CS 295: Optimal Control and Reinforcement Learning

Winter 2020

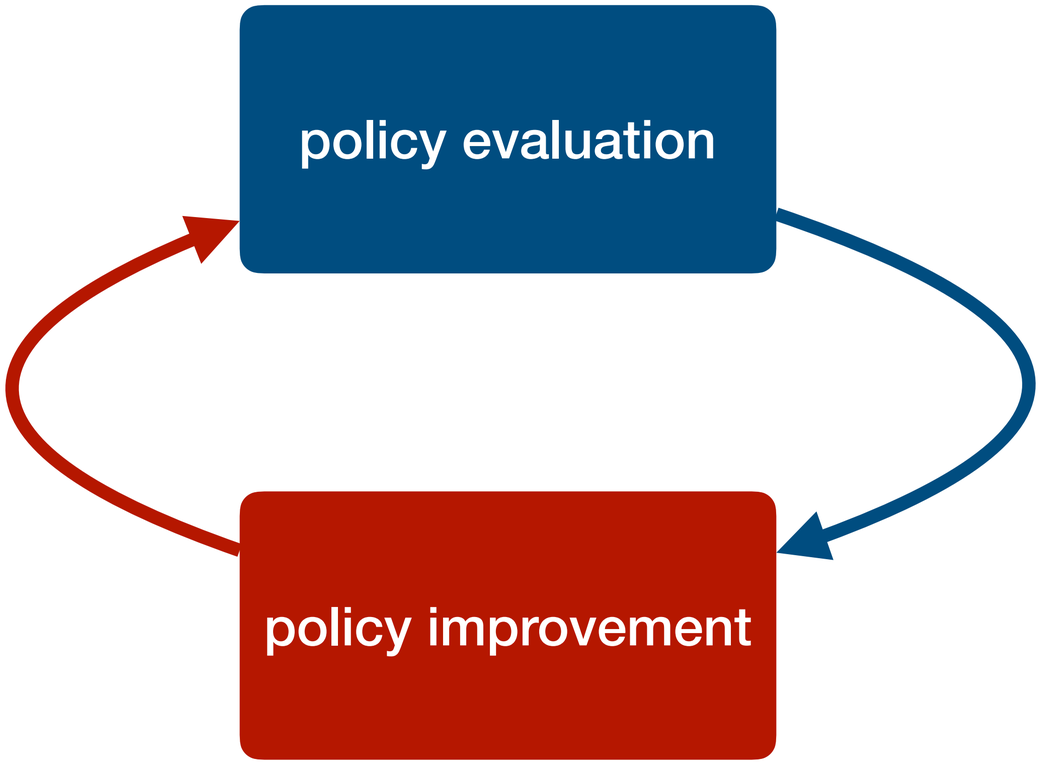
Lecture 7: Actor–Critic Methods

Roy Fox
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

Today's lecture

- Policy Gradient with learned value function
- Actor–Critic methods
- Advantage estimation
- n-step TD and TD(λ)

Policy-based methods



e.g. MC

differentiable policy

Policy Gradient (PG) with reward-to-go

$$
\begin{aligned}
\nabla_\theta \mathcal{J}_\theta &= \mathbb{E}_{\xi \sim p_\theta}[\nabla_\theta \log p_\theta(\xi) R] \\
&= \mathbb{E}_{\xi \sim p_\theta}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) R\right] \\
&= \sum_t \mathbb{E}_{s_t \sim p_\theta}[\mathbb{E}_{a_t|s_t \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t) R]] \\
&= \sum_t \mathbb{E}_{s_t \sim p_\theta}[\nabla_\theta \, \mathbb{E}_{a_t|s_t \sim \pi_\theta}[R]] \\
&= \sum_t \mathbb{E}_{s_t \sim p_\theta}\left[\nabla_\theta \, \mathbb{E}_{a_t|s_t \sim \pi_\theta}\left[\sum_{t'} r_{t'}\right]\right] \\
&= \sum_t \mathbb{E}_{s_t \sim p_\theta}\left[\nabla_\theta \, \mathbb{E}_{a_t|s_t \sim \pi_\theta}\left[\sum_{t' \geqslant t} r_{t'}\right]\right]
\end{aligned}
$$

- With finite horizon:
  - Gradient estimator

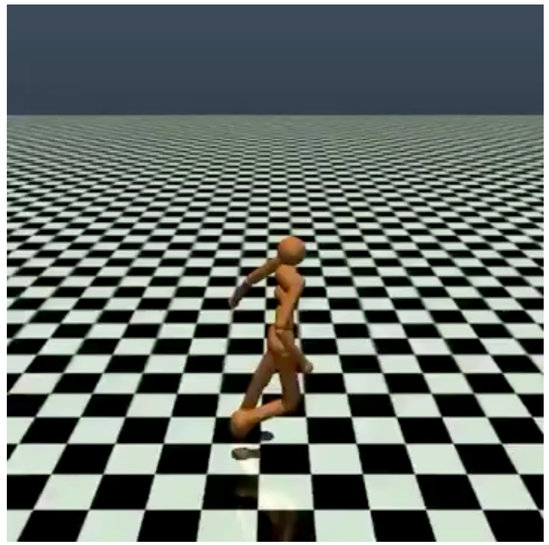
    independent of past rewards

Discounted case

$$\nabla_\theta \mathcal{J}_\theta = \sum_t \mathbb{E}_{s_t \sim p_\theta} \mathbb{E}_{a_t | s_t \sim \pi_\theta} \left[\nabla_\theta \log \pi_\theta(a_t | s_t) \sum_{t' \geqslant t} \gamma^{t'} r_{t'} \right]$$

$$= \sum_t \gamma^t \mathbb{E}_{s_t \sim p_\theta} \mathbb{E}_{a_t | s_t \sim \pi_\theta} \left[\nabla_\theta \log \pi_\theta(a_t | s_t) \sum_{t' \geqslant t} \gamma^{t'-t} r_{t'} \right]$$

- Does this make sense?
- Discounting is effectively finite-horizon
- If any state could be initial state
  - Discounting is just a computational / statistical trick
  - Don't discount the contribution of s_t to the loss
  - That's what most algorithms do



Policy-Gradient Theorem

$$\nabla_\theta \mathcal{J}_\theta = \sum_t \gamma^t \, \mathbb{E}_{s_t \sim p_\theta}[\mathbb{E}_{a_t|s_t \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t) R_{\geqslant t}]]$$

$$\overset{?}{=}\!\!\text{!} \sum_t \gamma^t \, \mathbb{E}_{s_t \sim p_\theta}[\mathbb{E}_{a_t|s_t \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t) Q_{\pi_\theta}(s_t, a_t)]]$$

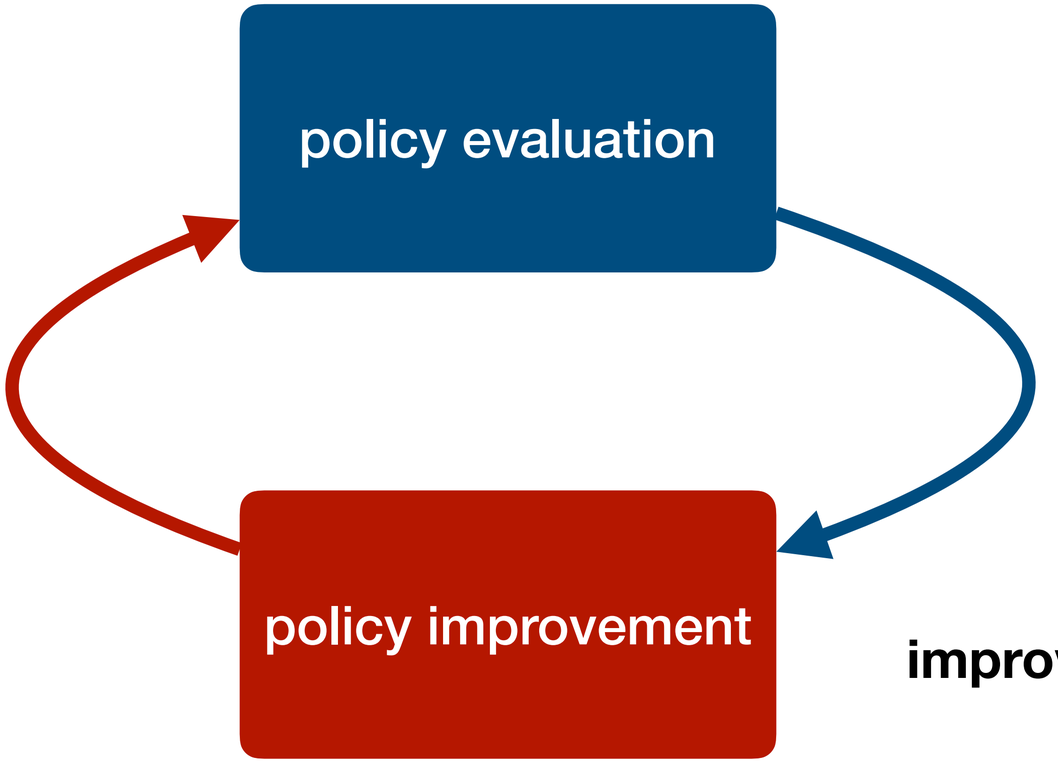
$$\begin{aligned}
\nabla_\theta V_{\pi_\theta}(s) &= \nabla_\theta \, \mathbb{E}_{a|s \sim \pi_\theta}[Q_{\pi_\theta}(s, a)] \\
&= \sum_a (\nabla_\theta \pi_\theta(a|s) Q_{\pi_\theta}(s, a) + \pi_\theta(a|s) \nabla_\theta Q_{\pi_\theta}(s, a)) \\
&= \mathbb{E}_{a|s \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s, a) + \nabla_\theta (r(s, a) + \gamma \, \mathbb{E}_{s'|s,a}[V_{\pi_\theta}(s')])] \\
&= \mathbb{E}_{a|s \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s, a) + \gamma \, \mathbb{E}_{s'|s,a}[\nabla_\theta V_{\pi_\theta}(s')]]
\end{aligned}$$

- Back-prop ~ Bellman recursion

Actor–Critic methods

$$\mathcal{L}_\theta(s,a) = -\nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a)$$

$$\mathcal{L}_\phi(s,a,r,s') = (r + \gamma\, \mathbb{E}_{a'|s' \sim \pi_\theta}[Q_{\bar{\phi}}(s',a')] - Q_\phi(s,a))^2$$



critic:
evaluates the actor

actor:
improves using the critic's advice

Baselines

$$\nabla_\theta \mathcal{J}_\theta = \mathbb{E}[\nabla_\theta \log \pi_\theta(a|s)(Q_{\pi_\theta}(s,a) - b)|s,a]$$

- b can be any variable independent of a given s
  - Can depend on the past, but not the future
- Previously, we used $b = \frac{1}{N}\sum_i R_i$
- This suggests using $b = V_{\pi_\theta}(s)$

Advantage estimation

$$\nabla_\theta \mathcal{J}_\theta = \mathbb{E}[\nabla_\theta \log \pi_\theta(a|s) A_{\pi_\theta}(s,a)|s,a]$$

- How to estimate $A_{\pi_\theta}(s,a)$?

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s) = r(s,a) + \gamma\, \mathbb{E}_{s'|s,a\sim p}[V_\pi(s')] - V(s)$$

$$\hat{A}(s,a) \approx r + \gamma V_\phi(s') - V_\phi(s)$$

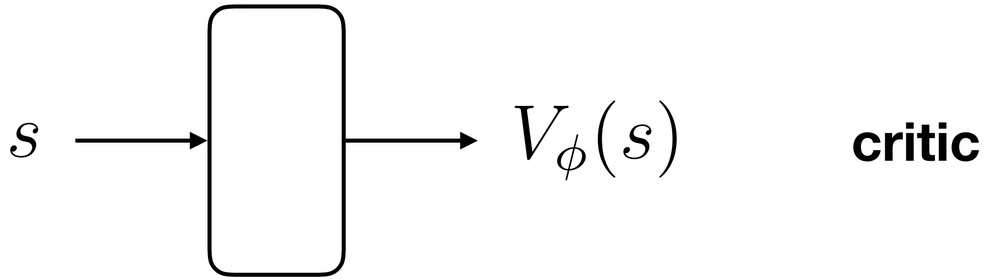
An Actor–Critic algorithm

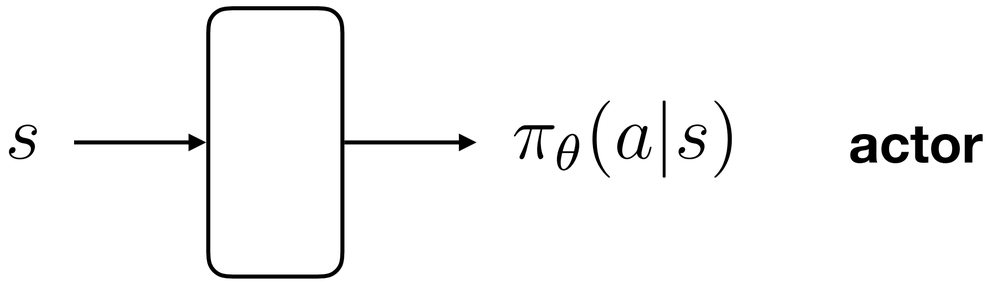
Algorithm 1 Actor–Critic

get on-policy sample (s, a, r, s')
take gradient step on $\mathcal{L}_\phi = (r + \gamma V_{\bar{\phi}}(s') - V_\phi(s))^2$
compute $\hat{A}(s, a) = r + \gamma V_\phi(s') - V_\phi(s)$
take gradient step $\nabla_\theta \log \pi_\theta(a|s)\hat{A}(s, a)$
repeat

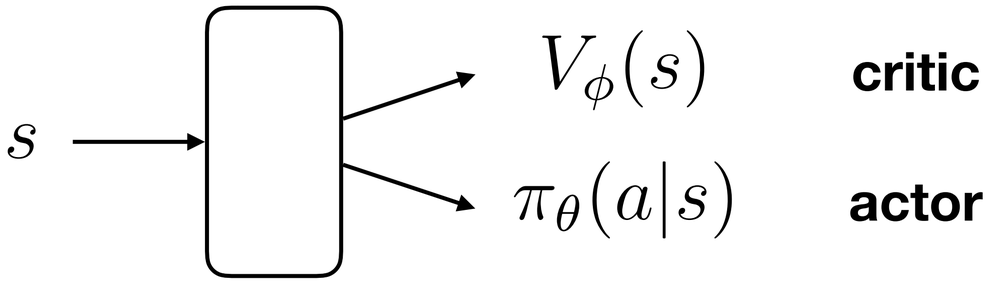
Practical considerations: param sharing

- Separate parameters:



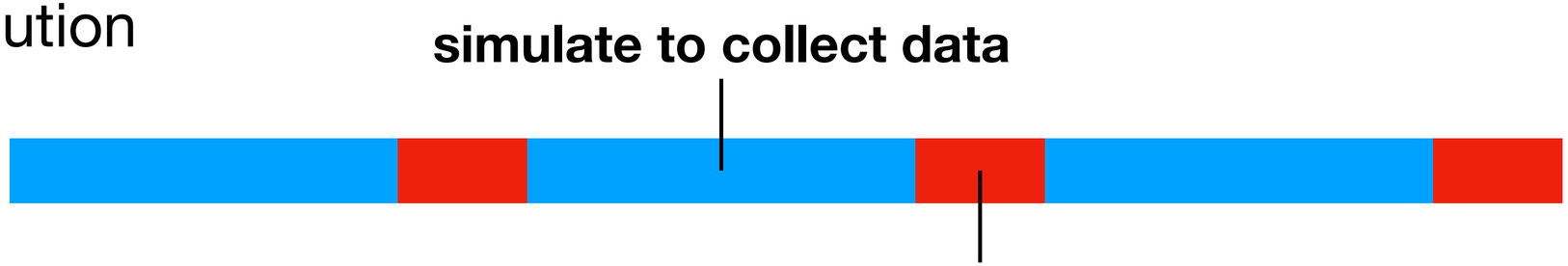


- Shared parameters:
  - Can be more data efficient
  - Can be less stable

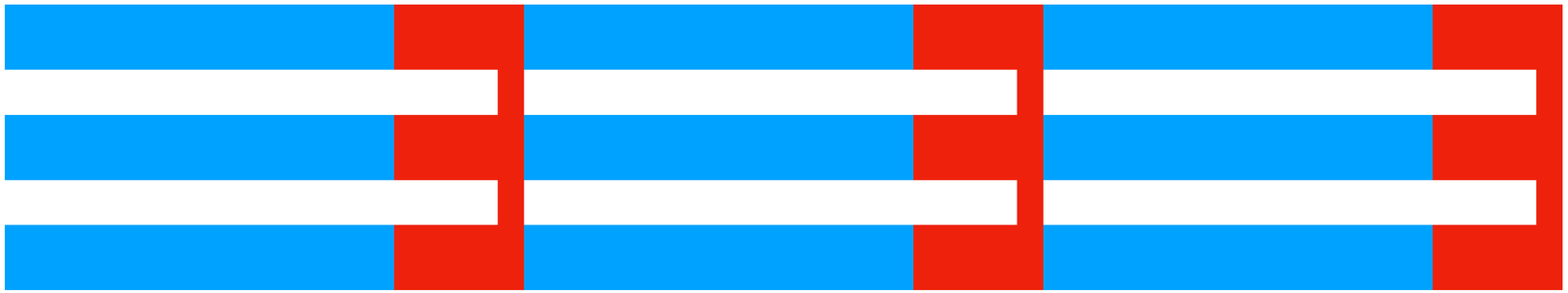


Practical considerations: distributed comp.

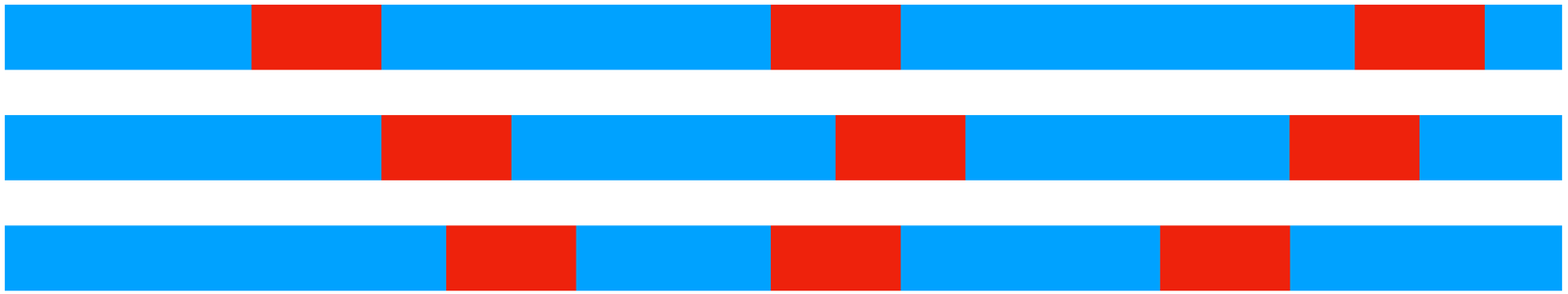
- Serial execution



- Synchronous parallel execution



- Asynchronous parallel execution



More advantage estimation

$$\hat{A}(s_t, a_t) \approx \sum_{t' \geqslant t} \gamma^{t'-t} r_{t'} - V_\phi(s_t)$$

- Asynchronous Advantage Actor Critic (A3C):
  - MC advantage estimation + asynchronous parallel execution
- Advantage Actor Critic (A2C): same but serial

Comparing advantage estimators

	bias	variance
$\nabla_\theta \mathcal{J}_\theta \approx \nabla_\theta \log \pi_\theta(a\|s) \left(\sum_{t' \geqslant t} \gamma^{t'-t} r_{t'} - b \right)$	none	high one grad per traj
$\nabla_\theta \mathcal{J}_\theta \approx \nabla_\theta \log \pi_\theta(a\|s)(r + \gamma V_\phi(s') - V_\phi(s))$	some approx value	lower
$\nabla_\theta \mathcal{J}_\theta \approx \nabla_\theta \log \pi_\theta(a\|s) \left(\sum_{t' \geqslant t} \gamma^{t'-t} r_{t'} - V_\phi(s) \right)$	none	mid state-dependent baseline

n-step TD

- 1-step TD: $\hat{A}_t^1 = r_t + \gamma V(s_{t+1}) - V(s_t)$
- 2-step TD: $\hat{A}_t^2 = r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2}) - V(s_t)$

...

- n-step TD: $\hat{A}_t^n = r_t + \cdots + \gamma^{n-1} r_{t+n-1} + \gamma^n V(s_{t+n}) - V(s_t)$
- In the limit: MC $\hat{A}_t^\infty = r_t + \gamma r_{t+1} + \cdots - V(s_t)$

TD(λ)

- How to choose n?
- Any specific n is hard truncation of the window of evidence we consider
- Instead, use "exponential window"
  - Take n-step TD with weight proportional to λ^{n-1}, where $0 \leqslant \lambda \leqslant 1$
- Now with $\hat{A}_t^n = \sum_{t'=0}^{n-1} \gamma^{t'} \hat{A}_{t+t'}^1 = \sum_{t'=0}^{n-1} \gamma^{t'} \left(r_{t+t'} + \gamma V\left(s_{t+t'+1}\right) - V\left(s_{t+t'}\right) \right)$

$$\hat{A}_t^\lambda = (1-\lambda) \sum_n \lambda^{n-1} \hat{A}_t^n = (1-\lambda) \sum_n \lambda^{n-1} \sum_{t'=0}^{n-1} \gamma^{t'} \hat{A}_{t+t'}^1$$

$$= (1-\lambda) \sum_{t'} \gamma^{t'} \hat{A}_{t+t'}^1 \sum_{n \geqslant t'+1} \lambda^{n-1} = \sum_{t'} (\lambda\gamma)^{t'} \hat{A}_{t+t'}^1$$

Generalized Advantage Estimation (GAE(λ))

$$\nabla_\theta \mathcal{J}_\theta \approx \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'} (\lambda\gamma)^{t'} \hat{A}^1_{t+t'}$$

$$\hat{A}^1_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

- GAE(0) = 1-step; GAE(1) = MC

Recap

- Policy-Gradient Theorem
- State-dependent baselines
- Actor–Critic methods
  - Advantage estimation
  - Practical considerations
- n-step TD and TD(λ)