

Notes Week 12

Sequential critic: Actor-Critic

For every update step:

$$\begin{aligned}
 A &\sim \pi(\cdot|S; \theta) \\
 &\text{Take action } A \text{ observe } S', R \\
 \theta &\leftarrow R + \gamma V(S'; \mathbf{w}) - V(S; \mathbf{w}) \\
 \mathbf{w} &\leftarrow \mathbf{w} + \lambda^{\mathbf{w}} \delta \nabla V(S; \mathbf{w}) \\
 \theta &\leftarrow \theta + \lambda^{\theta} I \delta \nabla \ln \pi(A|S; \theta)
 \end{aligned}$$

For an m -agent problem, we change this to:

$$A_1 \sim \pi_1(\cdot|S; \theta_1), A_2 \sim \pi_2(\cdot|S, A_1; \theta_2), \dots, A_m \sim \pi_m(\cdot|S, A_1, \dots, A_{m-1}; \theta_m)$$

Take actions A_1, \dots, A_m observe S', R

$$\delta_i \leftarrow R + \gamma V_i(S, A_1, \dots, A_{i-1}; \mathbf{w}_i) - V_{i-1}(S, A_1, \dots, A_{i-2}; \mathbf{w}_{i-1}), \quad \text{for } i = 1, \dots, m$$

$$\delta_m \leftarrow R(S, A_1, \dots, A_m) + \gamma V_1(S'; \mathbf{w}_1) - V_{m-1}(S, A_1, \dots, A_{m-1}; \mathbf{w}_{m-1})$$

$$\mathbf{w}_i \leftarrow \mathbf{w}_i + \lambda_i^{\mathbf{w}} \delta_i \nabla V_i(S, A_1, \dots, A_{i-1}; \mathbf{w}_i), \quad \text{for } i = 1, \dots, m$$

$$\theta_i \leftarrow \theta_i + \lambda_i^{\theta} I \delta_i \nabla \ln \pi_i(A|S, A_1, \dots, A_{i-1}), \quad \text{for } i = 1, \dots, m$$

Sequential critic: Soft Actor-Critic

Given the objective functions:

$$y(R, S', d) = R + \gamma(1 - d) \left(\min_{i=1,2} Q_i^{targ}(S', \tilde{A}'; \mathbf{w}_{targ}) - \alpha \log \pi(\tilde{A}'|S'; \theta) \right), \quad \tilde{A}' \sim \pi(\cdot|S'; \theta)$$

$$J_Q(\mathbf{w}) = \frac{1}{2} \sum_{i=1,2} (Q_i(S, A; \mathbf{w}_i) - y(R, S', d))^2$$

$$J_\pi(\theta) = \alpha \log \pi_\theta(\tilde{A}(S|\theta)|S; \theta) - \min_{i=1,2} Q_i(S, \tilde{A}(S|\theta); \mathbf{w})$$

$$J(\alpha) = -\alpha(\log \pi(A|S) + \mathcal{H})$$

For each gradient step:

$$A \sim \pi(\cdot|S; \theta)$$

Take action A observe S', R

$$\mathbf{w}_i \leftarrow \mathbf{w}_i - \lambda^{\mathbf{w}_i} \nabla_{\mathbf{w}_i} J_Q(\mathbf{w}_i) \quad \text{for } i = 1, 2$$

$$\theta \leftarrow \theta - \lambda^\theta \nabla_\theta J_\pi(\theta)$$

$$\alpha \leftarrow \alpha - \lambda^\alpha \nabla_\alpha J(\alpha)$$

$$\mathbf{w}_{targ,i} \leftarrow \rho \mathbf{w}_{targ,i} + (1 - \rho) \mathbf{w}_i, \quad \text{for } i = 1, 2$$

where:

- $y(R, S', d)$: Critic target.
- \mathbf{w} : Critic parameters (Two critics for clipped double Q, plus two target critics).
- θ : Actor parameters.
- α : Entropy temperature.
- $\tilde{A}(S|\theta)$: a sample from $\pi(\cdot|S; \theta)$ which is differentiable w.r.t. θ via the reparameterization trick:

$$\tilde{A}(S|\theta) = \tanh(\mu_\theta(S) + \sigma(S) \odot \xi), \quad \xi \sim N(0, I)$$

- \mathcal{H} : the entropy target, often taken as $-\dim(\mathcal{A})$, where \mathcal{A} is the action space.

For an m -agent problem, we change this to:

For each stage i in the sequential Q -calculation, from $i = 1, \dots, m$, we have a set

of two ($k = 1, 2$) Q -functions to perform clipped double Q -learning.

$$y_i(R_i, S, d) = R_i + \gamma(1 - d) \left(\min_{k=1,2} Q_{i,k}^{targ}(S, A_1, \dots, A_i; \mathbf{w}_{i,k}^{targ}) - \alpha_i \log \pi_i(A_i|S, A_1, \dots, A_{i-1}) \right)$$

$$y_m(R, S', d) = R + \gamma(1 - d) \left(\min_{k=1,2} Q_{1,k}^{targ}(S', A'_1; \mathbf{w}_{1,k}^{targ}) - \alpha_1 \log \pi_1(A'_1|S'; \theta_1) \right),$$

$$J_{Q_i}(\mathbf{w}_i) = \frac{1}{2} \sum_{k=1,2} (Q_{i,k}(S, A_1, \dots, A_i; \mathbf{w}_{i,k}) - y_{i+1}(R_i, S, d))^2$$

$$J_{Q_m}(\mathbf{w}_m) = \frac{1}{2} \sum_{k=1,2} (Q_{m,k}(S, A_1, \dots, A_m; \mathbf{w}_m) - y_m(R, S', d))^2$$

$$J_{\pi_i}(\theta_i) = \alpha_i \log \pi_i(\tilde{A}(S|\theta)|S, A_1, \dots, A_{i-1}; \theta) - \min_{k=1,2} Q_{i,k}(S, A_1, \dots, \tilde{A}_i(S|\theta); \mathbf{w})$$

$$J(\alpha_i) = -\alpha_i(\log \pi_i(A_i|S, A_1, \dots, A_{i-1}) + \mathcal{H}_i)$$

where:

- y : Critic target.
- \mathbf{w} : Critic parameters (Two critics for clipped double Q, plus two target critics).
- θ : Actor parameters.
- α : Entropy temperature.
- $\tilde{A}_i(S|\theta_i)$: a sample from $\pi_i(\cdot|S; \theta_i)$ which is differentiable w.r.t. θ_i via the reparameterization trick:

For each gradient step:

$$A_i \sim \pi_i(\cdot|S, A_1, \dots, A_{i-1}; \theta) \quad \text{for } i = 1, \dots, m-1$$

Take actions A_1, \dots, A_m observe S', R

$$A'_i \sim \pi_i(\cdot|S', A'_1, \dots, A'_{i-1}; \theta_i) \quad \text{for } i = 1, \dots, m-1$$

$$\mathbf{w}_{i,k} \leftarrow \mathbf{w}_{i,k} - \lambda^{\mathbf{w}_{i,k}} \nabla_{\mathbf{w}_{i,k}} J_{Q_i}(\mathbf{w}_{i,k}) \quad \text{for } k = 1, 2, \text{ and } i = 1, \dots, m$$

$$\theta_i \leftarrow \theta_i - \lambda^{\theta_i} \nabla_{\theta_i} J_{\pi_i}(\theta_i) \quad \text{for } i = 1, \dots, m$$

$$\alpha_i \leftarrow \alpha_i - \lambda_i^\alpha \nabla_{\alpha_i} J(\alpha_i) \quad \text{for } i = 1, \dots, m$$

$$\mathbf{w}_{i,k}^{targ} \leftarrow \rho \mathbf{w}_{i,k}^{targ} + (1 - \rho) \mathbf{w}_{i,k}, \quad \text{for } k = 1, 2, \text{ and } i = 1, \dots, m$$