\vdash

Notes Week 12

Sequential critic: Actor-Critic

For every update step:

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

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

$$\delta_{i} \leftarrow R + \gamma V_{i}(S,A_{1},\dots,A_{i-1};\mathbf{w}_{i}) - V_{i-1}(S,A_{1},\dots,A_{i-2};\mathbf{w}_{i-1}), \qquad \text{for } i=1,\dots,$$

$$\delta_{m} \leftarrow R(S,A_{1},\dots,A_{m}) + \gamma V_{1}(S';\mathbf{w}_{1}) - V_{m-1}(S,A_{1},\dots,A_{m-1};\mathbf{w}_{m-1})$$

$$\mathbf{w}_{i} \leftarrow \mathbf{w}_{i} + \lambda_{i}^{\mathbf{w}} \delta_{i} \nabla V_{i}(S,A_{1},\dots,A_{i-1};\mathbf{w}_{i}), \qquad \text{for } i=1,$$

$$\theta_{i} \leftarrow \theta_{i} + \lambda^{\theta_{i}} I \delta_{i} \nabla \ln \pi_{i}(A|S,A_{1},\dots,A_{i-1}), \qquad \text{for } i=1,$$

Sequential critic: Soft Actor-Critic

Given the objective functions:

$$egin{aligned} y(R,S',d) &= R + \gamma(1-d) \left(\min_{i=1,2} Q_i^{targ}(S', ilde{A}';\mathbf{w}_{targ}) - lpha \log \pi(ilde{A}'|S'; heta)
ight), \quad ilde{A}' \sim \pi(\cdot|S'; heta) \ J_Q(\mathbf{w}) &= rac{1}{2} \sum_{i=1,2} (Q_i(S,A;\mathbf{w}_i) - y(R,S',d))^2 \ J_\pi(heta) &= lpha \log \pi_ heta(ilde{A}(S| heta)|S; heta) - \min_{i=1,2} Q_i(S, ilde{A}(S| heta);\mathbf{w}) \ J(lpha) &= -lpha (\log \pi(A|S) + \mathcal{H}) \end{aligned}$$

For each gradient step:

$$egin{aligned} A &\sim \pi(\cdot|S; heta) \ & ext{Take action A observe S', R} \ & ext{ extbf{w}}_i \leftarrow ext{ extbf{w}}_i - \lambda^{ ext{ extbf{w}}_i}
abla_{ extbf{w}_i} J_Q(extbf{w}_i) & ext{for $i=1,2$} \ & heta \leftarrow heta - \lambda^{ heta}
abla_{ heta} J_{\pi}(heta) \ & ext{ extbf{\alpha}} \ & ext{ extbf{\alpha}} \leftarrow ext{ extbf{\alpha}}
abla_{ extbf{\alpha}} J_{\alpha}(extbf{\alpha}) \ & ext{ extbf{w}}_{targ,i} \leftarrow
ho extbf{ extbf{w}}_{targ,i} + (1-
ho) extbf{ extbf{w}}_i, & ext{for $i=1,2$} \end{aligned}$$

where:

- y(R, S', d): Critic target.
- 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:

$$ilde{A}(S| heta) = anh(\mu_{ heta}(S) + \sigma(S) \odot \xi), \quad \xi \sim N(0,I)$$

• \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 $i=1,\ldots,m$, we have a set

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

$$egin{aligned} y_i(R_i,S,d) &= R_i + \gamma(1-d) \left(\min_{k=1,2} Q_{i,k}^{targ}(S,A_1,\ldots,A_i;\mathbf{w}_{i,k}^{targ}) - lpha_i \log \pi_i(A_i|S,A_1,\ldots,A_{i-1};\mathbf{w}_{i,k}^{targ}) - lpha_i \log \pi_i(A_i|S,A_1,\ldots,A_{i-1};\mathbf{w}_{i,k}^{targ}) - lpha_i \log \pi_m(A_m'|S',A_1',\ldots,A_m';\mathbf{w}_{m,k}^{targ}) - lpha_m \log \pi_m(A_m'|S',A_1',\ldots,A_{i-1};\mathbf{w}_{i-1,k}) - y_i(R_i,S,d)
brace \\ J_{Q_i}(\mathbf{w}_i) &= \frac{1}{2} \sum_{k=1,2} (Q_{i-1,k}(S,A_1,\ldots,A_{i-1};\mathbf{w}_{i-1,k}) - y_i(R_i,S,d))^2 \\ J_{Q_m}(\mathbf{w}_m) &= \frac{1}{2} \sum_{k=1,2} (Q_{i,k}(S,A_1;\mathbf{w}_1) - y_m(R,S',d))^2 \\ J_{\pi_i}(heta_i) &= lpha_i \log \pi_i(\tilde{A}(S| heta)|S,A_1,\ldots,A_{i-1}; heta) - \min_{k=1,2} Q_{i,k}(S,A_1,\ldots,\tilde{A}_i(S| heta);\mathbf{w}) \\ J(lpha_i) &= -lpha_i (\log \pi_i(A_i|S,A_1,\ldots,A_{i-1}) + \mathcal{H}_i) \end{aligned}$$

where:

- y: Critic target.
- 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:

 $\mathbf{w}_{i,k}^{targ} \leftarrow \rho \mathbf{w}_{i,k}^{targ} + (1 - \rho) \mathbf{w}_{i,k},$

$$A_1 \sim \pi_1(\cdot|S; heta_1), A_2 \sim \pi_2(\cdot|S, A_1; heta_2), \ldots, A_m \sim \pi_m(\cdot|S, A_1, \ldots, A_{m-1}; heta_m)$$
Take actions A_1, \ldots, A_m observe S', R

$$A'_1 \sim \pi_1(\cdot|S'; heta_1), A'_2 \sim \pi_2(\cdot|S', A'_1; heta_2), \ldots, A'_m \sim \pi_m(\cdot|S', A'_1, \ldots, A'_{m-1}; heta_m)$$

$$\mathbf{w}_{i,k} \leftarrow \mathbf{w}_{i,k} - \lambda^{\mathbf{w}_{i,k}} \nabla_{\mathbf{w}_{i,k}} J_{Q_i}(\mathbf{w}_{i,k})$$
for $k = 1, 2$,
$$\theta_i \leftarrow \theta_i - \lambda^{\theta_i} \nabla_{\theta_i} J_{\pi_i}(\theta_i)$$

$$\alpha_i \leftarrow \alpha_i - \lambda^{\alpha}_i \nabla_{\alpha_i} J(\alpha_i)$$

for k = 1, 2.