* Actor

Reminder: causal policy-gradient estimator

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=t}^{T-1} r(s_{t'}, a_{t'})\right)\right]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T-1} \nabla_\theta \log \pi_\theta(a_{i,t}, s_{i,t})\left(\sum_{t'=t}^{T-1} r(s_{i,t'}, a_{i,t'})\right)$$

we can subtract $\underbrace{\text{a function of state}}_{\text{base, } b(s_t)}$ without adding bias

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=t}^{T-1} r(s_{t'}, a_{t'}) - b(s_{t'})\right)\right]$$

where $b(s_t)$ is baseline,

$$b(s_t) = \mathbb{E}\left[\sum_{t'=t}^{T-1} r(s_{t'}, a_{t'})\right]$$

Intuition: we only prioritize actions that do better than $\underbrace{\text{average}}_{b(s_t)}$

Actor-Critic

We can briefly write down overall transition,

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T-1} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \underbrace{\hat{Q}^\pi(s_{i,t}, a_{i,t})}_{\Downarrow} \qquad \left(\sum_{t'=t}^{T-1} r_{t'}\right) \Downarrow$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T-1} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \left(r(s_{i,t}, a_{i,t}) + \gamma \hat{V}^\pi_\phi(s_{i,t+1})\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T-1} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \Big(\underbrace{r(s_{i,t}, a_{i,t}) + \gamma \hat{V}^\pi_\phi(s_{i,t+1})}_{\hat{Q}^\pi_\phi(s_{i,t}, a_{i,t})} - \hat{V}^\pi_\phi(s_{i,t})\Big)$$

$$\hat{A}^\pi(s_{i,t}, a_{i,t}) = \hat{Q}^\pi_\phi(s_{i,t}, a_{i,t}) - \hat{V}^\pi_\phi(s_{i,t})$$

↑ Advantage function.

+ lower variance (due to critic)

− not unbiased (since critic is not perfect)

Policy Gradient

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \left(\sum_{t'=t}^{T-1} r(s_{t'}, a_{t'}) - b(s_{t'})\right)$$

+ no bias

− higher variance