Multi-Agent Systems

Introduction to Reinforcement Learning

Part 4: Policy-Gradient and Actor-Critic

Eric Pauwels (CWI & VU)

December 12, 2023

Outline

Policy gradient algorithms

Policy gradient algorithms

- optimise policy directly,
- NOT via value function (indirectly)

Ingredients:

1. Parametrised policy $\pi_\theta(a \mid s)$ (θ to be determined)
2. Objective function $J(\theta)$ to be maximised
3. Update rule: $\theta_{new} \leftarrow \theta_{old}$, specifically **gradient** ascent:

$$\theta_{new} \leftarrow \theta_{old} + \nabla_\theta J(\theta_{old})$$

Policy gradient: Objective Function

- Trajectory (episodic):

$$\tau = \{s_0, a_0, r_1, s_1, a_1, r_2, s_2, \ldots\ldots, s_{T-1}, a_{T-1}, r_T, s_T\}$$

- Cumulative return along trajectory

$$R(\tau) := \sum_{t=1}^{T} \gamma^{t-1} r_t \qquad \text{or} \qquad R_s(\tau) := \sum_{t=s}^{T} \gamma^{t-s} r_t$$

- Objective function $J(\theta)$: Quantifying policy performance:

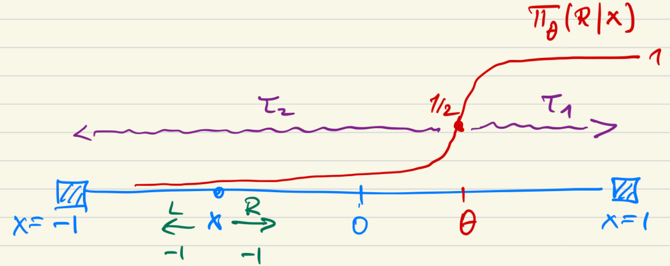
$$J(\theta) := \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)]$$

- **Goal**:

$$\max_{\theta} J(\theta)$$

Policy gradient: Example(1)

Absorbing states at $x = \pm 1$, absorption reward $= 0$



$$R(\tau_1) = -(1-\theta) \qquad p(\tau_1, \theta) = 1/2$$

$$R(\tau_2) = -(1+\theta) \qquad p(\tau_2 | \theta) = 1/2$$

$$\boxed{R(\tau_2) < R(\tau_1)}$$

Policy gradient theorem

- Objective function (as path integral and MC version)

$$J(\theta) := \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)] = \int R(\tau) p(\tau \mid \theta)\, d\tau \approx \frac{1}{N} \sum_{\tau \sim \pi_\theta} R(\tau)$$

(Path integral makes dependence on θ explicit)

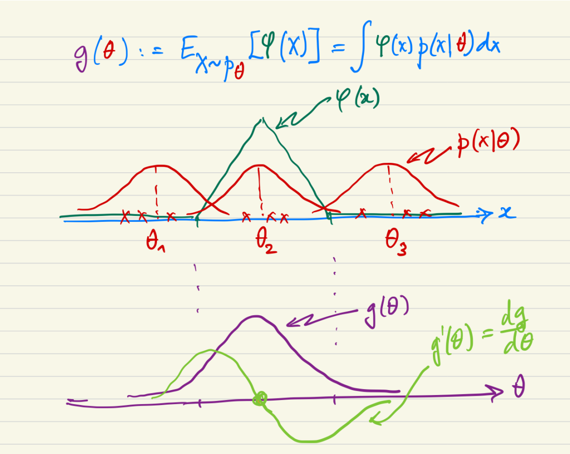
- In abstract terms (to simplify notation):

$$g(\theta) := \mathbb{E}_{X \sim p_\theta}[\phi(X)] = \int \phi(x) p(x \mid \theta)\, dx$$

- Need to compute derivative (to optimise):

$$\frac{d}{d\theta} g(\theta) = \frac{d}{d\theta} \int \phi(x) p(x \mid \theta)\, dx = \int \phi(x) \frac{d}{d\theta} \left(p(x \mid \theta)\right)\, dx$$

Policy gradient theorem



Optimal value $\theta^* = \theta_2$

Policy gradient theorem

- Maximise $g(\theta) = \mathbb{E}_{X\sim p_\theta}[\phi(X)] \approx \frac{1}{n}\sum_{X_i\sim p_\theta}\phi(X_i)$,
- Intuition: Sample $X_i \sim p(x\,|\,\theta)$
  - If $\phi(X_i)$ large: change θ to make X_i more likely;
  - If $\phi(X_i)$ small: change θ to make X_i less likely;
- Mathematics: Policy gradient theorem

$$\frac{dg}{d\theta} = \frac{d}{d\theta}\mathbb{E}_{X\sim p_\theta}[\phi(X)] = \mathbb{E}_{X\sim p_\theta}\left[\phi(X)\frac{d}{d\theta}\left(\log p(X\,|\,\theta)\right)\right]$$

Change θ such that

- If $\phi(X_i)$ large: make X_i more likely, i.e. $\dfrac{d(\log)p(X_i)}{d\theta} > 0$
- If $\phi(X_i)$ small: make X_i less likely, i.e. $\dfrac{d(\log)p(X_i)}{d\theta} < 0$

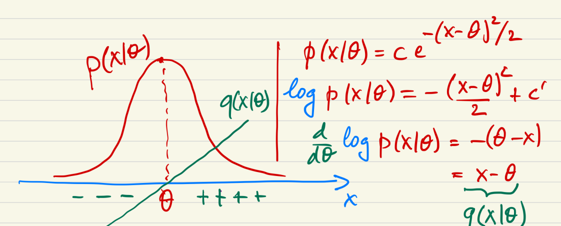
Policy gradient theorem

$$
\begin{aligned}
\frac{d}{d\theta} g(\theta) &= \int \phi(x) \frac{d}{d\theta}\left(p(x \mid \theta)\right) dx \\
&= \int \phi(x) \left[\frac{\frac{d}{d\theta}\left(p(x \mid \theta)\right)}{p(x \mid \theta)} \right] p(x \mid \theta)\, dx \\
&= \int \phi(x) \left[\frac{\frac{d}{d\theta}\left(p(x \mid \theta)\right)}{p(x \mid \theta)} \right] p(x \mid \theta)\, dx \\
&= \int \phi(x) \left[\frac{d}{d\theta}\left(\log p(x \mid \theta)\right) \right] p(x \mid \theta)\, dx \\
&= \mathbb{E}_{X \sim p_\theta} \left[\phi(X) \frac{d}{d\theta}\left(\log p(x \mid \theta)\right) \right]
\end{aligned}
$$

$$
\boxed{\frac{d}{d\theta} \mathbb{E}_{X \sim p_\theta} [\phi(X)] = \mathbb{E}_{X \sim p_\theta} \left[\phi(X) \frac{d}{d\theta}\left(\log p(X \mid \theta)\right) \right]}
$$

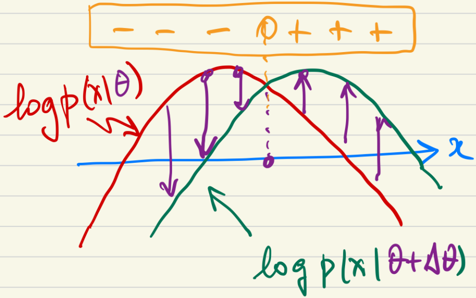
Policy gradient theorem

$$
\begin{aligned}
\frac{d}{d\theta} g(\theta) &= \frac{d}{d\theta} \mathbb{E}_{X \sim p_\theta} [\phi(X)] \\
&= \mathbb{E}_{X \sim p_\theta} \left[\phi(X) \frac{d}{d\theta} (\log p(X \mid \theta)) \right] \\
&\approx \frac{1}{n} \sum_{X_i \sim p_\theta} \phi(X_i) \frac{d}{d\theta} (\log p(X_i \mid \theta)) \\
&= \frac{1}{n} \sum_{X_i \sim p_\theta} \phi(X_i) q(X_i \mid \theta)
\end{aligned}
$$



Policy gradient theorem

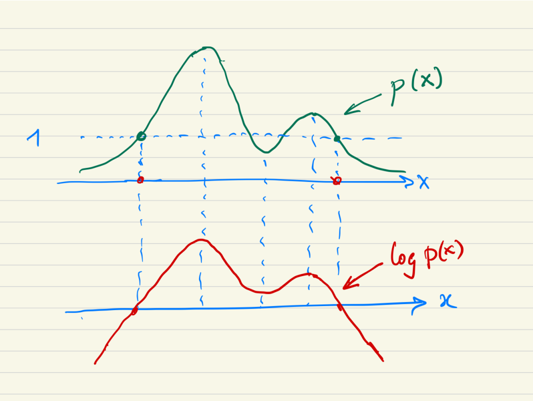
$$\nabla_\theta g(\theta) = \mathbb{E}_{X\sim\pi_\theta}\left[\varphi(X)\,\nabla_\theta \log p(X;\theta)\right]$$



at given X
how does $\log p(X|\theta)$
change if θ increases
i.e. $\theta \to \theta + \Delta\theta$

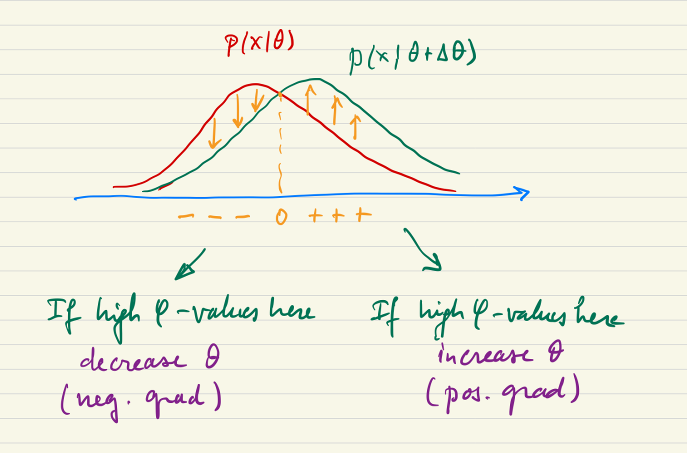
$p(x)$ vs $\log p(x)$

- $p(x)$ vs $\log p(x)$ have same local extremes, increasing and decreasing behaviour
- $sgn(\nabla_\theta p(x \mid \theta)) = sgn(\nabla_\theta \log p(x \mid \theta))$

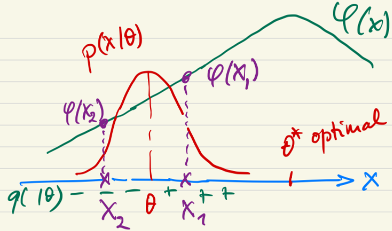


Policy gradient theorem

$$\frac{dg}{d\theta} \approx \frac{1}{n} \sum_{X_i \sim p_\theta} \phi(X_i) \frac{d}{d\theta} (\log p(X_i \mid \theta))$$



Policy gradient theorem

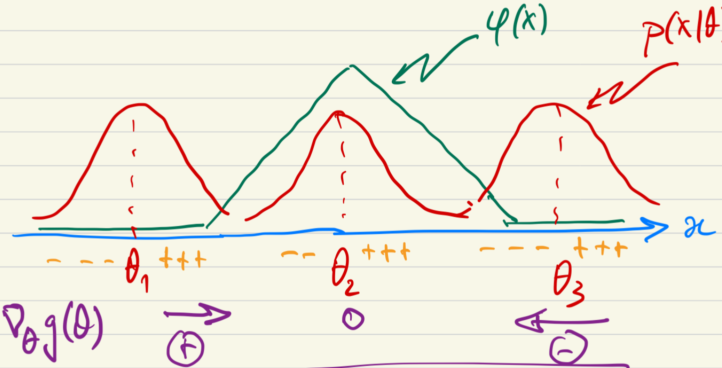


$$\nabla_\theta g(\theta) = \mathbb{E}_{X \sim p_\theta}\left[\varphi(X)\, \nabla_\theta \log p(X|\theta)\right]$$

$$\simeq \frac{1}{n} \sum_{X_i \sim p_\theta} \varphi(X_i)\, \underbrace{\nabla_\theta \log p(X_i|\theta)}_{q(X_i|\theta)}]$$

$$\sim \varphi(X_1)\, \underbrace{q(X_1|\theta)}_{+} + \varphi(X_2)\, \underbrace{q(X_2|\theta)}_{-} > 0$$

Policy gradient theorem



$$\boxed{\theta_{new} \leftarrow \theta + \alpha \nabla_\theta g(\theta)}$$

Policy gradient theorem

make positive — MAXIMISATION

$$\nabla_\theta g(\theta) \approx \frac{1}{n} \sum_{x_i \sim p_\theta} \varphi(x_i) \underbrace{\nabla_\theta \log p(x_i|\theta)}$$

change θ s.t.

if $\varphi(x_i)$ HIGH $\longrightarrow$ $p(x_i|\theta)$ increases ($\nabla_\theta \log p > 0$)

if $\varphi(x_i)$ LOW $\longrightarrow$ $p(x_i|\theta)$ decreases ($\nabla_\theta \log p < 0$)

Policy gradient theorem

General result

$$\nabla_\theta \mathbb{E}_{X\sim p_\theta}\left[\phi(X)\right] = \mathbb{E}_{X\sim p_\theta}\left[\phi(X)\nabla_\theta \log p(X \mid \theta)\right]$$

Policy Gradient Theorem

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}\left[R(\tau)\right] = \mathbb{E}_{\tau\sim\pi_\theta}\left[R(\tau)\nabla_\theta \log p(\tau \mid \theta)\right]$$

- If $R(\tau)$ high, change θ to make τ MORE likely, i.e. $p(\tau \mid \theta) \uparrow$
- If $R(\tau)$ low, change θ to make τ LESS likely, i.e. $p(\tau \mid \theta) \downarrow$

Policy gradient theorem

$$p(\tau \mid \theta) = \prod_{t \geq 0} \underbrace{p(s_{t+1} \mid s_t, a_t)}_{\text{MDP env.}} \underbrace{\pi_\theta(a_t \mid s_t)}_{\text{policy}}$$

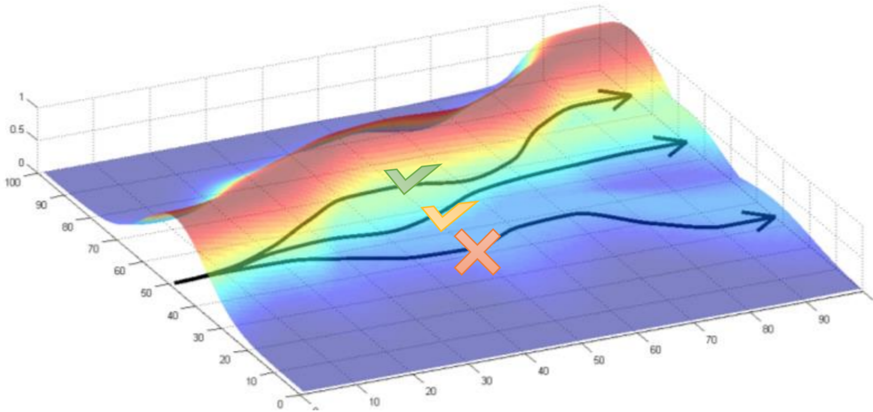
$$\log p(\tau \mid \theta) = \sum_{t \geq 0} [\log p(s_{t+1} \mid s_t, a_t) + \log \pi_\theta(a_t \mid s_t)]$$

$$\nabla_\theta \log p(\tau \mid \theta) = \sum_{t \geq 0} \nabla_\theta \log \pi_\theta(a_t \mid s_t)$$

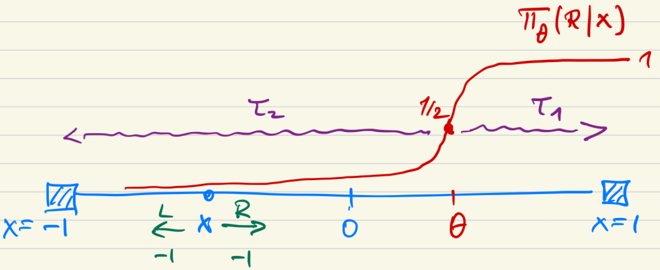
Policy Gradient Theorem

$$\begin{aligned} \nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)] = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau) \nabla_\theta \log p(\tau \mid \theta)] \\ &= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} R_t(\tau) \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right] \end{aligned}$$

Policy gradient illustration



Policy gradient: Example(1)

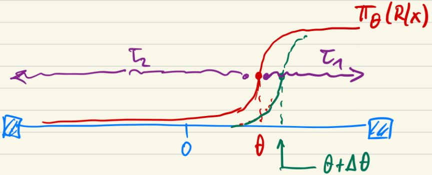


$$R(\tau_1) = -(1-\theta) \qquad p(\tau_1,\theta) = 1/2$$

$$R(\tau_2) = -(1+\theta) \qquad p(\tau_2|\theta) = 1/2$$

$$\boxed{R(\tau_2) < R(\tau_1)}$$

Policy gradient: Example(2)



$$p(\tau_1|\theta+\Delta\theta) < p(\tau_1|\theta) \longrightarrow \nabla_\theta \log p(\tau_1|\theta) < 0 \quad (=-\delta)$$

$$p(\tau_2|\theta+\Delta\theta) > p(\tau_2|\theta) \longrightarrow \nabla_\theta \log p(\tau_2|\theta) > 0 \quad (=+\delta)$$

$$\nabla_\theta J \simeq R(\tau_1)\cdot\underbrace{\nabla_\theta \log p(\tau_1|\theta)}_{-\delta} + R(\tau_2)\underbrace{\nabla_\theta \log p(\tau_2|\theta)}_{+\delta}$$

$$\simeq \delta\left[-R(\tau_1) + R(\tau_2)\right]$$

$$\simeq \delta\left[|R(\tau_1)| - |R(\tau_2)|\right] < 0 \implies \theta\downarrow$$

Recall: reward(path) = - length(path)

REINFORCE algorithm

Example of policy gradient algo

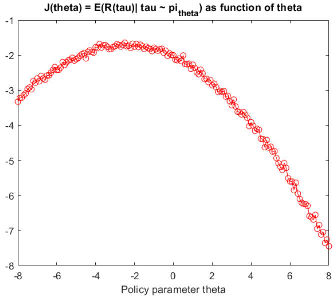
1. Initialise parameter θ of policy π_θ, learning rate α
2. **for** episode $= 1 \ldots$ NR_EPISODES:
3. Sample trajectory $\tau = \{s_0, a_0, r_1, s_1, \ldots, r_T, s_T\}$
4. Set $\nabla_\theta J(\theta) = 0$
5. # add gradient contributions along trajectory
6. **for** $t = 0, 1, \ldots, T$:
7. $R_t(\tau) = \sum_{u=t}^{T} \gamma^{u-t} r_u$,
8. $\nabla_\theta J(\theta) = \nabla_\theta J(\theta) + R_t(\tau) \nabla_\theta \log \pi_\theta(a_t \mid s_t)$
9. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$ # update policy parameter

NOTE: REINFORCE algo is on-policy:

τ cannot be re-used when policy (i.e. θ) changes!

Worked example

- Linear state space (-10:10), left and right absorbing
- $R_{left} = 0, R_{right} = 5, R_{NT} = -1$



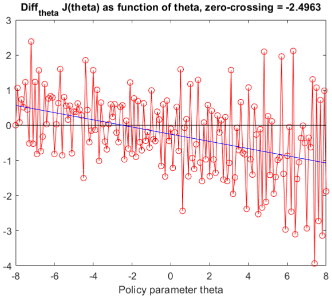


Figure: Left: MC computation for each value of θ, (Right) Numerical gradient (diff), zero crossing at $\theta = -2.5$

Worked example

- Linear state space (-10:10), left and right absorbing
- $R_{left} = 0, R_{right} = 5, R_{NT} = -1$

$$\nabla_\theta J(\theta) = \sum_t R_t(\tau) \nabla_\theta \log \pi_\theta(a_t \mid s_t)$$

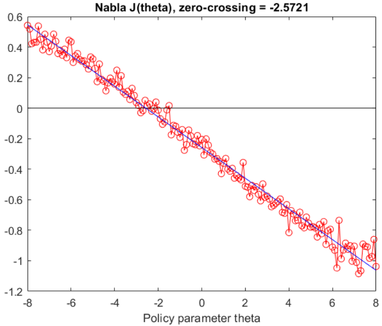


Figure: $\nabla_\theta J(\theta)$ via policy gradient theorem, zero crossing at $\theta = -2.57$

Improving REINFORCE algorithm

- Policy gradient estimate has high variance
  (trajectories might be very different!)

$$\nabla_\theta J(\theta) \approx \sum_{t=0}^{T} R_t(\tau) \nabla_\theta \log \pi_\theta(a_t \mid s_t)$$

- Variance reduction by introducing action-independent baseline:

$$\nabla_\theta J(\theta) \approx \sum_{t=0}^{T} (R_t(\tau) - b(s_t)) \nabla_\theta \log \pi_\theta(a_t \mid s_t)$$

- Example: Actor-Critic algorithm:

$$R_t(\tau) = q_{\pi_\theta}(s_t, a_t) \qquad \text{and} \qquad b(s_t) = v_{\pi_\theta}(s_t)$$