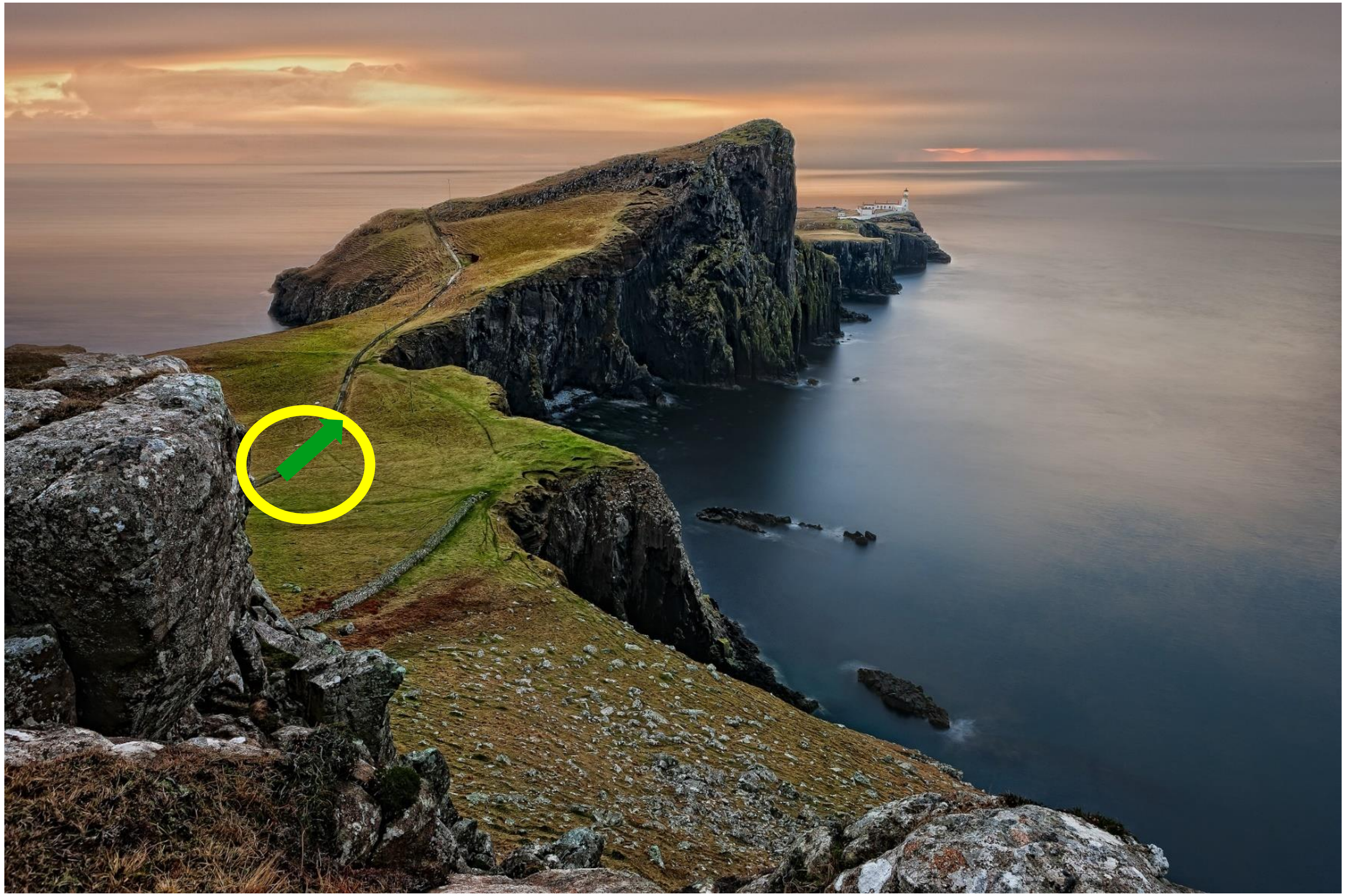


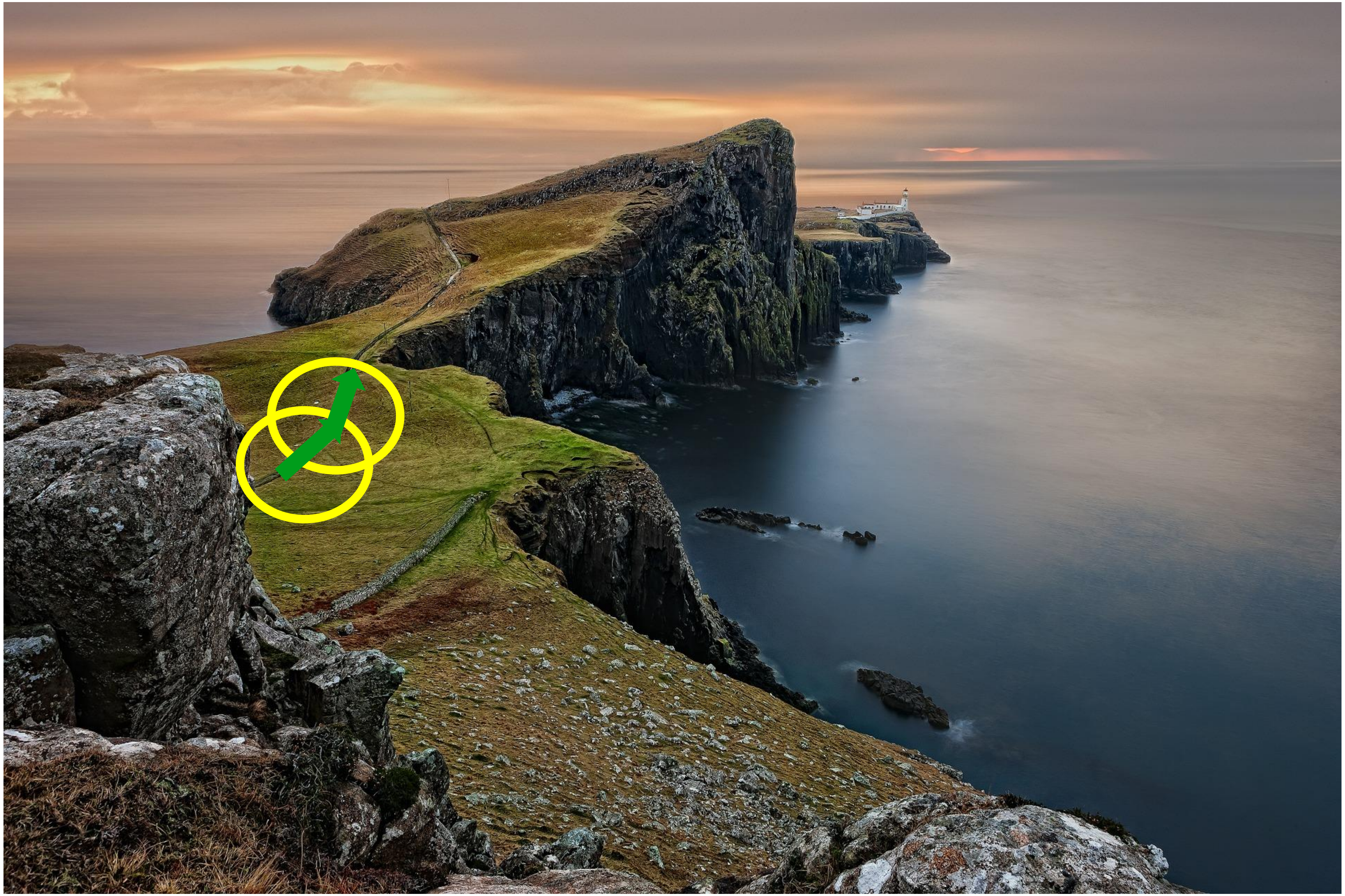


$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

$$M_i(\pi) = L_{\pi_i}(\pi) - C D_{KL}^{max}(\pi_i, \pi)$$

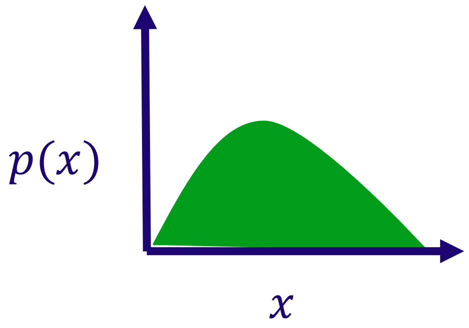
Trust Region

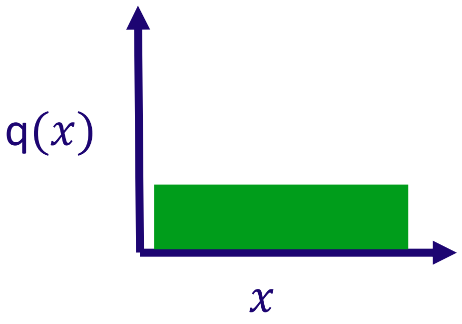




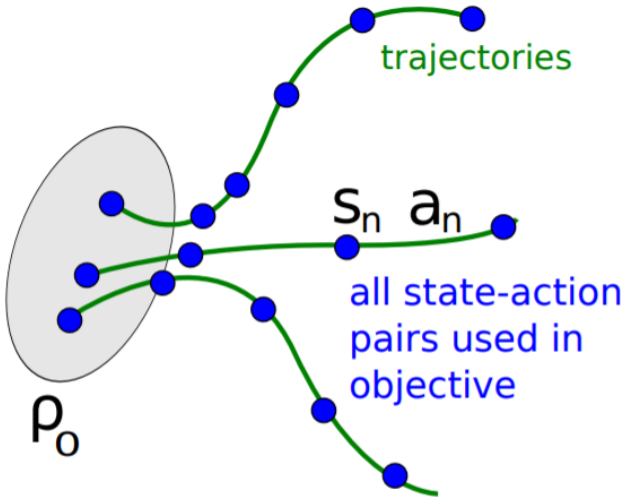
$$\underset{\theta}{\text{maximize}}\, L_{\theta_{old}}(\theta)$$
$$\textit{subject to } \overline{D}_{KL}^{\rho\theta_{old}}(\theta_{old}, \theta) \leq \delta$$

Importance Sampling





$$\mathrm{E}_q[f^*(x)] = \mathrm{E}_q\left[\frac{p(x)}{q(x)} f(x)\right] = \sum_x q(x)\frac{p(x)}{q(x)} f(x) = \sum_x p(x) f(x) = E_p[f(x)]$$



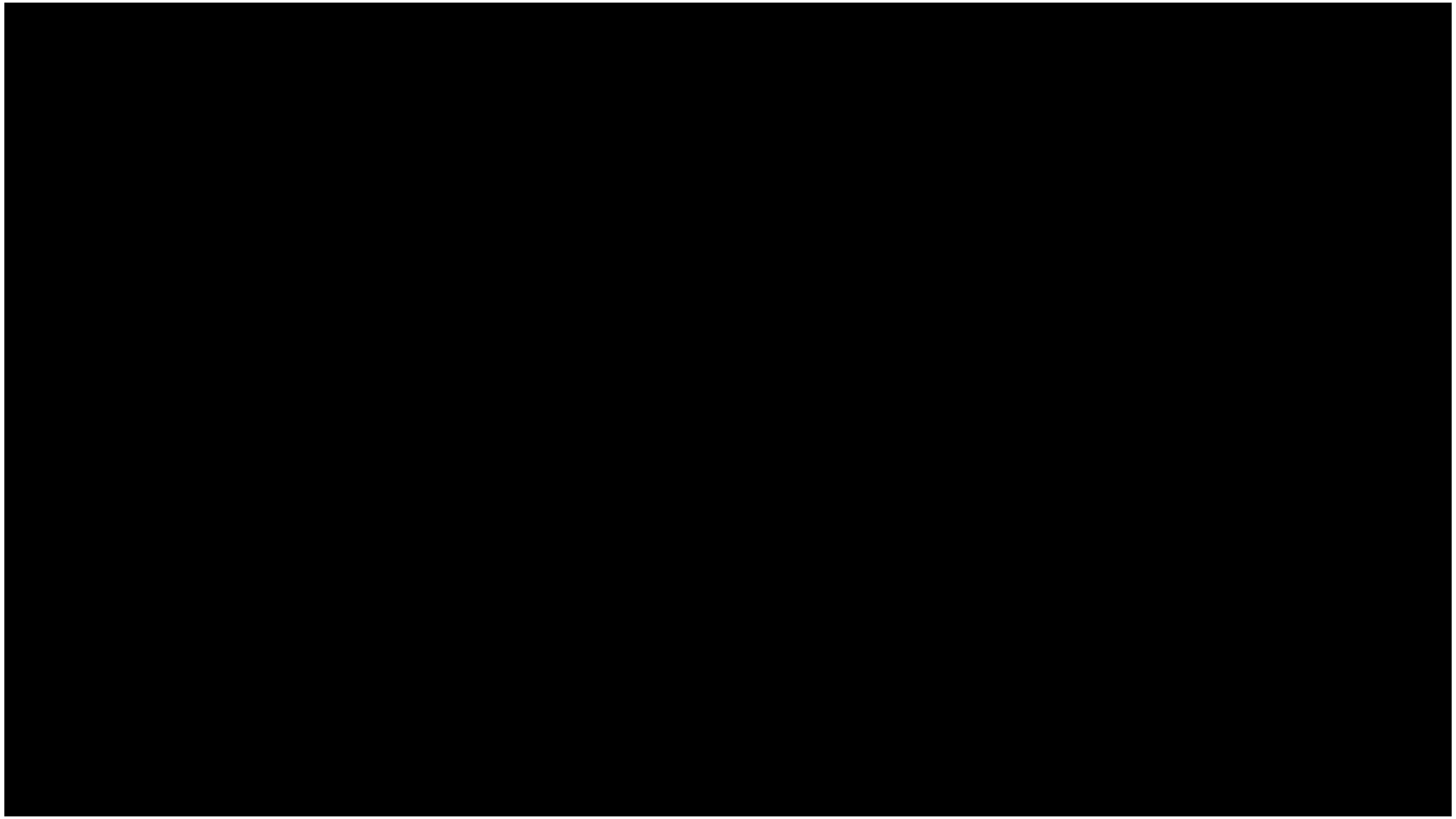
single path procedure

$$\underset{\theta}{\text{maximize}}\ \mathrm{E}_{s\sim\rho_{\theta_{old}},a\sim q}\left[\frac{\pi_\theta(a|s)}{q(a|s)}Q_{\theta_{old}}(s,a)\right]$$

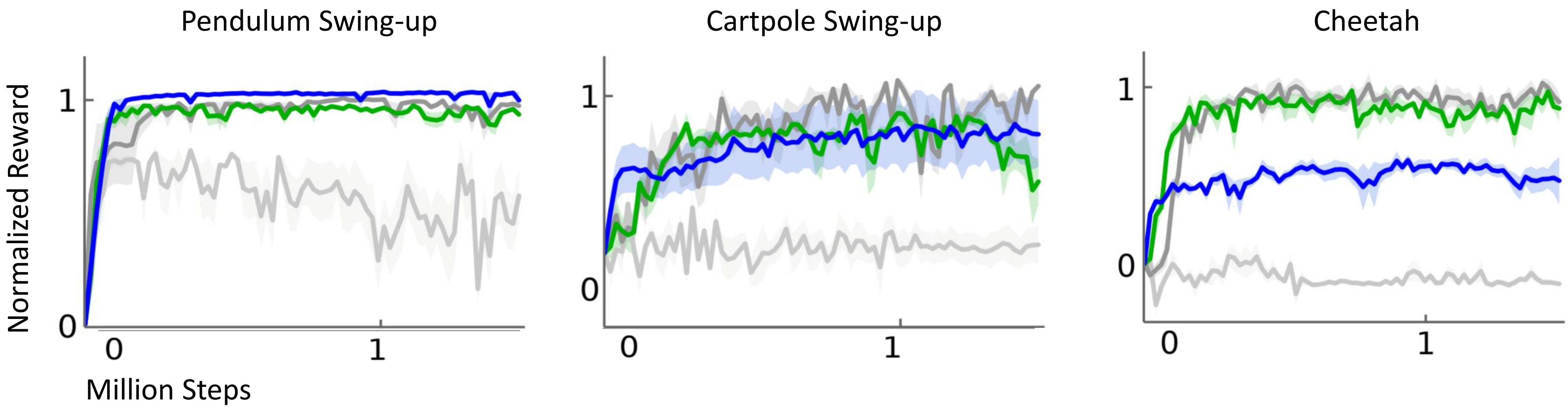
$$\textit{subject to}\ \mathrm{E}_{s\sim\rho_{\theta_{old}}}\left[D_{KL}(\pi_{\theta_{old}}(\cdot|s)\parallel\pi_\theta(\cdot|s))\right]\le\delta$$

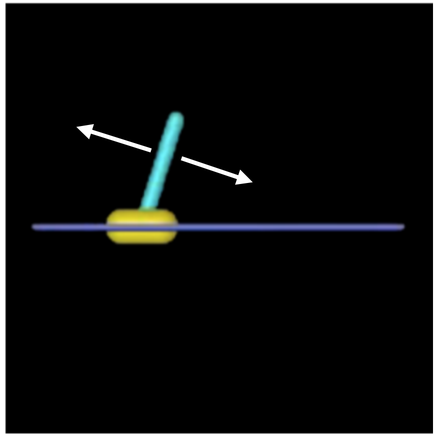
Solve by Approximation
(Fisher Information Matrix)

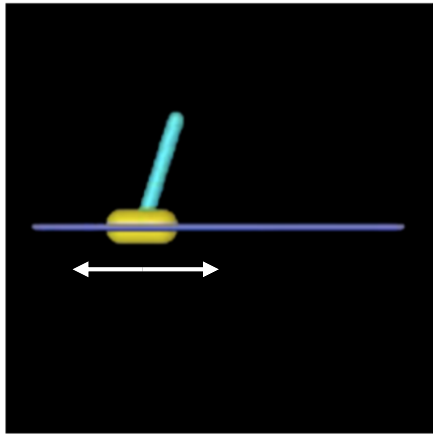
Results

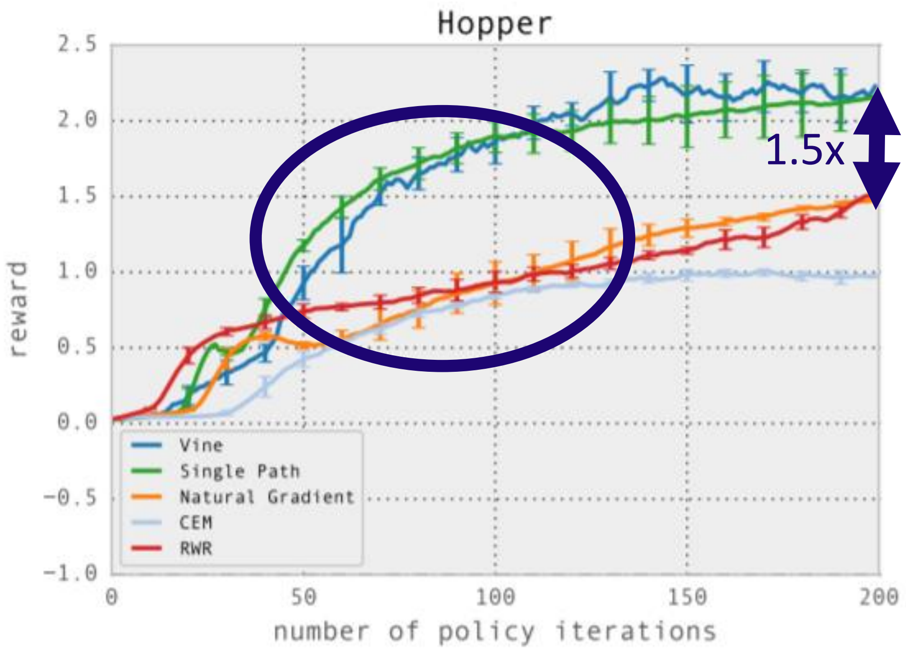


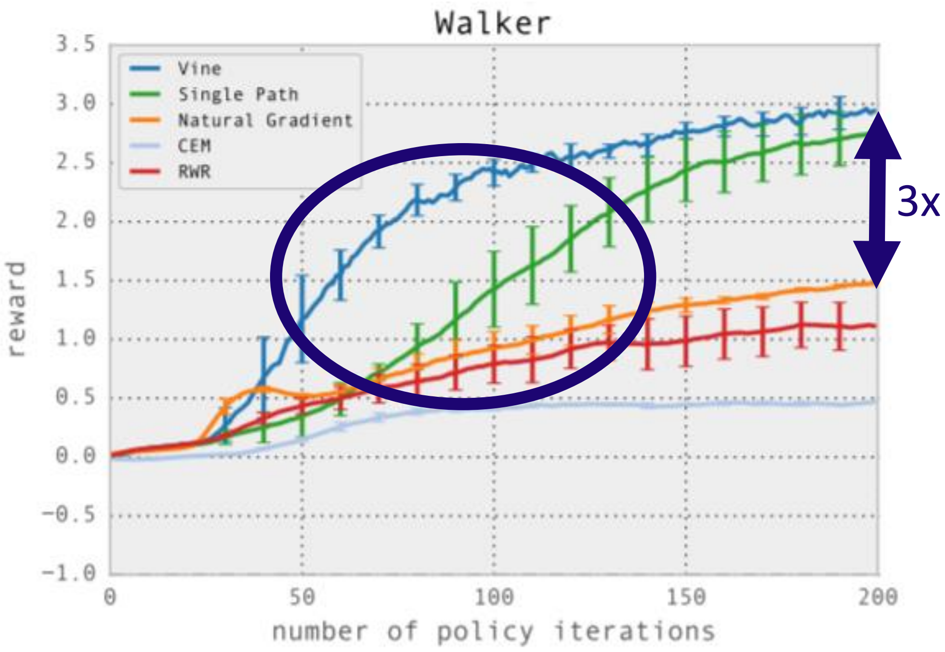
(light grey) – minibatch NFQCA , (grey) – DDPG, (green) – DDPG minibatch, (blue) – DDPG pixel only inputs

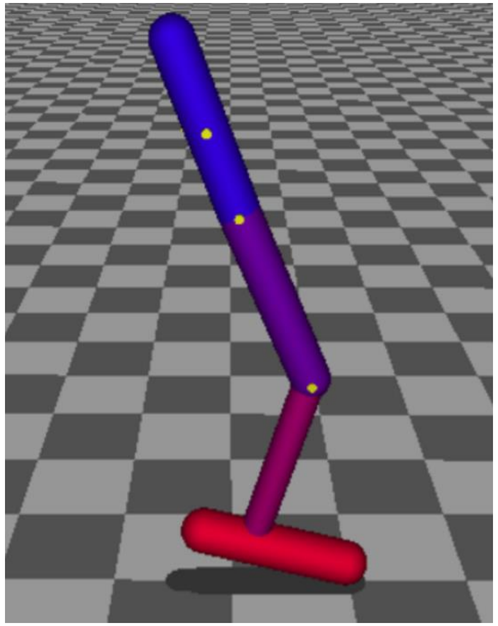


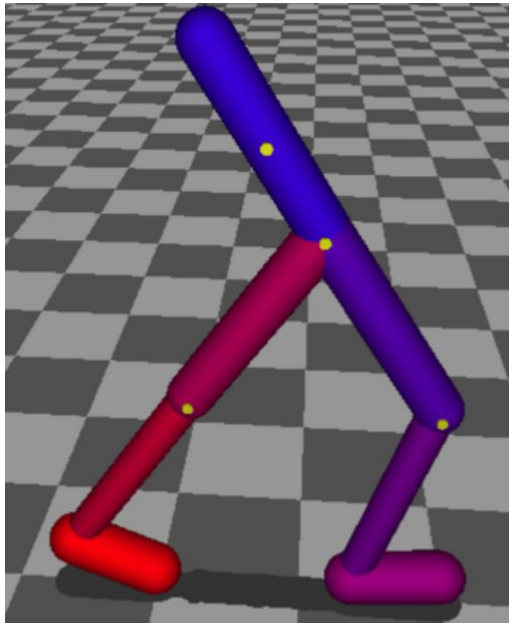




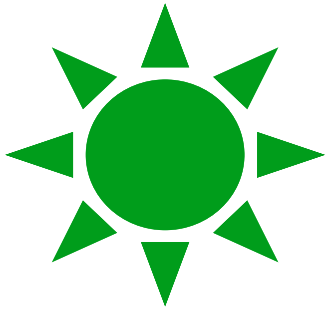

Hopper
reward
number of policy iterations
Vine
Single Path
Natural Gradient
CEM
RWR
1.5x


Walker
reward
number of policy iterations
Vine
Single Path
Natural Gradient
CEM
RWR
3x

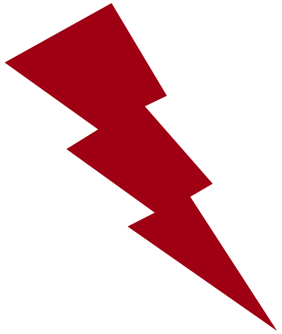




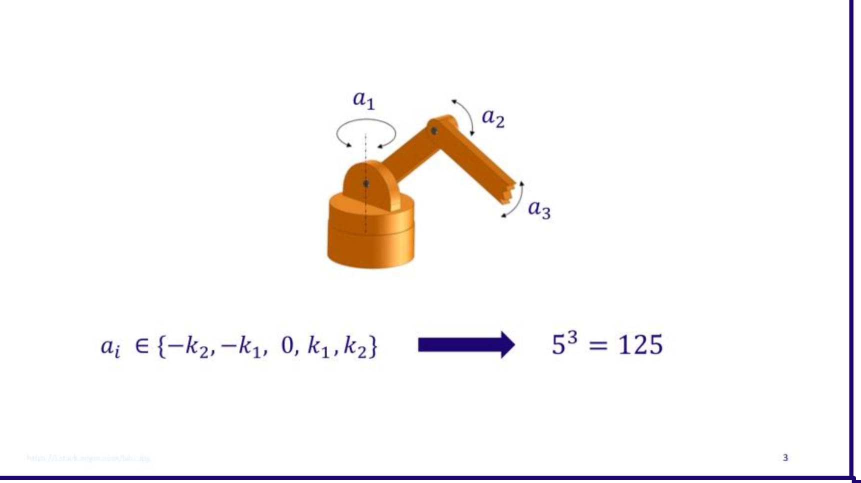
Conclusion

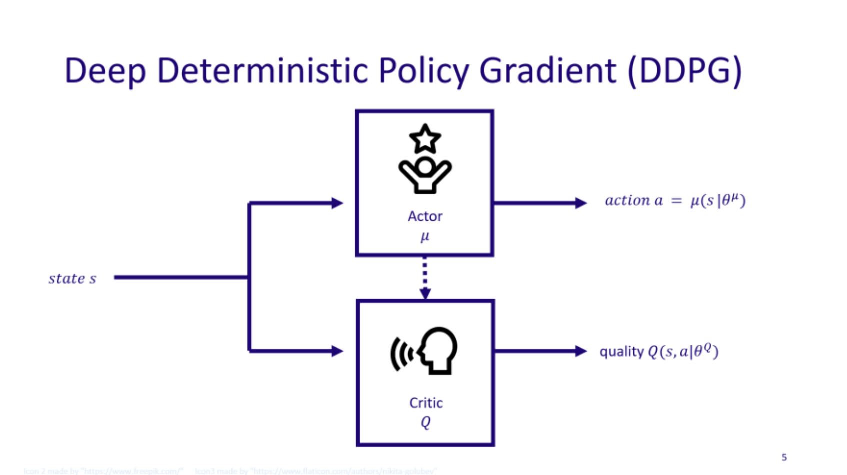


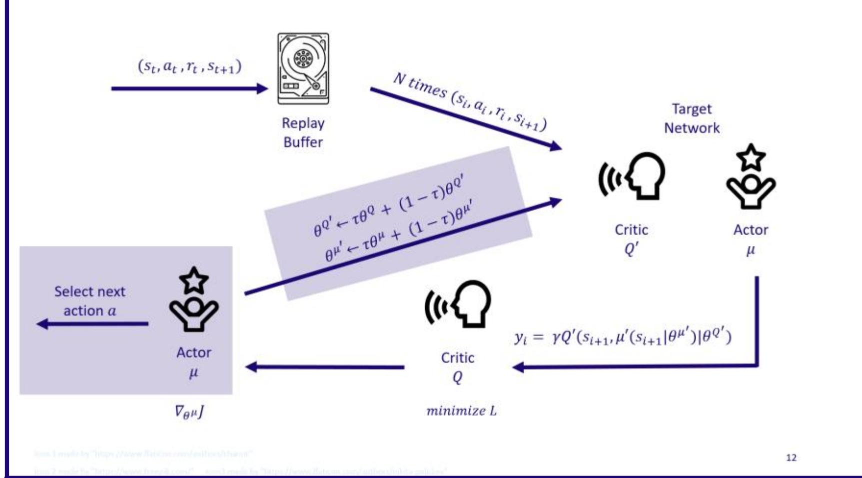
- General approaches
- Basis for further research activities

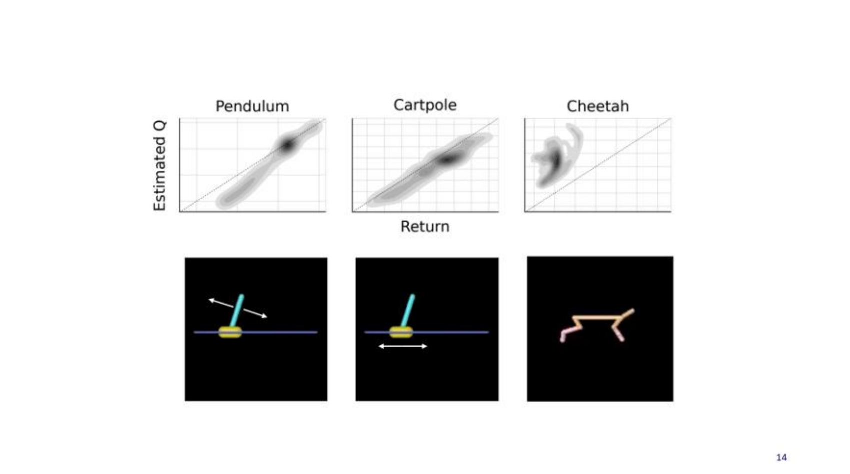


- DDPG → reliability on step size
- TRPO → issues with scalability (computation)

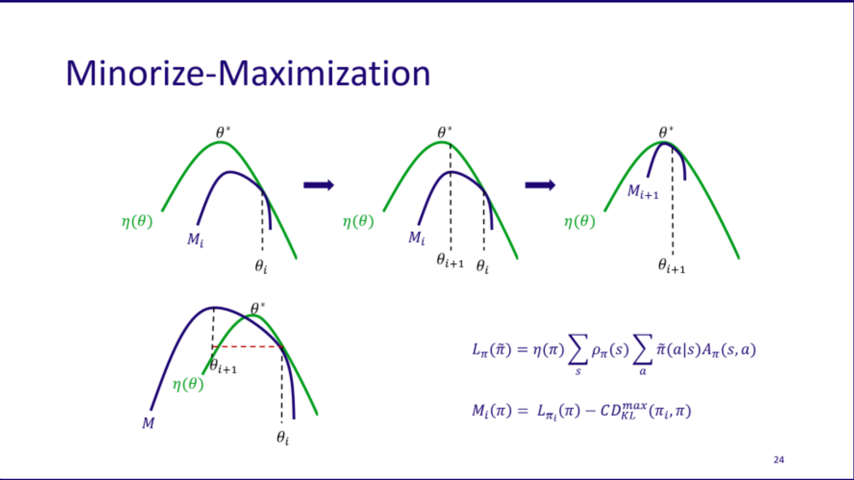

a_1
a_2
a_3
$a_i \in \{-k_2, -k_1, 0, k_1, k_2\}$
$5^3 = 125$
3


Deep Deterministic Policy Gradient (DDPG)
state s
Actor μ
Critic Q
action $a = \mu(s|\theta^\mu)$
quality $Q(s,a|\theta^Q)$
5

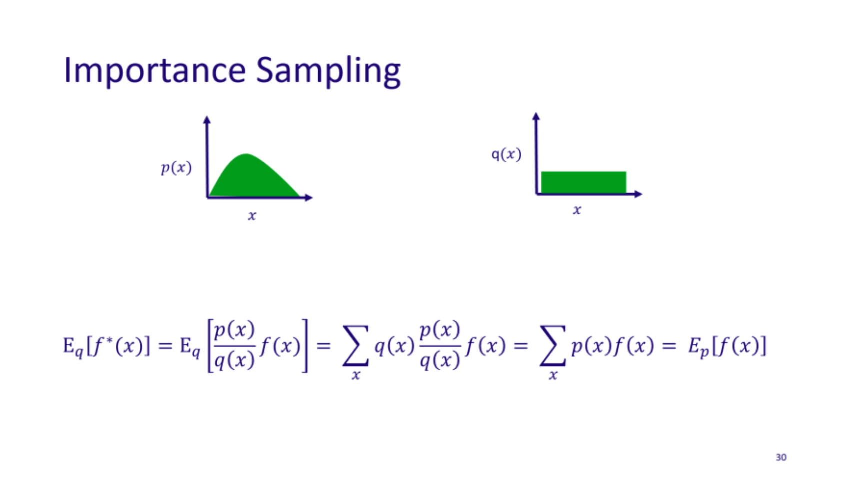

(s_t, a_t, r_t, s_{t+1})
Replay Buffer
N times (s_i, a_i, r_i, s_{i+1})
Target Network
Critic Q'
Actor μ
$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$
$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$
Select next action a
Actor μ
Critic Q
$y_i = \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
$\nabla_{\theta^\mu} J$
minimize L
12

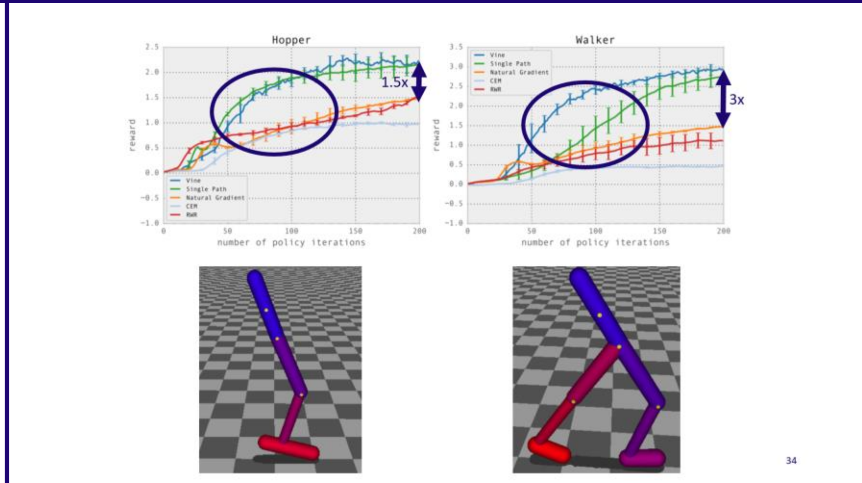

Pendulum
Cartpole
Cheetah
Estimated Q
Return
14

? Questions ?


Minorize-Maximization
θ^*
$\eta(\theta)$
M_i
θ_i
θ_{i+1}
M_{i+1}
M
$L_\pi(\tilde{\pi}) = \eta(\pi) \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$
$M_i(\pi) = L_{\pi_i}(\pi) - C D_{KL}^{max}(\pi_i, \pi)$
24


$\underset{\theta}{\text{maximize}}\ L_{\theta_{old}}(\theta)$
subject to $\bar{D}_{KL}^{\rho\theta_{old}}(\theta_{old}, \theta) \le \delta$
29


Importance Sampling
$p(x)$
$q(x)$
x
$E_q[f^*(x)] = E_q\left[\frac{p(x)}{q(x)} f(x)\right] = \sum_x q(x) \frac{p(x)}{q(x)} f(x) = \sum_x p(x) f(x) = E_p[f(x)]$
30


Hopper
Walker
1.5x
3x
reward
number of policy iterations
Vine
Single Path
Natural Gradient
CEM
RWR
34