RL 2022 - lec 3 : policy gradient

The setup :

$$\max_{\rho} J^{\rho}(x_0) = \mathbb{E}\left[\sum_{k=0}^{N} \gamma^k r(X_k, U_k) \,\middle|\, X_0 = x_0 \right]$$

s.t. $X_{k+1} \sim f(x_{k+2} / x_k, u_k)$

$U_k \sim \rho^{\theta}(u_k / x_k)$

θ — policy par. (tensor)

$$Z_k := (X_k, U_k)$$

$\Rightarrow$ can rewrite the env. as

$$Z_{k+1} \sim P^{\theta}(z_{k+2} / z_k)$$

state-action trajectory :

$$\bar{Z}_k^N := \{ Z_k, Z_{k+2}, \dots, Z_{k+N-1} \}$$

The accumulated reward (return):

$$G_k^N(\bar{Z}_k^N) := \sum_{i=k}^{k+N-1} \gamma^i r(Z_i)$$

Let's unwrap the objective:

$$J^\theta(x_0) = \mathbb{E}\left[G_0^N(\bar{Z}_0^N) \,\middle|\, X_0 = x_0 \right]$$

$$= \int_{\mathbb{X}^N \times \mathbb{U}^N} G_0^N(\bar{z}_0^N) \, \mathbb{P}\left[d\bar{z}_0^N \,\middle|\, X_0 = x_0 \right]$$

$\mathbb{X}$: state space; $\mathbb{U}$: action space

↑ integration w.r.t. prob-ty measure

In finite state and action space case, that integral would amount to

$$\sum_j G_0^N(\bar{z}_0^{N(j)}) \, \mathbb{P}\left[\bar{z}_0^{N(j)} \right]$$

↑ index of an outcome

An example policy (Gaussian)

① $$\rho^\theta(u|x) = \frac{1}{\sigma(x,\theta)\sqrt{2\pi}} e^{-\frac{(u-\mu(x,\theta))^2}{2\sigma(x,\theta)^2}},$$

where

feature vector

$$\mu(x,\theta) = \theta_\mu^T \psi_\mu(x),$$

$$\sigma(x,\theta) = e^{\theta_\sigma^T \psi_\sigma(x)}$$

feature vector

$$\theta = (\theta_\mu, \theta_\sigma)$$

② In case of finite action space, say,

soft - argmax

$$\rho^\theta(u,x) = \frac{e^{K(u|x,\theta)}}{\sum_{s\in U} e^{K(s|x,\theta)}},$$

where $K(u|x,\theta) = \theta^T \psi(u,x)$

Getting back:

$$J^\theta(x_0) = \int_{\mathbb{X}^N \times \mathbb{U}^N} G_0^N(\bar{z}_0^N)\, \mathbb{P}\left[d\bar{z}_0^N\right]$$

$$= \int_{\mathbb{X}^N \times \mathbb{U}^N} G_0^N(\bar{z}_0^N)\, P_0(x_0) \cdot \prod_{k=0}^{N} \rho^\theta(u_k|x_k) \cdot f(x_{k+1}|x_k, u_k)\, d\bar{z}_0^N$$

Let's use gradient ascent rule for the policy update:

$$\theta_{i+1} := \theta_i + \alpha \nabla_\theta J^\theta(x_0)\Big|_{\theta=\theta_i}$$

↑ gradient step index

↑ learning rate

Log-likelihood trick:

$$\nabla_\theta P^\theta(z) = P^\theta(z)\, \nabla_\theta \ln P^\theta(z)$$

Now, apply the trick to our case:

$$\nabla_\theta \left(\int_{\mathcal{X}^N \times \mathcal{U}^N} G_0^N(\bar{z}_0^N) P_0(x_0) \cdot \prod_{k=0}^{N} \rho^\theta(u_k/x_k) \cdot f(x_{k+1}/x_k, u_k) \, d\bar{z}_0^N \right) =$$

$$\int_{\mathcal{X}^N \times \mathcal{U}^N} G_0^N(\bar{z}_0^N) P_0(x_0) \cdot \prod_{k=0}^{N} \rho^\theta(u_k/x_k) \cdot f(x_{k+1}/x_k, u_k)$$

$$\nabla_\theta \ln \left(P_0(x_0) \cdot \prod_{k=0}^{N} \rho^\theta(u_k/x_k) \cdot f(x_{k+1}/x_k, u_k) \right) d\bar{z}_0^N =$$

Observe the following:

$$\mathbb{E}[h(X)] = \int_{\mathcal{X}} p(x) h(x) \, dx = \int_{\mathcal{X}} h(x) \, \mathbb{P}[dx]$$

$$\int_{\mathcal{X}^N \times \mathcal{U}^N} G_0^N(\bar{z}_0^N) P_0(x_0) \cdot \prod_{k=0}^{N} \rho^\theta(u_k/x_k) \cdot f(x_{k+1}/x_k, u_k)$$

$$\nabla_\theta \ln \left(P_0(x_0) \cdot \prod_{k=0}^{N} \rho^\theta(u_k/x_k) \cdot f(x_{k+1}/x_k, u_k) \right) d\bar{z}_0^N =$$

$$\mathbb{E}\left[G_0^N(\bar{Z}_0^N) \sum_{k=0}^{N} \nabla_\theta \ln \rho^\theta(U_k | X_k) \right]$$

Plugging this into the gradient ascent rule gives REINFORCE:

$$\theta_{i+1} := \theta_i + \alpha \, \mathbb{E}\left[G_0^N(\bar{Z}_0^N) \sum_{k=0}^{N} \nabla_\theta \ln \rho^\theta(U_k | X_k) \right]$$

$$= \theta_i + \alpha \, \mathbb{E}\left[\sum_{k=0}^{N} \nabla_\theta \ln \rho^\theta(U_k | X_k) \cdot \sum_{i=0}^{N} \gamma^i r(X_i, U_i) \right]$$