

Actor-Critic Methods

Chen-Yu Wei

Review: Full-Information Policy Learning in MDPs

$$\theta_{k+1} = \operatorname{argmax}_{\theta} \underbrace{\left(V^{\pi_{\theta}}(\rho) - V^{\pi_{\theta_k}}(\rho) - \frac{1}{\eta} D(\theta, \theta_k) \right)}$$

$$\approx \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) (\pi_{\theta}(a|s) - \pi_{\theta_k}(a|s)) Q^{\pi_{\theta_k}}(s, a) = \mathbb{E}_{(s_i, a_i) \sim \pi_{\theta_k}} \left[\frac{\pi_{\theta}(a_i|s_i) - \pi_{\theta_k}(a_i|s_i)}{\pi_{\theta_k}(a_i|s_i)} \boxed{Q^{\pi_{\theta_k}}(s_i, a_i)} \right]$$

$$\approx (\theta - \theta_k)^{\top} \underbrace{\sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \left(\nabla_{\theta} \pi_{\theta}(a|s) \Big|_{\theta=\theta_k} \right) Q^{\pi_{\theta_k}}(s, a)}_{\text{PG/NPG}} \\ = \mathbb{E}_{(s_i, a_i)} \left[\frac{\nabla_{\theta} \pi_{\theta}(a_i|s_i) \Big|_{\theta=\theta_k}}{\pi_{\theta_k}(a_i|s_i)} \boxed{Q^{\pi_{\theta_k}}(s_i, a_i)} \right]$$

PG/NPG: Estimate them using the empirical sum of reward in the trajectory (i.e., Monte Carlo estimator)

We can also use other estimators to balance bias and variance

Actor-Critic Methods

Use value function approximation to estimate $Q^{\pi_{\theta_k}}(s_i, a_i)$ or $A^{\pi_{\theta_k}}(s_i, a_i)$

Use $V_{\phi}(s)$ to approximate $V^{\pi_{\theta_k}}(s)$

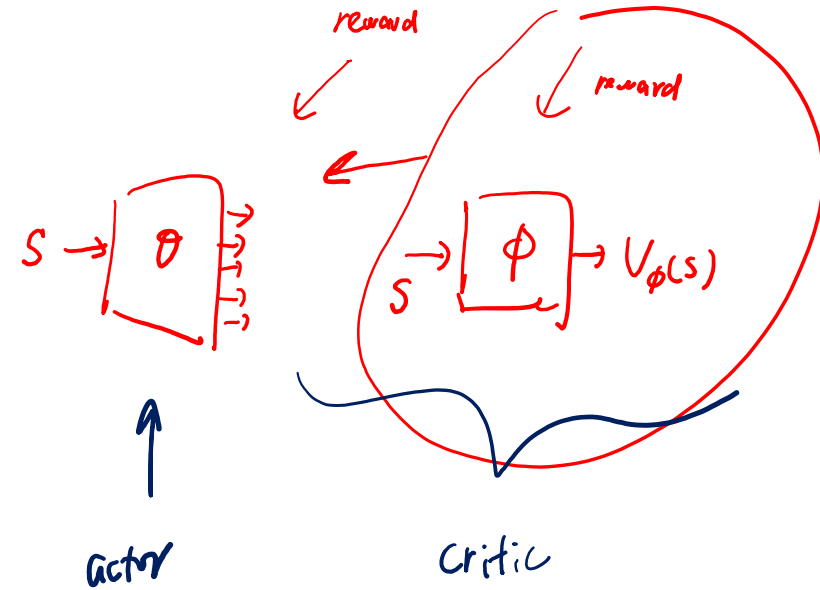
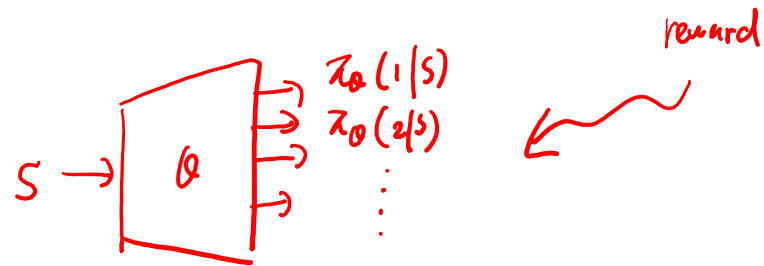
Use $Q_{\phi}(s, a)$ to approximate $Q^{\pi_{\theta_k}}(s, a)$

Possible estimators for $A^{\pi_{\theta_k}}(s, a)$:

Let $(s_1, a_1, r_1, s_2, a_2, r_2, \dots)$ be a trajectory starting from $s_1 = s, a_1 = a$

	$Q_{\phi}(s_1, a_1) - \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot s)}[Q_{\phi}(s_1, a')]$
$r_1 + \gamma V_{\phi}(s_2) - V_{\phi}(s_1)$	$r_1 + \gamma Q_{\phi}(s_2, a_2) - \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot s)}[Q_{\phi}(s_1, a')]$
$r_1 + \gamma r_2 + \gamma^2 V_{\phi}(s_3) - V_{\phi}(s_1)$	$r_1 + \gamma r_2 + \gamma^2 Q_{\phi}(s_3, a_3) - \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot s)}[Q_{\phi}(s_1, a')]$
\vdots	\vdots

Pure Policy-Based Methods vs. Actor-Critic Methods



Actor-Critic with Q_ϕ

(find π^*)
(given π)

(off-policy)

$\Rightarrow Q^*$
Q-learning: $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha [r + \max_{a'} Q(s',a')]$
TD-learning: $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha [r + \sum_{a'} \pi(a'|s) Q(s',a')]$

(on-policy)

\downarrow
 Q^π

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

$$(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}), \dots, (s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)})$$

Define

$$g = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \frac{\nabla_{\theta} \pi_{\theta} (a_h^{(i)} | s_h^{(i)})}{\pi_{\theta_k} (a_h^{(i)} | s_h^{(i)})} \overset{\text{red } Q}{\parallel} Q_{\phi_k} (s_h^{(i)}, a_h^{(i)}) \text{ or } \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \sum_a \nabla_{\theta} \pi_{\theta} (a | s_h^{(i)}) \big|_{\theta=\theta_k} Q_{\phi_k} (s_h^{(i)}, a)$$

Perform updates

$$\theta_{k+1} \leftarrow \theta_k + \eta g \quad \phi_{k+1} \leftarrow \phi_k - \lambda \nabla_{\phi} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \left(Q_{\phi} (s_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - \gamma Q_{\phi_k} (s_{h+1}^{(i)}, a_{h+1}^{(i)}) \right)^2 \bigg|_{\phi=\phi_k}$$

Advantage Actor-Critic (A2C) = PG + V_ϕ

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

$$(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}), \dots, (s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)})$$

Define

$$g = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \nabla_{\theta} \log \pi_{\theta} (a_h^{(i)} | s_h^{(i)}) \Big|_{\theta=\theta_k} \underbrace{\left(r_h^{(i)} + \gamma V_{\phi_k}(s_{h+1}^{(i)}) - V_{\phi_k}(s_h^{(i)}) \right)}_{\text{or any other advantage estimator in the previous slide}}$$

Handwritten notes: $\frac{\nabla_{\theta} \log \pi_{\theta}(a_h^{(i)} | s_h^{(i)})}{\log \pi_{\theta}(a_h^{(i)} | s_h^{(i)})} \approx A^{\pi_k}(s_h^{(i)}, a_h^{(i)})$ and $\mathbb{E}(\cdot) = \sum_{s,a} d_{\rho}^{\pi_k}(s) \nabla_{\theta} \log \pi_{\theta}(a|s) A^{\pi_k}(s,a)$

Perform updates

$$\theta_{k+1} \leftarrow \theta_k + \eta g \quad \phi_{k+1} \leftarrow \phi_k - \lambda \nabla_{\phi} \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \left(V_{\phi}(s_h^{(i)}) - r_h^{(i)} - \gamma V_{\phi_k}(s_{h+1}^{(i)}) \right)^2 \Big|_{\phi=\phi_k}$$

Handwritten note: $V_{\phi} \approx V^{\pi_k}$

Proximal Policy Optimization (PPO) = NPG + V_ϕ

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \dots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

Perform updates

$$\theta_{k+1} \leftarrow \operatorname{argmax}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \frac{\pi_{\theta} \left(a_h^{(i)} | s_h^{(i)} \right)}{\pi_{\theta_k} \left(a_h^{(i)} | s_h^{(i)} \right)} \left(r_h^{(i)} + \gamma V_{\phi_k} \left(s_{h+1}^{(i)} \right) - V_{\phi_k} \left(s_h^{(i)} \right) \right) \right. \\ \left. - \frac{1}{\eta} \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \operatorname{KL} \left(\pi_{\theta} \left(\cdot | s_h^{(i)} \right), \pi_{\theta_k} \left(\cdot | s_h^{(i)} \right) \right) \right\}$$

or any other advantage estimator in the previous slide

$$\phi_{k+1} \leftarrow \phi_k - \lambda \nabla_{\phi} \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \left(V_{\phi} \left(s_h^{(i)} \right) - r_h^{(i)} - \gamma V_{\phi_k} \left(s_{h+1}^{(i)} \right) \right)^2 \Bigg|_{\phi=\phi_k}$$

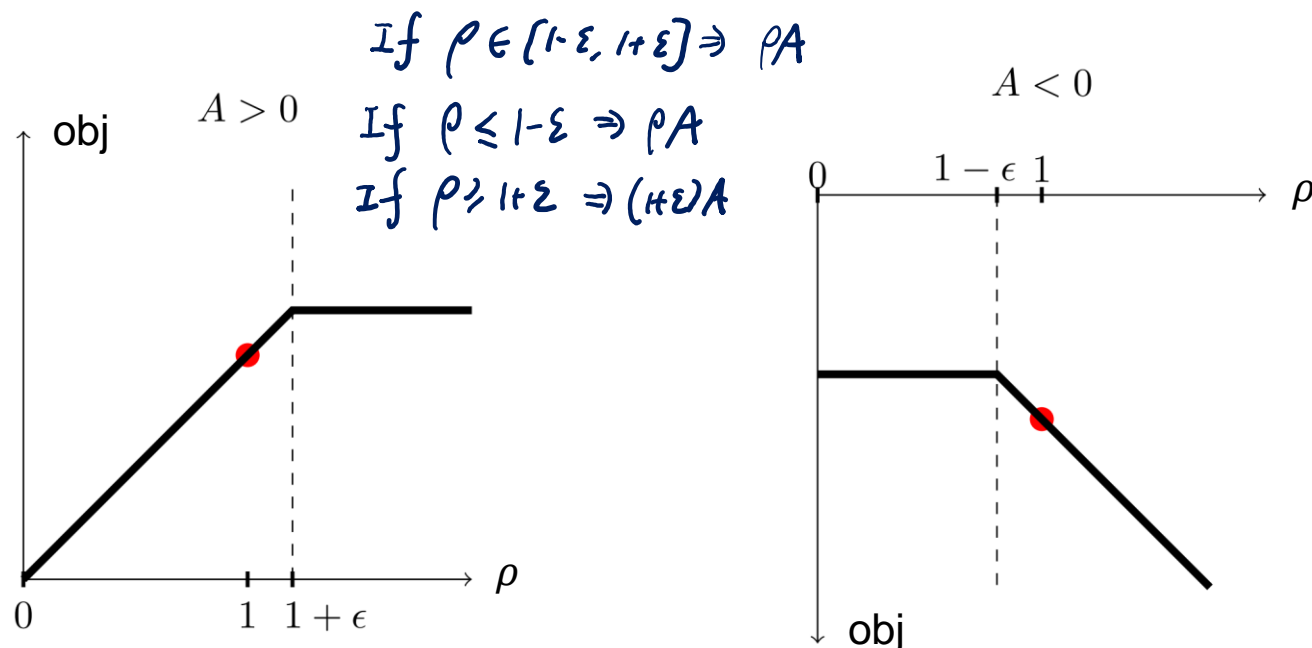
$$\frac{r(a) \mathbb{I}(a_+ = a)}{p_+(a)} \quad \frac{(r(a) - 1) \mathbb{I}(a_+ = a)}{p_+(a)}$$

Additional Technique 1: Clipped Objective (for PPO)

$$\rho := \frac{\pi_{\theta} (a_h^{(i)} | s_h^{(i)})}{\pi_{\theta_k} (a_h^{(i)} | s_h^{(i)})} \quad A := (r_h^{(i)} + \gamma V_{\phi_k} (s_{h+1}^{(i)}) - V_{\phi_k} (s_h^{(i)}))$$

$$\text{clip}_{[1-\epsilon, 1+\epsilon]}(\rho) = \min(\max(\rho, 1-\epsilon), 1+\epsilon)$$

Instead of using ρA as the objective, use $\min\{\rho A, \text{clip}_{[1-\epsilon, 1+\epsilon]}(\rho) A\}$



If $\rho \in (1-\epsilon, 1+\epsilon] \Rightarrow \rho A$

If $\rho \leq 1-\epsilon \Rightarrow (1-\epsilon)A$ (strange case)

If $\rho \geq 1+\epsilon \Rightarrow \rho A$

algorithm	avg. normalized score
No clipping or penalty	-0.39
Clipping, $\epsilon = 0.1$	0.76
Clipping, $\epsilon = 0.2$	0.82
Clipping, $\epsilon = 0.3$	0.70
Adaptive KL $d_{\text{target}} = 0.003$	0.68
Adaptive KL $d_{\text{target}} = 0.01$	0.74
Adaptive KL $d_{\text{target}} = 0.03$	0.71
Fixed KL, $\beta = 0.3$	0.62
Fixed KL, $\beta = 1.$	0.71
Fixed KL, $\beta = 3.$	0.72
Fixed KL, $\beta = 10.$	0.69

Additional Technique 2: Entropy Bonus

In the objective of policy update, add a bonus term

$$H(\pi_{\theta}(\cdot | s)) = \sum_a \pi_{\theta}(a|s) \ln \frac{1}{\pi_{\theta}(a|s)}$$

For PPO:

$$\operatorname{argmax}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \frac{\pi_{\theta}(a_h^{(i)} | s_h^{(i)})}{\pi_{\theta_k}(a_h^{(i)} | s_h^{(i)})} A_h^{(i)} - \frac{1}{\eta} \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \operatorname{KL}(\pi_{\theta}(\cdot | s_h^{(i)}), \pi_{\theta_k}(\cdot | s_h^{(i)})) + c \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \underbrace{H(\pi_{\theta}(\cdot | s_h^{(i)}))}_{\text{Entropy Bonus}} \right\}$$

$-\operatorname{KL}(\pi_{\theta}(\cdot | s_h^{(i)}), \pi_{\text{unif}}(\cdot | s_h^{(i)}))$

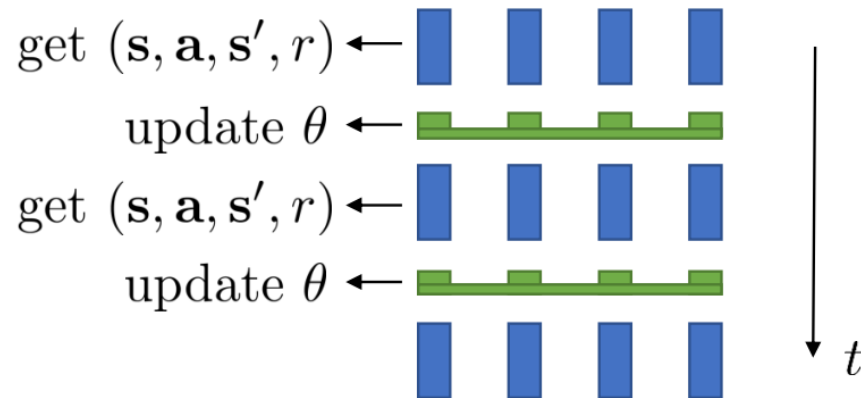
For A2C:

$$g = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \nabla_{\theta} \log \pi_{\theta}(a_h^{(i)} | s_h^{(i)}) \Big|_{\theta=\theta_k} A_h^{(i)} + c \nabla_{\theta} \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} H(\pi_{\theta}(\cdot | s_h^{(i)}))$$

Additional Technique 3: Parallel Sample Collection

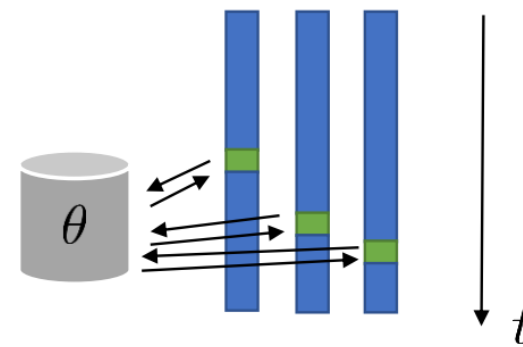
A2C

synchronized parallel actor-critic



A3C

asynchronous parallel actor-critic



Actor-Critic Summary

PG \longrightarrow A2C

NPG \longrightarrow PPO

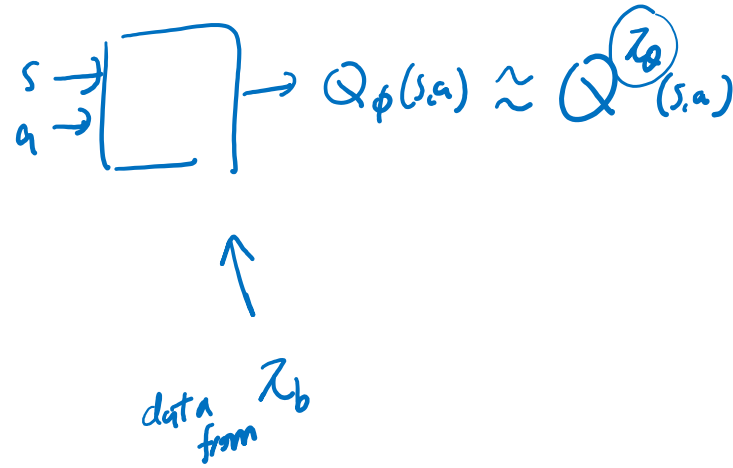
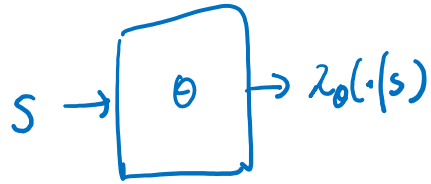
$$s \rightarrow \boxed{\theta} \rightarrow \pi(\cdot|s)$$

$$s \rightarrow \boxed{\theta} \rightarrow \pi(\cdot|s) \quad s \rightarrow \boxed{\phi} \rightarrow V_{\phi}(s)$$

$$s \xrightarrow{a} \boxed{\phi} \rightarrow Q_{\phi}(s,a)$$

Off-Policy Actor-Critic

- Leveraging **off-policy evaluation** → allow reusing data



Review: Full-Information Policy Learning in MDPs

$$\begin{aligned}\theta_{k+1} &= \operatorname{argmax}_{\theta} \left(\underbrace{V^{\pi_{\theta}}(\rho) - V^{\pi_{\theta_k}}(\rho)}_{\text{}} - \frac{1}{\eta} D(\theta, \theta_k) \right) \\ &\approx \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \left(\pi_{\theta}(a|s) - \pi_{\theta_k}(a|s) \right) Q^{\pi_{\theta_k}}(s, a) \\ &\approx (\theta - \theta_k)^{\top} \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \left(\nabla_{\theta} \pi_{\theta}(a|s) \Big|_{\theta=\theta_k} \right) Q^{\pi_{\theta_k}}(s, a)\end{aligned}$$

Use any off-policy policy evaluation methods to find ϕ_k such that $Q_{\phi_k}(s, a) \approx Q^{\pi_{\theta_k}}(s, a)$

Suppose that our (s_i, a_i) samples are obtained from $\hat{\pi}$

Off-Policy Actor-Critic

$$\begin{aligned}
 \theta_{k+1} &= \operatorname{argmax}_{\theta} \left(\underbrace{V^{\pi_{\theta}}(\rho) - V^{\pi_{\theta_k}}(\rho)}_{\text{}} - \frac{1}{\eta} D(\theta, \theta_k) \right) \\
 &\approx \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \left(\pi_{\theta}(a|s) - \pi_{\theta_k}(a|s) \right) Q_{\phi_k}(s, a) = \mathbb{E}_{s \sim d_{\hat{\rho}}} \left[\frac{\cancel{d_{\rho}^{\pi_{\theta_k}}(s)}}{\cancel{d_{\rho}^{\hat{\pi}}(s)}} \sum_a \left(\pi_{\theta}(a|s) - \pi_{\theta_k}(a|s) \right) Q_{\phi_k}(s, a) \right] \\
 &\approx (\theta - \theta_k)^{\top} \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \left(\nabla_{\theta} \pi_{\theta}(a|s) \Big|_{\theta=\theta_k} \right) Q_{\phi_k}(s, a) = (\theta - \theta_k)^{\top} \mathbb{E}_{s \sim d_{\hat{\rho}}} \left[\frac{\cancel{d_{\rho}^{\pi_{\theta_k}}(s)}}{\cancel{d_{\rho}^{\hat{\pi}}(s)}} \sum_a \nabla_{\theta} \pi_{\theta}(a|s) \Big|_{\theta=\theta_k} Q_{\phi_k}(s, a) \right]
 \end{aligned}$$

Use any off-policy policy evaluation methods to find ϕ_k such that $Q_{\phi_k}(s, a) \approx Q^{\pi_{\theta_k}}(s, a)$

Suppose that our (s_i, a_i) samples are obtained from $\hat{\pi}$

Actor-Critic + Replay Buffer

For $k = 1, 2, \dots$

Collect samples using π_{θ_k} , and place them in the replay buffer

Sample a batch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ from replay buffer

Define

$$g = \frac{1}{n} \sum_{i=1}^n \sum_a \nabla_{\theta} \pi_{\theta}(a|s_i) \Big|_{\theta=\theta_k} Q_{\phi_k}(s_i, a) \quad \text{Note: not using } a_i \text{ here}$$

Perform updates

$$\theta_{k+1} \leftarrow \theta_k + \eta g$$

$$\phi_{k+1} \leftarrow \phi_k - \lambda \nabla_{\phi} \frac{1}{n} \sum_{i=1}^n \left(Q_{\phi}(s_i, a_i) - r_i - \gamma \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s'_i)} [Q_{\phi_k}(s'_i, a')] \right)^2 \Big|_{\phi=\phi_k}$$

Off-policy TD \rightarrow unstable (more on this later)

Dealing with Continuous Action Sets

Review: Linear Bandits and One-Point Gradient Estimator

Feasible set $A \subseteq \mathbb{R}^d$

For $t=1, \dots, T$:

Learner choose $a_t \in A$

Environment reveals $f_t(a_t)$, where $f_t: A \rightarrow \mathbb{R}$

Idea/update

$$a_{t+1} \leftarrow a_t + \eta \nabla f_t(a_t)$$

$$\overset{\cdot}{a_t - v} \quad \overset{\times}{a_t} \quad \overset{\cdot}{a_t + v}$$

(1d)

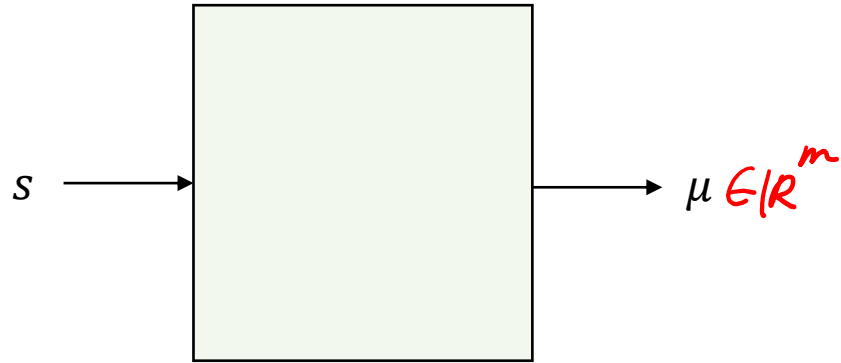
$$\nabla f_t(a_t) \approx \frac{f_t(a_t+v) - f_t(a_t-v)}{2v}$$

$$\mathbb{E} \frac{f_t(\tilde{a}_t) S}{v} = \frac{f_t(\tilde{a}_t) (\tilde{a}_t - a_t)}{v^2}$$

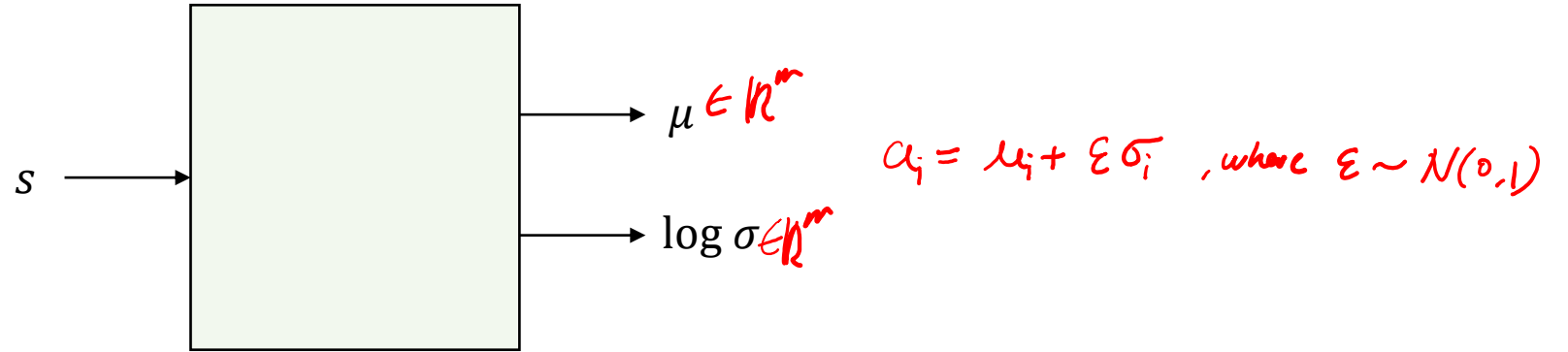
$$\text{where } \tilde{a}_t = \begin{cases} a_t + v, & \text{w.p. } \frac{1}{2} \\ a_t - v, & \text{w.p. } \frac{1}{2} \end{cases}$$

$$S = \begin{cases} 1, & \tilde{a}_t = a_t + v \\ -1, & \tilde{a}_t = a_t - v \end{cases}$$

Policy Network for Continuous Action Sets



Policy Network for Continuous Action Sets



A2C / PPO with Continuous Action Sets

$$\int_a \pi_\theta(a|s) da = 1$$

μ_θ

$$\pi_\theta(a) = \text{const} \cdot \exp\left(\frac{-(a-\mu_\theta)^2}{2\sigma^2}\right)$$

$$g = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \log \pi_{\theta}(a_i | s_i) \Big|_{\theta=\theta_k} A_i$$

$$\log \pi_\theta(a) = \text{const} - \frac{(a-\mu_\theta)^2}{2\sigma^2}$$

$$\nabla_{\theta} \log \pi_\theta(a) = \nabla_{\theta} \left[\frac{(a-\mu_\theta)^2}{2\sigma^2} \right] \in \mathbb{R}^n$$

$\mathbb{R}^{n \times m}$ m

$$\mu_\theta \in \mathbb{R}^m$$

$$\theta \in \mathbb{R}^n$$

$$\theta_{k+1} \leftarrow \operatorname{argmax}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\pi_{\theta}(a_i | s_i)}{\pi_{\theta_k}(a_i | s_i)} A_i - \frac{1}{\eta} \frac{1}{n} \sum_{i=1}^n \text{KL}(\pi_{\theta}(\cdot | s_i), \pi_{\theta_k}(\cdot | s_i)) \right\}$$



Recall: Actor-Critic with Q_ϕ Critic



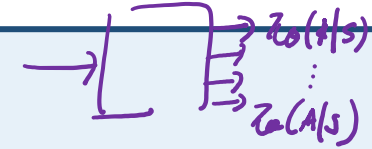
For $k = 1, 2, \dots$

Use π_{θ_k} to collect samples $\{(s_i, a_i, r_i, s_i')\}_{i=1}^n$

Define $g = \frac{1}{n} \sum_{i=1}^n \sum_a \nabla_\theta \pi_\theta(a|s_i) \Big|_{\theta=\theta_k} Q_{\phi_k}(s_i, a)$

Perform updates

$$\theta_{k+1} \leftarrow \theta_k + \eta g \qquad \phi_{k+1} \leftarrow \phi_k - \lambda \nabla_\phi \sum_{i=1}^n \left(Q_\phi(s_i, a_i) - r_i - \gamma \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s_i')} [Q_{\phi_k}(s_i', a')] \right)^2 \Big|_{\phi=\phi_k}$$



Deterministic Policy Gradient Theorem

Policy: $\mu_\theta(s) \in \mathbb{R}^n$

$$\begin{aligned} V^{\pi_{\theta+\delta\theta}}(\rho) - V^{\pi_\theta}(\rho) &= \sum_s d_\rho^{\pi_{\theta+\delta\theta}}(s) \sum_a \left(\underline{\pi_{\theta+\delta\theta}(a|s)} - \underline{\pi_\theta(a|s)} \right) Q^{\pi_\theta}_{(s,a)} \\ &= \sum_s d_\rho^{\pi_{\theta+\delta\theta}}(s) \left(Q^{\pi_\theta}(s, \mu_{\theta+\delta\theta}(s)) - Q^{\pi_\theta}(s, \mu_\theta(s)) \right) \end{aligned}$$

$$\begin{aligned} \Rightarrow \nabla_\theta V^{\pi_\theta}(\rho) &= \sum_s d_\rho^{\pi_\theta}(s) \nabla_\theta \left[Q^{\pi_\theta}(s, \mu_\theta(s)) \right] \\ &\quad \left\{ \sum_s d_\rho^{\pi_\theta}(s) \left[\sum_a \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right] \right\} \end{aligned}$$

Deterministic Policy Gradient

A.C.

$$Q_{\phi_k} \approx Q^{\pi_{\theta_k}}$$

For $k = 1, 2, \dots$

Use μ_{θ_k} to collect samples $\{(s_i, a_i, r_i, s_i')\}_{i=1}^n$

Q-learning

$$Q \leftarrow (1-\alpha)Q + \alpha(r + \gamma \max_{a'} Q(s', a'))$$

Define $g = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} Q_{\phi_k}(s_i, \mu_{\theta_k}(s_i)) \Big|_{\theta=\theta_k}$

Perform updates

$$\theta_{k+1} \leftarrow \theta_k + \eta g \quad \phi_{k+1} \leftarrow \phi_k - \lambda \nabla_{\phi} \sum_{i=1}^n \left(Q_{\phi}(s_i, a_i) - r_i - \gamma Q_{\phi_k}(s'_i, \mu_{\theta_k}(s'_i)) \right)^2 \Big|_{\phi=\phi_k}$$

$\mu_{\theta_k}(s) \approx \underset{a}{\operatorname{argmax}} Q_{\phi_k}(s, a)$

Two Viewpoints for the Deterministic PG Algorithm

Deep Deterministic Policy Gradient (DDPG)

For $k = 1, 2, \dots$

Use $\mu_{\theta_k}(s) + \mathcal{N}(0, \sigma^2)$ to collect samples and place them in replay buffer

Sample a batch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ from the replay buffer

$$\theta \leftarrow \theta + \eta \sum_{i=1}^n \nabla_{\theta} Q_{\phi}(s_i, \mu_{\theta}(s_i))$$

$$\phi \leftarrow \phi - \lambda \nabla_{\phi} \sum_{i=1}^n \left(Q_{\phi}(s_i, a_i) - r_i - \gamma Q_{\phi_{\text{tar}}}(s'_i, \mu_{\theta_{\text{tar}}}(s'_i)) \right)^2$$

$$\theta_{\text{tar}} \leftarrow \tau \theta + (1 - \tau) \theta_{\text{tar}}$$

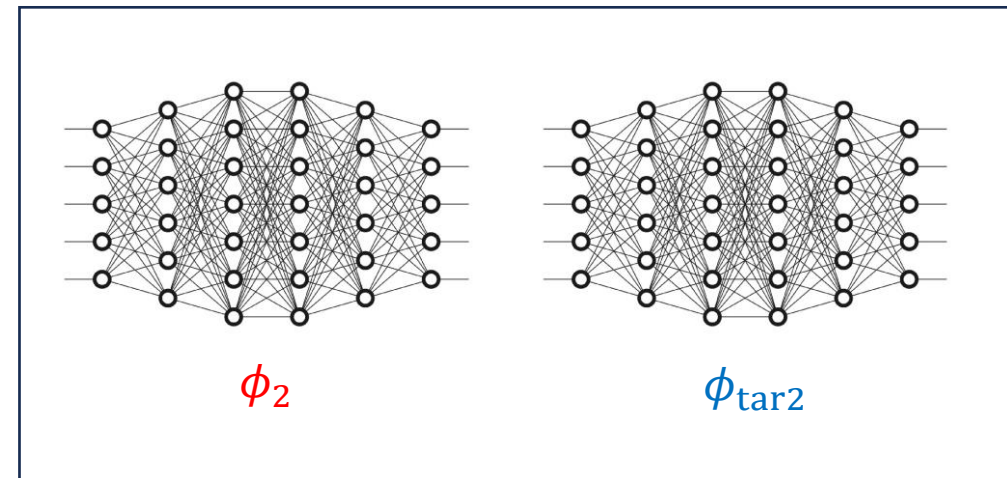
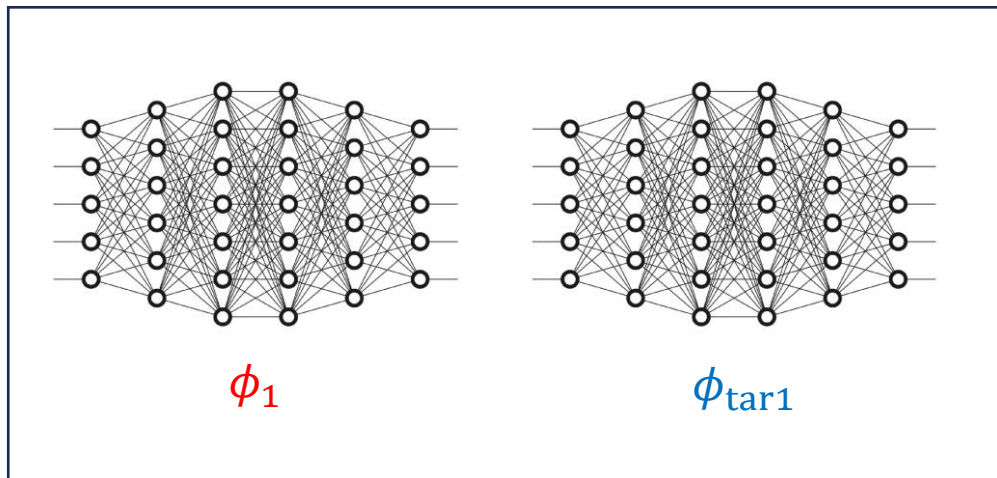
$$\phi_{\text{tar}} \leftarrow \tau \phi + (1 - \tau) \phi_{\text{tar}}$$

Elements: replay buffer, target network, action noise

Lillicrap et al., Continuous control with deep reinforcement learning. 2015.

Further Stabilizing DDPG (1/3)

- Double Q-learning



Double Q-learning: When training ϕ_1 , instead of using $Q_{\phi_{tar1}}$ to evaluate the regression target, use $Q_{\phi_{tar2}}$

TD3: $\min \{Q_{\phi_{tar1}}, Q_{\phi_{tar2}}\}$

Double Q-learning: Use independent samples to train ϕ_1 and ϕ_2

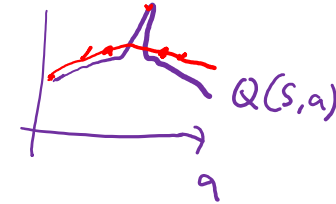
TD3: Use the same set of samples
(the independence between ϕ_1 and ϕ_2 only comes from random initialization)

Further Stabilizing DDPG (2/3)

- Target policy smoothing

DDPG: use $Q_{\phi_{\text{tar}}}(s', \mu_{\theta_{\text{tar}}}(s'))$ as the regression target

TD3: sample $a' = \mu_{\theta_{\text{tar}}}(s') + \mathcal{N}(0, \sigma^2)$
use $Q_{\phi_{\text{tar}}}(s', a')$ as the regression target



Further Stabilizing DDPG (3/3)

- Delayed policy updates: running multiple steps of value updates before running one step of policy update

Twin Delayed DDPG (TD3)

For $k = 1, 2, \dots$

Use $\mu_\theta(s) + \mathcal{N}(0, \sigma^2)$ to collect samples and place them in replay buffer

Sample a batch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ from the replay buffer

For each sample i , draw $a'_i \sim \mu_{\theta_{\text{tar}}}(s'_i) + \mathcal{N}(0, \sigma^2 I)$

$$\phi_j \leftarrow \phi_j - \lambda \nabla_{\phi_j} \sum_{i=1}^n \left(Q_{\phi_j}(s_i, a_i) - r_i - \gamma \min_{\ell=1,2} Q_{\phi_{\text{tar}\ell}}(s'_i, a'_i) \right)^2 \quad \forall j = 1, 2$$

If $k \bmod M = 0$:

$$\theta \leftarrow \theta + \eta \sum_{i=1}^n \nabla_{\theta} Q_{\phi}(s_i, \mu_{\theta}(s_i))$$

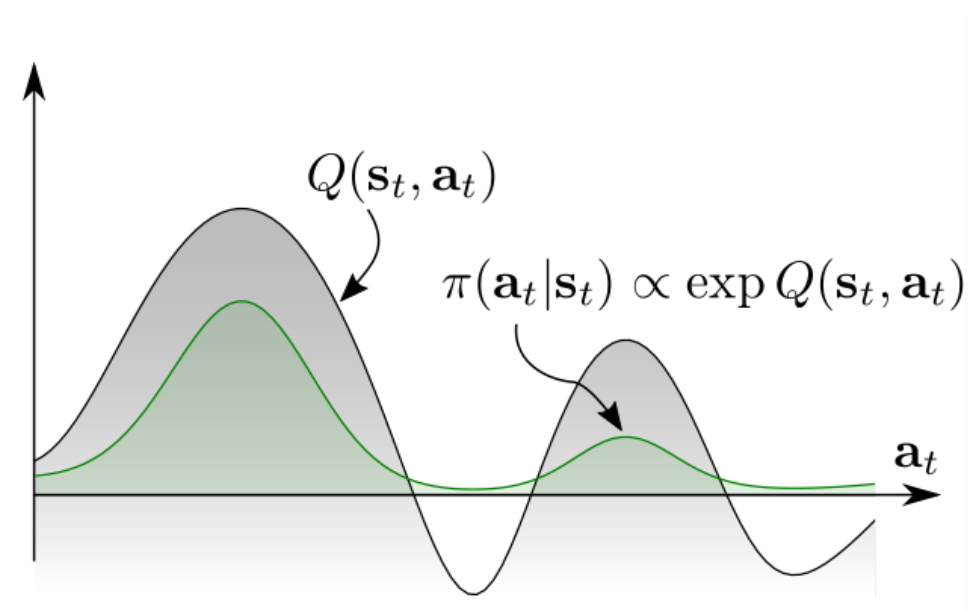
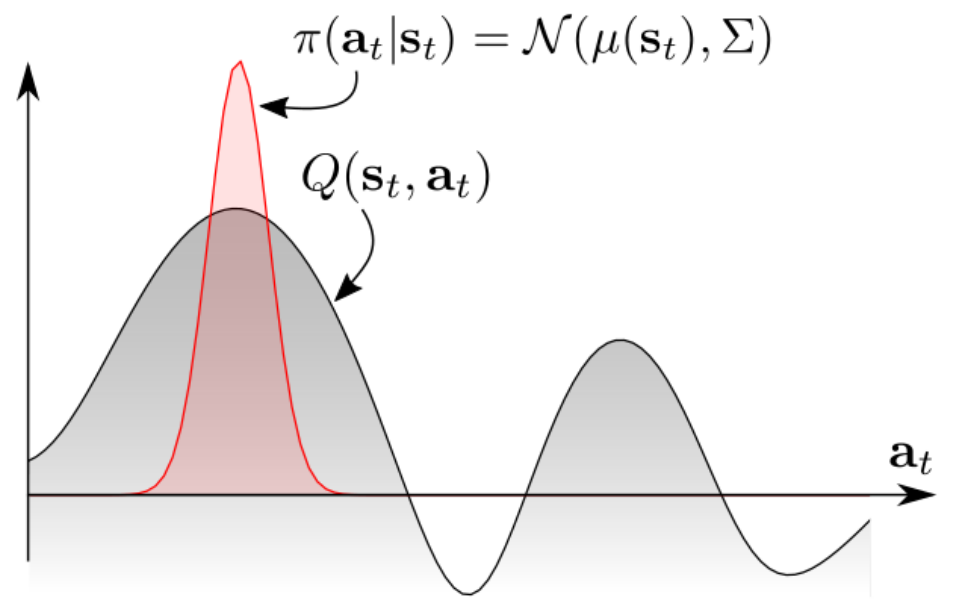
$$\theta_{\text{tar}} \leftarrow \tau \theta + (1 - \tau) \theta_{\text{tar}}$$

$$\phi_{\text{tar}j} \leftarrow \tau \phi_j + (1 - \tau) \phi_{\text{tar}j} \quad \forall j = 1, 2$$

Soft Actor-Critic (SAC)

- TD3 / DDPG: modeling a deterministic policy + additional noise for exploration
- SAC: modeling a randomized policy (by adding entropy as an exploration bonus)
- TD3 / DDPG vs. SAC is similar to ϵ -greedy vs. Boltzmann exploration

Entropy Bonus



TD3 vs. SAC

- Policy update

TD3: Do not view $-\alpha \log \pi_{\theta}(a|s)$ as part of the reward

Simply perform $\theta \leftarrow \theta + \eta \nabla_{\theta} Q_{\phi}(s, \mu_{\theta}(s))$

SAC: View $-\alpha \log \pi_{\theta}(a|s)$ as part of the reward

Perform the following:

Let $a_{\theta}(s) = \mu_{\theta}(s) + \epsilon \sigma_{\theta}(s)$ where $\epsilon \sim \mathcal{N}(0,1)$

Perform $\theta \leftarrow \theta + \eta \nabla_{\theta} (Q_{\phi}(s, a_{\theta}(s)) - \alpha \log \pi_{\theta}(a_{\theta}(s)|s))$

“reparameterization trick”

TD3 vs. SAC

- Value update

TD3: Sample $a' \sim \mu_{\theta}(s') + \mathcal{N}(0, \sigma^2)$

Use $Q_{\phi_{\text{tar}}}(s', a')$ as the regression target

SAC: Sample $a' \sim \pi_{\theta}(\cdot | s') = \mu_{\theta}(s') + \mathcal{N}(0, \sigma_{\theta}^2(s'))$

Use $Q_{\phi_{\text{tar}}}(s', a') - \alpha \log \pi_{\theta}(a' | s')$ as the regression target

Soft Actor-Critic (SAC)

For $k = 1, 2, \dots$

Use $\mu_\theta(s) + \mathcal{N}(0, \sigma^2)$ to collect samples and place them in replay buffer

Sample a batch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ from the replay buffer

For each sample i , draw $a'_i \sim \mu_\theta(s'_i) + \mathcal{N}(0, \sigma_\theta^2(s'_i))$

$$\phi_j \leftarrow \phi_j - \lambda \nabla_{\phi_j} \sum_{i=1}^n \left(Q_{\phi_j}(s_i, a_i) - r_i - \gamma \min_{\ell=1,2} Q_{\phi_{\text{tar}^\ell}}(s'_i, a'_i) + \alpha \log \pi_\theta(a'_i | s'_i) \right)^2 \quad \forall j = 1, 2$$

Let $a_\theta(s_i) = \mu_\theta(s_i) + \epsilon \sigma_\theta(s_i)$ where $\epsilon \sim \mathcal{N}(0, I)$

Perform $\theta \leftarrow \theta + \eta \sum_{i=1}^n \nabla_\theta (Q_\phi(s, a_\theta(s_i)) - \alpha \log \pi_\theta(a_\theta(s_i) | s_i))$

$$\phi_{\text{tar}^j} \leftarrow \tau \phi_j + (1 - \tau) \phi_{\text{tar}^j} \quad \forall j = 1, 2$$