UC San Diego
JACOBS SCHOOL OF ENGINEERING

ECE/SIOC 228 Machine Learning for Physical Applications
Lecture 15: Deep Reinforcement Learning II

Yuanyuan Shi

Assistant Professor, ECE

University of California, San Diego

Reference & Acknowledgement: David Silver's Course on RL https://www.davidsilver.uk/teaching/
Sergey Levine, Deep Reinforcement Learning https://rail.eecs.berkeley.edu/deeprlcourse/

Episodic task: $J(\theta) = E_{\tau \sim p_\theta(\tau)}[\underbrace{\sum_{t=0}^{T-1} \gamma^t R(s_t, a_t)}_{r(\tau)}]$

$r(\tau)$ is the return (total discounted reward) of an episode

$\tau = s_0, a_0, s_1, a_1, s_2, a_2, \cdots, s_{T-1}, a_{T-1}, s_T$

Find the best θ that optimize $J(\theta)$:

$$\theta^* = \arg\max_\theta J(\theta) = \arg\max_\theta E_{\tau \sim p_\theta(\tau)}[r(\tau)]$$

Review: Policy Gradient

UC San Diego

- Use Gradient ascent: $\theta_{t+1} \leftarrow \theta_t + \alpha \nabla_\theta J(\theta)$
- $\nabla_\theta J(\theta) = \nabla_\theta E_z[r(z)] = \nabla_\theta \int p_\theta(z) r(z) dz = \int \nabla_\theta p_\theta(z) r(z) dz$

$$= \int p_\theta(z) \frac{\nabla_\theta p_\theta(z)}{p_\theta(z)} r(z) dz = E_{z\sim p_\theta(z)}[\nabla_\theta \log p_\theta(z)\, r(z)]$$

$$= E_{z\sim p_\theta(z)}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\, r(z)\right]$$

A property used in derivation: $p_\theta(z) = p(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t)\, p(s_{t+1}|s_t, a_t)$

Note:
- In some reference, time index starts from $t=1, \cdots, T$
- If T is finite, we can also choose $\gamma \in (0, 1]$

The REINFORCE Algorithm

UC San Diego

Standard policy gradient:

$$\nabla_\theta J(\theta) = E_{\tau \sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\, r(\tau)\right]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i)\sum_{t=0}^{T-1}\gamma^t r(s_t^i, a_t^i)$$

★ Causality: policy at time t' cannot affect reward at time $t < t'$

★ Fix:

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

return: G_t^i

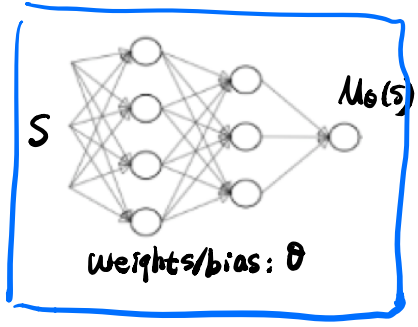
REINFORCE Algorithm

1. sample $\{\tau^i\}$ from $\pi_\theta(a_t|s_t)$ (run the policy)
2. $\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_i \sum_t \nabla_\theta \log \pi_\theta\left(a_t^i|s_t^i\right) G_t^i$
3. $\theta \leftarrow \theta + \alpha\, \nabla_\theta J(\theta)$

score function.

- For continous action space, a Gaussian policy is natural
- Gaussian policy: $a \sim N(\mu_\theta(s), \sigma^2)$
- $\pi_\theta(a|s) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{a-\mu_\theta(s)}{\sigma}\right)^2\right]$

  $\log \pi_\theta(a|s) = -\log \sigma\sqrt{2\pi} - \frac{1}{2}\left(\frac{a-\mu_\theta(s)}{\sigma}\right)^2$



- Score function:

  $\nabla_\theta \log \pi_\theta(a|s) = \frac{(a-\mu_\theta(s))\,\nabla_\theta \mu_\theta(s)}{\sigma^2}$

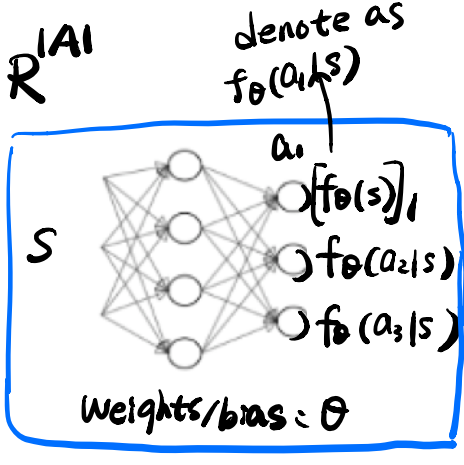
- For discrete action space, we commonly use a

Categorical policy (use softmax activation)

$$\text{Softmax}(z) = \frac{e^{z}}{\sum_{k=1}^{|A|} e^{z_i}} \leftarrow \text{probabilities} \in \mathbb{R}^{|A|}$$

- For example, $\pi_\theta(a_i|s) = \frac{e^{f_\theta(a_i|s)}}{\sum_{k=1}^{|A|} e^{f_\theta(a_k|s)}}$

$$\log \pi_\theta(a_i|s) = f_\theta(a_i|s) - \underbrace{\log \sum_{k=1}^{|A|} e^{f_\theta(a_k|s)}}_{\text{RHS}}$$



Score function: $\nabla_\theta \log \pi_\theta(a|s) = \nabla_\theta f_\theta(a|s) - \nabla_\theta \text{RHS}$

$$\nabla_\theta \text{RHS} = \nabla_\theta \log \sum_{k=1}^{|A|} e^{f_\theta(a_k|s)}$$

$$= \frac{\nabla_\theta \sum_{k=1}^{|A|} e^{f_\theta(a_k|s)}}{\sum_{k=1}^{|A|} e^{f_\theta(a_k|s)}} = \frac{\sum_{k=1}^{|A|} e^{f_\theta(a_k|s)} \cdot \nabla_\theta f_\theta(a_k|s)}{\sum_{k=1}^{|A|} e^{f_\theta(a_k|s)}}$$

$$= \sum_{k=1}^{|A|} \pi_\theta(a_k|s) \nabla_\theta f(a_k|s)$$

$$\nabla \log \pi_\theta(a|s) = \nabla_\theta f_\theta(a|s) - E_{a \sim \pi_\theta(\cdot|s)} [\nabla_\theta f_\theta(a|s)]$$

Implementing Policy Gradients

UC San Diego

Pseudocode example:

REINFORCE Algorithm:

```
# Given:
# states: [N*T, state_dim] tensor of states
# actions:  [N*T, action_dim] tensor of actions
# returns: [N*T, 1] tensor of returns

update_policy(states, actions, returns):
    logprob = calculate_logprob(states, actions) #this should return [N*T, 1] actions log probability
    loss = -torch.mean(logprob*returns)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
```

$\log \pi_\theta(a_t^i | s_t^i)$

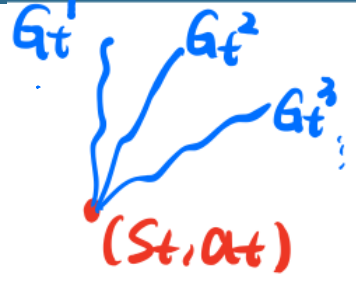
$\log \pi_\theta(a_t^i | s_t^i) \cdot G_t^i$

UC San Diego

REINFORCE Algorithm

$$\nabla_\theta J(\theta) = E_{\tau \sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) G_t\right]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^i | s_t^i) G_t^i$$



REINFORCE with Baseline

Can we do that??

$$\nabla J(\theta) = E_{\tau \sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(a_t|s_t)\left(G_t - b(s_t)\right)\right]$$

To see why, we will show

$$E_{\tau \sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\, b(s_t)\right] = 0$$

Due to linearity of expectation, it suffices to show that for any time step t, $E_{\tau \sim p_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\, b(s_t)\right] = 0$

$$E_{\tau \sim p_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\, b(s_t)\right] \qquad s_0, a_0, s_1, a_1, \cdots, s_t \,|\, a_t, \cdots, s_{T-1}, a_{T-1}, s_T$$

$$= E_{s_{0:t}, a_{0:t-1}}\left[E_{s_{t+1:T}, a_{t:T}}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\, b(s_t)\right]\right]$$

$$= E_{s_{0:t}, a_{0:t-1}}\left[b(s_t) E_{s_{t+1:T}, a_{t:T}}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\right]\right]$$

$$= E_{s_{0:t}, a_{0:t-1}}\left[b(s_t)\, E_{a_t}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\right]\right]$$

$$E_{a_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)] = \int \pi_\theta(a_t|s_t) \frac{\nabla_\theta \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} da_t$$

$$= \nabla_\theta \int \pi_\theta(a_t|s_t) da_t$$

$$= \nabla_\theta 1 = 0$$

Thus: $E_{\tau \sim p_\theta(\tau)}[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)] = 0, \forall t$

✱ Substracting a baseline $b(s_t)$ is unbiased in expectation

$$E_{\tau \sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(a_t|s_t)(G_t - b(s_t))\right] = E_{\tau \sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla \log \pi_\theta(a_t|s_t) G_t\right]$$

Then... Why substract baseline $b(S_t)$? $\Rightarrow$ reduce variance

$$\mathrm{Var}\left(\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\,(G_t - b(s_t))\right)$$

$$\approx \sum_{t=0}^{T-1} \mathrm{Var}\left(\nabla_\theta \log \pi_\theta(a_t|s_t)\,(G_t - b(s_t))\right) \quad \text{(assume independence)}$$

$$\mathrm{Var}[x] = E[x^2] - E[x]^2$$

$$= \sum_{t=0}^{T-1} \underbrace{E\left[\left(\nabla_\theta \log \pi_\theta(a_t|s_t)\,(G_t - b(s_t))\right)^2\right]}_{①} - \underbrace{\left(E\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\,(G_t - b(s_t))\right]\right)^2}_{②}$$

$$② = E\left[\nabla_\theta \log \pi_\theta(S_t|a_t)\, G_t\right]^2$$

$$① = E\left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t)\right)^2 (G_t - b(S_t))^2\right]$$

$$= E\left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t)\right)^2 \left(G_t^2 - 2G_t b(S_t) + b(S_t)^2\right)\right]$$

$$= \underbrace{E\left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t)\right)^2 G_t^2\right]}_{①a} \underbrace{- 2E\left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t)\right)^2 G_t\, b(S_t)\right]}_{①b}$$

$$+ \underbrace{E\left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t)\right)^2 b^2(S_t)\right]}_{①c}$$

$$\mathrm{Var}(\nabla_\theta \log \pi_\theta(a_t|s_t) G_t) = (1a) - (2)$$ REINFORCE

$$\mathrm{Var}(\nabla_\theta \log \pi_\theta(a_t|s_t)(G_t - b(s_t)))$$ REINFORCE with Baseline

$$= (1a) - (1b) + (1c) - (2)$$

Choose $b^*(s_t)$ that minimize $(-(1b) + (1c))$, thus to obtain variance reduction

$$-\underbrace{\nabla b}_{\ell} + \nabla C = E_{S_{0:t}, a_{0:t-1}} \Big\{ -2\, b(S_t)\, E_{a_{t:T-1},\, S_{t+1:T}} \left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t) \right)^2 G_t \right] + b^2(S_t)\, E_{a_{t:T-1},\, S_{t+1:T}} \left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t) \right)^2 \right] \Big\}$$

Let $\alpha = E_{a_{t:T-1},\, S_{t+1:T}} \left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t) \right)^2 | S_t \right]$

$$\beta = E_{a_{t:T-1},\, S_{t+1:T}} \left[\left(\nabla_\theta \log \pi_\theta(a_t|S_t) \right)^2 G_t | S_t \right]$$

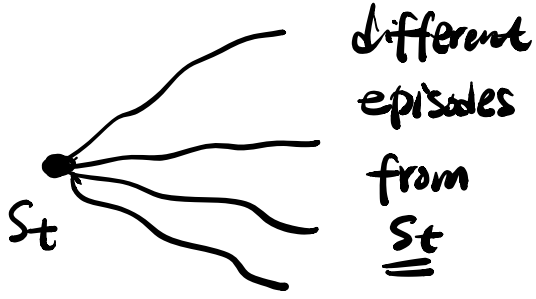
$$\min_{b(S_t)} -2\, b(S_t) \cdot \beta + 2\, b^2(S_t) \Rightarrow b^*(S_t) = \frac{\beta}{2}$$

$$b^*(S_t) = \frac{E_{A_{t:T-1}, S_{t+1:T}}\left[\left(\nabla_\theta \log \pi_\theta(A_t|S_t)\right)^2 G_t \mid S_t\right]}{E_{A_{t:T-1}, S_{t+1:T}}\left[\left(\nabla_\theta \log \pi_\theta(A_t|S_t)\right)^2 \mid S_t\right]}$$ ← expected return from S_t, weighted by $\left[\nabla_\theta \log \pi_\theta(A_t|S_t)\right]^2$

- Therefore, if we drop the "weight terms", we obtain a commonly used baseline: $b(S_t) = V^{\pi_\theta}(S_t) = E_{\pi_\theta}[G_t|S_t] = E_{A_{t:T-1}, S_{t+1:T}}[G_t|S_t]$
- With $b^*(S_t)$, $\ell^* = -\frac{\beta^2}{2} < 0 \Rightarrow$ variance reduction!

$$V^{\pi_\theta}(s_t) = E_{\pi_\theta}[G_t | s_t]$$

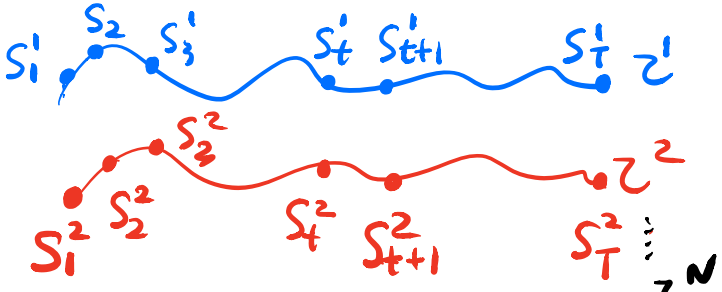
$$= E_{a_{t:T-1}, s_{t+1:T}}[G_t | s_t]$$



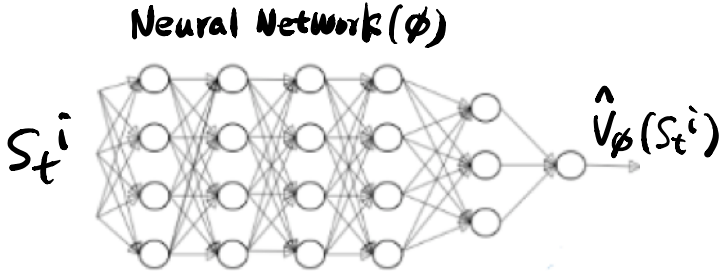
(sample)

$$\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}^i, a_{t'}^i)$$

- Require us to reset simulator to state s_t and generate N episodes.
- If $N=1$. $V^{\pi_\theta}(s_t) \approx \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})$ (can be noisy)



N episodes with length T

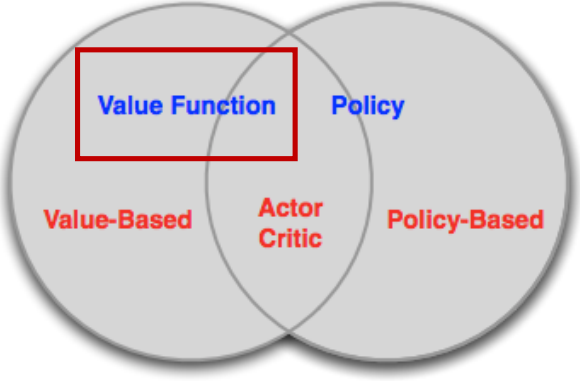


Training data: $\left\{ \left(s_t^i, \underbrace{\sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}^i, a_{t'}^i)}_{G_t^i} \right) \right\}$

Supervised learning $L(\phi) = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \left(\hat{V}_\phi^{\pi_\theta}(s_t^i) - G_t^i \right)^2$

Value Function Fitting

UC San Diego



- **More on value-based deep reinforcement learning after 2 guest lectures:**
  - Machine Learning for Power Systems
  - Machine Learning for Robotics

- We will also talk about combining policy-based + value-based DRL → **actor-critic method**!