

Reinforcement Learning and REINFORCE

David S. Rosenberg

NYU: CDS

April 11, 2021

Contents

- 1 Markov Decision Processes
- 2 Episodic Learning
- 3 Policies and Value Functions
- 4 REINFORCE
- 5 The Policy Gradient Theorem
- 6 Monte Carlo for implementation
- 7 Proof of Policy Gradient Theorem
- 8 Proof of Policy Gradient Theorem II

Markov Decision Processes

[Online] Stochastic k -armed contextual bandit

Stochastic k -armed contextual bandit

- 1 Environment samples **context** and **rewards vector** jointly, iid, for each round:

$$(X, R), (X_1, R_1), \dots, (X_T, R_T) \in \mathcal{X} \times \mathbb{R}^k \text{ i.i.d. from } P,$$

where $R_t = (R_t(1), \dots, R_t(k)) \in \mathbb{R}^k$.

- 2 For $t = 1, \dots, T$,

- 1 Our algorithm **selects action** $A_t \in \mathcal{A} = \{1, \dots, k\}$ based on X_t and history

$$\mathcal{D}_t = \left((X_1, A_1, R_1(A_1)), \dots, (X_{t-1}, A_{t-1}, R_{t-1}(A_{t-1})) \right).$$

- 2 Our algorithm **receives reward** $R_t(A_t)$.

- We **never observe** $R_t(a)$ for $a \neq A_t$.

Generalizing from contextual bandits

- Contextual bandits: contexts X_1, \dots, X_T are i.i.d.
- What about playing a video game, driving a car, moving a robot arm?
- Next context depends on previous context and action selected.
- We now want to allow dependence between consecutive X_i 's.
- This is the **main difference** between reinforcement learning and contextual bandits.

Markov decision processes (MDPs)

“MDPs are a mathematically idealized form of the reinforcement learning problem for which precise theoretical statements can be made.” [SB18, p. 47]

Markov decision processes

- Learner / decision maker is called the **agent**
- Agent interacts with the **environment**
- Each round $t = 0, 1, 2, 3, \dots$,
 - agent receives a **state** $X_t \in \mathcal{X}$.
 - agent selects an action $A_t \in \mathcal{A}$
 - agent receives a reward $R_t \in \mathbb{R}$
- We get a **trajectory**: $X_0, A_0, R_0, X_1, A_1, R_1, X_2, A_2, R_2, X_3, \dots$

MDPs, continued

- The **dynamics** of the MDP are given by

$$\mathbb{P}(X_{t+1} = x', R_t = r \mid X_t = x, A_t = a) = p(x', r \mid x, a),$$

for any $x', x \in \mathcal{X}$, $r \in \mathbb{R}$, $a \in \mathcal{A}$.

- Gives distribution of reward and next state given previous state and action.
- Note: For simplicity, below we assume that rewards and states are discrete
 - The final algorithms will not require this. (Still need finite action space.)

Key points

- 1 The reward and the next state are **generated jointly**.
 - Why? e.g. allows next state to contain information about reward
- 2 Note that the transition probabilities have no explicit dependence on time.
 - Though we can always include time into the state x .

Episodic Learning

Episodic learning

- Often problem breaks up into “**episodes**” or “**trials**”.
- For an episode there is a final time step T
 - need not be the same in every episode
 - it's typically random.
- Sometimes the task just continues, without natural breaks.
- These are called **continuing tasks**.
- In episodic learning, we typically update our policy after every episode.
- In continuing tasks, we have to update as we go
- We'll consider the episodic case, but things are similar for continuing case.

- We can denote the trajectories for each episode as

Episode 1: $X_{1,0}, A_{1,0}, R_{1,0}, X_{1,1}, A_{1,1}, R_{1,1}, X_{1,2}, A_{1,2}, R_{1,2}, X_{1,3}$

Episode 2: $X_{2,0}, A_{2,0}, R_{2,0}, X_{2,1}, A_{2,1}, R_{2,1}, X_{2,2}, A_{2,2}, R_{2,2}, X_{2,3}, A_{2,3}, R_{2,3}, X_{2,4}$

Episode 3: $X_{3,0}, A_{3,0}, R_{3,0}, X_{3,1}, A_{3,1}, R_{3,1}, X_{3,2}$

\vdots \vdots

- However, we'll find we usually only need to refer to one episode at a time.
- So we'll usually leave off the episode subscript, and just use a subscript for roundya

- I think of each episode as the analogue of a single round of a contextual bandit. In fact, if each episode ends after round 1, it's exactly the contextual bandit setting (assuming we set things up as described in a previous note, where round 0 starts in a fixed start state, but the state distribution in round 1 is the same as the context distribution in the contextual bandit). So an episode is kind of an expanded version of a contextual bandit round.

Start and terminal states

- For simplicity (and w.l.o.g.), assume we always start in a special **start state** $x_0 \in \mathcal{X}$.
- We'll also assume we have a **terminal state** $x_{\text{stop}} \in \mathcal{X}$.
- The terminal state is an “absorbing” state: once we arrive, we never leave.
- We get no reward in the terminal state.
- Formally, this means:

$$p(x', r \mid x_{\text{stop}}, a) = \mathbb{1}[x' = x_{\text{stop}}] \mathbb{1}[r = 0].$$

- So we'll say that T is the last round of the MDP if $X_T \neq x_{\text{stop}}$ and

$$\begin{aligned} X_{T+1} &= X_{T+2} = \cdots = x_{\text{stop}} \\ R_{T+1} &= R_{T+2} = \cdots = 0 \end{aligned}$$

- How can we say that starting in start state x_0 is not a loss in generality? Suppose we want to start in a random state given by $p_0(x)$. Then we can define $p(x_1, r_0 | x_0, a_0) = p_0(x_1) \mathbb{1}[r_0 = 0]$. In words, no matter what action is taken in round 0, the state distribution in round 1 is $p_0(x)$, as desired, and the reward received in round 0 is 0. That way the MDP is equivalent to the MDP that starts at round 1 with initial state distribution $p_0(x)$.
- Note that with our stop state convention, we can write the total reward received in an episode in two ways:

$$\sum_{t=0}^T R_t = \sum_{t=0}^{\infty} R_t$$

Assumption: Bounded episode lengths

- We will assume there is some known integer $T_0 < \infty$ such that

$$\mathbb{P}(T \leq T_0) = 1.$$

- In words: every episode terminates at or before T_0 rounds.
- This seems reasonable from a practical perspective. We can take T_0 arbitrarily large.
- From a theoretical perspective, I don't see another clean way to make the derivations rigorous.
- Specifically, the points of concern would be
 - interchanging expectations with a sum over the rounds of a random episode and
 - solving the recurrence relation in the proof of the Policy Gradient Theorem.

Policies and Value Functions

- A policy for an MDP at round t
 - gives a conditional distribution over action A_t
 - conditioned on the state X_t .
- In this module, we consider policies parameterized by θ : $\pi_\theta(a | x)$, for $\theta \in \mathbb{R}^d$.
- At round t , action $A_t \in \mathcal{A} = \{1, \dots, k\}$ is chosen according to

$$\mathbb{P}(A_t = a | X_t = x) = \pi_\theta(a | x).$$

- Our policy parameter θ will be **fixed** for each episode.
- However, our policy can still “learn”, in a certain sense, within an episode.
- Unlike contextual bandit setting, in each round of an episode,
 - the state X_t can summarize the history of play since the beginning of the episode.

The state-value function

- In contextual bandits, the **value** of a policy is the expected reward.
- In MDPs, we define a couple different value functions for a policy.

Definition (State-value function)

The **state-value function** for policy π , denoted $v_\pi(x)$ is the expected reward starting in state x and following π thereafter:

$$v_\pi(x) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} R_k \mid X_0 = x \right] \quad \forall x \in \mathcal{X}.$$

- With the convention that $X_0 = x_0$, the value of a policy is $v_\pi(x_0)$.

The action-value function

Definition (Action-value function)

The **action-value function** for policy π , denoted $q_\pi(x, a)$ is the expected reward starting in state x , taking action a , and following π thereafter:

$$q_\pi(x, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} R_k \mid X_0 = x, A_0 = a \right] \quad \forall x \in \mathcal{X}, a \in \mathcal{A}.$$

- Since the dynamics are time-indepent, it would be equivalent to make the definition

$$q_\pi(x, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} R_{k+t} \mid X_t = x, A_t = a \right],$$

and similarly for the definition of the state-value function.

The value functions

- Exercise: Write $v_\pi(x)$ in terms of $q_\pi(x, a)$. (Let $G = \sum_{t=0}^{\infty} R_t$.)

$$\begin{aligned}v_\pi(x) &= \mathbb{E}_\pi[G \mid X_0 = x] \\&= \mathbb{E}_\pi[\mathbb{E}_\pi[G \mid A_0, X_0 = x] \mid X_0 = x] \\&= \sum_a \pi(a \mid x) \mathbb{E}_\pi[G \mid A_0 = a, X_0 = x] \\&= \sum_a \pi(a \mid x) q_\pi(x, a)\end{aligned}$$

- Concept checks: In this inner expectation: $\mathbb{E}_{\pi}[G \mid A_0, X_0 = x]$, why did we indicate a dependency on π in the expectation?
 - Answer: Although the reward R_0 has nothing to do with the policy distribution, since we're conditioning on A_0 and X_0 , all subsequent rewards will be affected by the policy distribution.

Intuition builder / lemma for later

Show: $q_\pi(x, a) = \mathbb{E}[R_t | (X_t, A_t) = (x, a)] + \sum_{x'} p(x' | x, a) v_\pi(x')$.

Proof: Then

$$\begin{aligned} q_\pi(x, a) &= \mathbb{E}_\pi \left[R_0 + \sum_{k=1}^{\infty} R_k | (X_0, A_0) = (x, a) \right] \\ &= \mathbb{E}_\pi \left[\mathbb{E}_\pi \left[R_0 + \sum_{k=1}^{\infty} R_k | X_1, R_0, (X_0, A_0) = (x, a) \right] | (X_0, A_0) = (x, a) \right] \\ &= \mathbb{E}_\pi \left[R_0 + \mathbb{E}_\pi \left[\sum_{k=1}^{\infty} R_k | X_1 \right] | (X_0, A_0) = (x, a) \right] \\ &= \mathbb{E}[R_0 | (X_0, A_0) = (x, a)] + \mathbb{E}[v_\pi(X_1) | (X_0, A_0) = (x, a)] \\ &= \mathbb{E}[R_0 | (X_0, A_0) = (x, a)] + \sum_{x'} p(x' | x, a) v_\pi(x') \end{aligned}$$

REINFORCE

Policy gradient for contextual bandits

- We took a “policy gradient” approach to contextual bandits.
- The idea was to find the policy $\pi_{\theta}(a | x)$ that optimized

$$J(\theta) = \mathbb{E}_{\theta} [R(A)].$$

- We found that

$$R_t(A_t) \nabla_{\theta} \log \pi_{\theta_t}(A_t | X_t)$$

was an unbiased estimate of $\nabla J(\theta)$.

- We used that to form an SGD-style optimization algorithm:

$$\theta_{t+1} \leftarrow \theta_t + \eta R_t(A_t) \nabla_{\theta} \log \pi_{\theta_t}(A_t | X_t)$$

Policy gradient for MDPs

- What if we think about each action in an episode as a separate round of a contextual bandit?
- Then our update would be

$$\theta_{t+1} \leftarrow \theta_t + \eta R_t \nabla_{\theta} \log \pi_{\theta_t}(A_t | X_t).$$

- The problem: action now may affect reward many rounds later, but update does not reflect this
- Another approach: use the total episode reward for each round of an episode:

$$\theta_{t+1} \leftarrow \theta_t + \eta \left[\sum_{i=1}^{\infty} R_t \right] \nabla_{\theta} \log \pi_{\theta_t}(A_t | X_t).$$

- This could work...

- But one thing doesn't seem quite right with

$$\theta_{t+1} \leftarrow \theta_t + \eta \left[\sum_{i=1}^{\infty} R_t \right] \nabla_{\theta} \log \pi_{\theta_t}(A_t | X_t).$$

- Action A_t can be penalized by poor rewards received at time $t-1$.
- Seems to make more sense to only include rewards received after A_t :

$$\theta_{t+1} \leftarrow \theta_t + \eta \left[\sum_{i=t}^{\infty} R_t \right] \nabla_{\theta} \log \pi_{\theta_t}(A_t | X_t).$$

- This is the basic REINFORCE update, which we will derive in the next section.

The Policy Gradient Theorem

Policy gradient theorem for MDPs (I)

- The policy gradient theorem states¹ that

$$\nabla J(\theta) = \sum_x \eta(x) \sum_a [q_\theta(x, a) \nabla_\theta \pi_\theta(a | x)]$$

where

$$\eta(x) := \mathbb{E}_\theta \left[\sum_{k=0}^{\infty} \mathbb{1}[X_k = x] \mid X_0 = x_0 \right].$$

- Note that $\eta(x)$ is the expected number of visits to state x in an episode,
 - when we start in state $X_0 = x_0$ and
 - select actions according to π_θ .

¹Our convention here and below is that \sum_x excludes x_{stop} .

Interpretation (I)

- For any state x , $\nabla_{\theta}\pi_{\theta}(a|x)$ is the direction to move θ
 - to make a more likely (in state x).
- $q_{\theta}(x, a)$ is the expected future rewards for action a in state x , and $A \sim \pi_{\theta}$ after that.
- So $\sum_a [q_{\theta}(x, a)\nabla_{\theta}\pi_{\theta}(a|x)]$ is a weighted average of policy updates
 - where we make action a more likely (in state x)
 - in proportion to the future rewards associated with that action.
- That's a sensible improvement to the policy π_{θ} for state x .
- How do we improve the policy for all states?

$$\nabla J(\theta) = \sum_x \eta(x) \sum_a [q_{\theta}(x, a)\nabla_{\theta}\pi_{\theta}(a|x)]$$

takes a weighted average of the updates that improve each state x , in proportion to how often we expect to be in state x .

Policy gradient theorem for MDPs (II)

- We'll also show that

$$\nabla J(\theta) = \mathbb{E}_{\theta} \left[\sum_{t=0}^{T_0} \sum_a [q_{\theta}(X_t, a) \nabla_{\theta} \pi_{\theta}(a | X_t)] \right],$$

where the expectation is over a single episode X_1, \dots, X_T played according to π_{θ} .

- Recall that $T_0 < \infty$ is our assumed maximum episode length.
- This is the form of the policy gradient theorem most amenable to deriving REINFORCE.
 - (At least that I'm aware of.)

Monte Carlo for implementation

Episode-level Monte Carlo

- Consider

$$\nabla J(\theta) = \mathbb{E}_{\theta} \left[\sum_{t=0}^{T_0} \sum_a q_{\theta}(X_t, a) \nabla_{\theta} \pi_{\theta}(a | X_t) \right].$$

where the expectation is over a single episode X_1, \dots, X_T played according to π_{θ} .

- We can do a one-episode Monte Carlo estimate of $\nabla J(\theta)$:

$$\sum_{t=0}^T \sum_a q_{\theta}(X_t, a) \nabla_{\theta} \pi_{\theta}(a | X_t).$$

- This will be an unbiased estimate of $\nabla J(\theta)$.

All-actions method

- We don't know $q_\theta(X_t, a)$, but we can plug-in an action-value estimate $\hat{q}_\theta(x, a)$, fit to historical data:

$$\sum_{t=0}^T \sum_a \hat{q}_\theta(X_t, a) \nabla_\theta \pi_\theta(a | X_t).$$

- This is called an **all-actions** method.
- This estimate is biased, since \hat{q}_θ will generally be biased,
 - but we expect it to have lower variance than the REINFORCE method discussed next.
- If the action space is too large to sum over,
 - we can sample actions $A_t \sim \pi_\theta(a | X_t)$ as we did for contextual bandits.

- For an unbiased estimate, we use our “clever trick” with logs:

$$\begin{aligned}
 \nabla J(\theta) &= \mathbb{E}_{\theta} \left[\sum_{t=0}^T \sum_a q_{\theta}(X_t, a) \nabla_{\theta} \pi_{\theta}(a | X_t) \right] \\
 &= \mathbb{E}_{\theta} \left[\sum_{t=0}^T \sum_a q_{\theta}(X_t, a) \pi_{\theta}(a | X_t) \nabla_{\theta} \log \pi_{\theta}(a | X_t) \right] \\
 &= \mathbb{E}_{\theta} \left[\sum_{t=0}^T \mathbb{E}_{A_t \sim \pi_{\theta}(a|X_t)} [q_{\theta}(X_t, A_t) \nabla_{\theta} \log \pi_{\theta}(A_t | X_t) | X_t] \right] \\
 &= \mathbb{E}_{\theta} \left[\sum_{t=0}^T \mathbb{E}_{A_t \sim \pi_{\theta}(a|X_t)} \left[\mathbb{E}_{\theta} \left[\sum_{k=t}^{\infty} R_k | X_t, A_t \right] \nabla_{\theta} \log \pi_{\theta}(A_t | X_t) | X_t \right] \right]
 \end{aligned}$$

REINFORCE (II)

$$\begin{aligned}\nabla J(\theta) &= \mathbb{E}_{\theta} \left[\sum_{t=0}^T \mathbb{E}_{A_t \sim \pi_{\theta}(a|X_t)} \left[\mathbb{E}_{\theta} \left[\sum_{k=t}^{\infty} R_k \mid X_t, A_t \right] \nabla_{\theta} \log \pi_{\theta}(A_t \mid X_t) \mid X_t \right] \right] \\&= \sum_{t=0}^{T_0} \mathbb{E}_{\theta} \left[\mathbb{E}_{A_t \sim \pi_{\theta}(a|X_t)} \left[\mathbb{E}_{\theta} \left[\nabla_{\theta} \log \pi_{\theta}(A_t \mid X_t) \sum_{k=t}^{\infty} R_k \mid X_t, A_t \right] \mid X_t \right] \right] \\&= \sum_{t=0}^{T_0} \mathbb{E}_{\theta} \left[\nabla_{\theta} \log \pi_{\theta}(A_t \mid X_t) \sum_{k=t}^{\infty} R_k \right] \quad (\text{Adam's rule}) \\&= \mathbb{E}_{\theta} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(A_t \mid X_t) \sum_{k=t}^{\infty} R_k \right]\end{aligned}$$

REINFORCE (III)

- We've derived

$$\nabla J(\theta) = \mathbb{E}_{\theta} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(A_t | X_t) \sum_{k=t}^{\infty} R_k \right]$$

- The expectation is over an episode played according to π_{θ} , starting in $X_0 = x_0$.
- We can get a one-episode Monte Carlo unbiased estimate of $\nabla J(\theta)$ as

$$\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(A_t | X_t) \sum_{k=t}^{\infty} R_k.$$

REINFORCE in Sutton and Barto

- Our proposed REINFORCE makes a single update per episode:

$$\theta \leftarrow \theta + \eta \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(A_t | X_t) \sum_{k=t}^{\infty} R_k$$

- REINFORCE in [SB18, p. 328] has an update for every round of the episode,
 - but after the full episode has been run with parameter setting θ_0 .
- For each round of the episode, they make an update

$$\theta_{t+1} \leftarrow \theta_t + \eta \nabla_{\theta} \log \pi_{\theta_t}(A_t | X_t) \sum_{k=t}^{\infty} R_k.$$

- One concern: each A_t is sampled from $\pi_{\theta_0}(a | X_t)$,
 - but treating it like it was sampled from π_{θ_t} .

Proof of Policy Gradient Theorem

The objective

- Consider policy space $\pi_\theta(a | x)$.
- We'd like to find θ maximizing

$$\begin{aligned} J(\theta) &= \mathbb{E}_{\pi_\theta} \left[\sum_{i=0}^{\infty} R_i \mid X_0 = x_0 \right] \\ &= v_{\pi_\theta}(x_0). \end{aligned}$$

- Since we're only dealing with policies π_θ , we'll write

$$v_\theta(x) := v_{\pi_\theta}(x) \quad q_\theta(x, a) := q_{\pi_\theta}(x, a) \quad \mathbb{E}_\theta := \mathbb{E}_{\pi_\theta}$$

Policy gradient theorem: product rule

- Recall: $q_{\theta}(x, a) = \mathbb{E}[R_t \mid (X_t, A_t) = (x, a)] + \sum_{x'} p(x' \mid x, a) v_{\theta}(x')$.
- So $\nabla_{\theta} q_{\theta}(x, a) = \sum_{x'} p(x' \mid x, a) \nabla_{\theta} v_{\theta}(x')$.
- Then

$$\begin{aligned}\nabla_{\theta} v_{\theta}(x) &= \nabla_{\theta} \left[\sum_a \pi_{\theta}(a \mid x) q_{\theta}(x, a) \right] \\ &= \sum_a [q_{\theta}(x, a) \nabla_{\theta} \pi_{\theta}(a \mid x) + \pi_{\theta}(a \mid x) \nabla_{\theta} q_{\theta}(x, a)] \\ &= \sum_a \left[q_{\theta}(x, a) \nabla_{\theta} \pi_{\theta}(a \mid x) + \pi_{\theta}(a \mid x) \sum_{x'} p(x' \mid x, a) \nabla_{\theta} v_{\theta}(x') \right]\end{aligned}$$

- Note that this is a recurrence relation! ($\nabla_{\theta} v_{\theta}(\cdot)$ shows up on the LHS and RHS).

Cleaning up the recurrence

- Let $\mathbb{P}_\theta(x \rightarrow x', k)$ be the probab of being in state x' in k steps:
 - conditioned on starting in state x (under policy π_θ).

$$\mathbb{P}_\theta(x \rightarrow x', k) := \mathbb{P}_\theta(X_k = x' \mid X_0 = x)$$

- Let $\phi(x) = \sum_a [q_\theta(x, a) \nabla_\theta \pi_\theta(a \mid x)]$. Then

$$\begin{aligned}\nabla_\theta v_\theta(x) &= \sum_a \left[q_\theta(x, a) \nabla_\theta \pi_\theta(a \mid x) + \pi_\theta(a \mid x) \sum_{x'} p(x' \mid x, a) \nabla_\theta v_\theta(x') \right] \\&= \phi(x) + \sum_a \pi_\theta(a \mid x) \sum_{x'} p(x' \mid x, a) \nabla_\theta v_\theta(x') \\&= \phi(x) + \sum_{x'} \left[\sum_a p(x' \mid x, a) \pi_\theta(a \mid x) \right] \nabla_\theta v_\theta(x') \\&= \phi(x) + \sum_{x'} \mathbb{P}_\theta(x \rightarrow x', 1) \nabla_\theta v_\theta(x')\end{aligned}$$

Unrolling the recurrence

$$\begin{aligned}\nabla_{\theta} v_{\theta}(x) &= \phi(x) + \sum_{x'} \mathbb{P}_{\theta}(x \rightarrow x', 1) \nabla_{\theta} v_{\theta}(x') \\&= \phi(x) + \sum_{x'} \mathbb{P}_{\theta}(x \rightarrow x', 1) \left[\phi(x') + \sum_{x''} \mathbb{P}_{\theta}(x' \rightarrow x'', 1) \nabla_{\theta} v_{\theta}(x'') \right] \\&= \phi(x) + \sum_{x'} \mathbb{P}_{\theta}(x \rightarrow x', 1) \phi(x') + \sum_{x''} \left[\sum_{x'} \mathbb{P}_{\theta}(x \rightarrow x', 1) \mathbb{P}_{\theta}(x' \rightarrow x'', 1) \right] \nabla_{\theta} v_{\theta}(x'') \\&= \phi(x) + \sum_{x'} \mathbb{P}_{\theta}(x \rightarrow x', 1) \phi(x') + \sum_{x''} \mathbb{P}_{\theta}(x \rightarrow x'', 2) \nabla_{\theta} v_{\theta}(x'')\end{aligned}$$

Putting it together

$$\begin{aligned}\nabla_{\theta} v_{\theta}(x) &= \phi(x) + \sum_{x'} \mathbb{P}_{\theta}(x \rightarrow x', 1) \phi(x') + \sum_{x''} \mathbb{P}_{\theta}(x \rightarrow x'', 2) \phi(x'') \\ &\quad + \sum_{x'''} \mathbb{P}_{\theta}(x \rightarrow x''', 3) \phi(x''') + \sum_{x''''} \mathbb{P}_{\theta}(x \rightarrow x'''', 4) \nabla_{\theta} v_{\theta}(x''') \\ &= \sum_{k=0}^{T_0} \sum_{x'} \mathbb{P}_{\theta}(x \rightarrow x', k) \phi(x') + \sum_{x'} \underbrace{\mathbb{P}_{\theta}(x \rightarrow x', T_0 + 1)}_{=0} \nabla_{\theta} v_{\theta}(x') \\ &= \sum_{k=0}^{T_0} \sum_{x'} \mathbb{P}_{\theta}(x \rightarrow x', k) \phi(x')\end{aligned}$$

- To get the 2nd equality, we continue to expand the recursion for $T_0 + 1$ steps.
- In the last equality, we use our assumption that for $t > T_0$ we're always in state x_{stop} .
- And the our sum over states excludes the stop state.

Back to the objective

- We now bring in the start state:

$$\begin{aligned}\nabla J(\theta) = \nabla_{\theta} v_{\theta}(x_0) &= \sum_x \left(\sum_{k=0}^{T_0} \mathbb{P}_{\theta}(x_0 \rightarrow x, k) \right) \phi(x) \\ &= \sum_x \left(\sum_{k=0}^{T_0} \mathbb{P}_{\theta}[X_k = x \mid X_0 = x_0] \right) \phi(x) \\ &= \sum_x \left(\sum_{k=0}^{T_0} \mathbb{E}[\mathbb{1}[X_k = x] \mid X_0 = x_0] \right) \phi(x) \\ &= \sum_x \left(\mathbb{E}_{\theta} \left[\sum_{k=0}^{T_0} \mathbb{1}[X_k = x] \mid X_0 = x_0 \right] \right) \phi(x),\end{aligned}$$

where the inner expectation is over a full episode X_1, \dots, X_T played according to π_{θ} .

Conclusion (I)

- Recalling the definitions of $\eta(x)$ and then $\phi(x)$, we can write

$$\begin{aligned}\nabla J(\theta) = \nabla_{\theta} v_{\theta}(x_0) &= \sum_x \left(\mathbb{E}_{\theta} \left[\sum_{k=0}^{T_0} \mathbb{1}[X_k = x] \mid X_0 = x_0 \right] \right) \phi(x) \\ &= \sum_x \eta(x) \phi(x) \\ &= \sum_x \eta(x) \sum_a [q_{\theta}(x, a) \nabla_{\theta} \pi_{\theta}(a \mid x)]\end{aligned}$$

- The last expression is the first part of our Policy Gradient Theorem.

Proof of Policy Gradient Theorem II

Towards writing as an expectation

- We can write

$$\begin{aligned}\nabla J(\theta) &= \sum_x \eta(x) \phi(x) \\ &= \left[\frac{\sum_{x' \in \mathcal{X}'} \eta(x')}{\sum_{x' \in \mathcal{X}'} \eta(x')} \right] \sum_x \eta(x) \phi(x) \\ &= \left[\sum_{x'} \eta(x') \right] \sum_x \frac{\eta(x)}{\sum_{x' \in \mathcal{X}'} \eta(x')} \phi(x) \\ &= \left[\sum_{x'} \eta(x') \right] \sum_x \mu(x) \phi(x),\end{aligned}$$

where $\mu(x) := \eta(x) / \sum_{x' \in \mathcal{X}'} \eta(x')$.

- How should we interpret $\mu(x)$?

Interpreting $\mu(x)$ (I)

- Suppose we run E episodes with policy π_θ .
- Take the states visited in all those episodes and put them into a bag.
- Let X_E be a state drawn randomly from this bag. Let $\mu_E(x) := \mathbb{P}(X_E = x)$.
- Let \mathcal{D}_E be all the trajectories in those E episodes. Then

$$\begin{aligned}\mathbb{P}(X_E = x) &= \mathbb{E}[\mathbb{1}[X_E = x]] &= \mathbb{E}[\mathbb{E}[\mathbb{1}[X_E = x] \mid \mathcal{D}_E]] \\ &= \mathbb{E}[\mathbb{P}(X_E = x \mid \mathcal{D}_E)] \\ &= \mathbb{E}\left[\frac{\sum_{e=1}^E (\# \text{ of visits to state } x \text{ in episode } e)}{\sum_{e=1}^E T(e)}\right],\end{aligned}$$

where $T(e) = (\# \text{ rounds in episode } e)$.

- Isn't sampling from $\mu(x)$ the same as sampling a random round from a single random episode? Why do we have to say all this stuff about "putting all rounds from all episodes into a bag?"
- Suppose we have two types of episodes that occur with equal probability:
 - Type 1: Episode ends immediately after the start state x_0 .
 - Type 2: Episode has length 1000, state x_0 followed by 999 other states, not x_0 .
- Then the probability of state x_0 under $\mu(x)$ is $\mu(x_0) = \frac{1}{1001/2} = \frac{2}{1001}$.
- The probability of state x_0 under the second approach is $\frac{1}{2} \left(1 + \frac{1}{1000}\right) = \frac{1001}{2000} \approx \frac{1}{2}$.
- VERY DIFFERENT.
- Second approach makes states that occur in shorter episodes are more likely.

Interpreting $\mu(x)$ (II)

- So $\mathbb{P}(X_E = x) = \mathbb{E}[V_E(x)/L_E]$ where

$$V_E(x) = \frac{1}{E} \sum_{e=1}^E (\# \text{ of visits to state } x \text{ in episode } e)$$

$$L_E = \frac{1}{E} \sum_{e=1}^E T(e).$$

- By the SLLN, as $E \rightarrow \infty$, $V_E(x) \xrightarrow{\text{a.s.}} \eta(x)$ and $L_E \xrightarrow{\text{a.s.}} \sum_x \eta(x)$.
- Since $L_E(x) > 0$, the continuous mapping theorem implies $\frac{V_E(x)}{L_E(x)} \xrightarrow{\text{a.s.}} \frac{\eta(x)}{\sum_x \eta(x)} = \mu(x)$.
- Since $|V_E(x)/L_E| \leq 1$, by the dominated convergence theorem, we get

$$\lim_{E \rightarrow \infty} \mu_E(x) = \lim_{E \rightarrow \infty} \mathbb{P}(X_E = x) = \lim_{E \rightarrow \infty} \mathbb{E}[V_E(x)/L_E] = \mu(x).$$

- So drawing X from $\mu(x)$ is like sampling from the bag above, when $E \rightarrow \infty$.

Expectation w.r.t. $\mu(x)$

- Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be any function. Then

$$\begin{aligned}\mathbb{E}_{X \sim \mu(x)} f(X) &= \sum_x \mu(x) f(x) = \sum_{x'} \frac{\eta(x')}{\sum_x \eta(x)} f(x') \\&= \frac{1}{\sum_x \eta(x)} \sum_x \eta(x) f(x) \\&= \frac{1}{\sum_x \eta(x)} \sum_x f(x) \mathbb{E}_\theta \left[\sum_{k=0}^{T_0} \mathbb{1}[X_k = x] \mid X_0 = x_0 \right] \\&= \frac{1}{\sum_x \eta(x)} \mathbb{E}_\theta \left[\sum_{k=0}^{T_0} \sum_x f(x) \mathbb{1}[X_k = x] \mid X_0 = x_0 \right] \\&= \frac{1}{\sum_x \eta(x)} \mathbb{E}_\theta \left[\sum_{k=0}^{T_0} f(X_k) \mid X_0 = x_0 \right]\end{aligned}$$

The policy gradient in terms of an episode

- Applying the previous result to $\phi(x)$, we get

$$\begin{aligned}\nabla J(\theta) &= \left[\sum_{x'} \eta(x') \right] \sum_x \mu(x) \phi(x) \\ &= \left[\sum_{x'} \eta(x') \right] \frac{1}{\sum_x \eta(x)} \mathbb{E}_\theta \left[\sum_{k=0}^{T_0} \phi(X_k) \mid X_0 = x_0 \right] \\ &= \mathbb{E}_\theta \left[\sum_{k=0}^{T_0} \sum_a [q_\theta(X_k, a) \nabla_\theta \pi_\theta(a \mid X_k)] \mid X_0 = x_0 \right]\end{aligned}$$

where the expectation is over a single episode X_1, \dots, X_T played according to π_θ .

Policy gradient theorem for MDPs

- Summarizing our results, we have

$$\nabla J(\theta) = \sum_x \eta(x) \sum_a [q_\theta(x, a) \nabla_\theta \pi_\theta(a | x)],$$

where $\eta(x) := \mathbb{E}_\theta [\sum_{k=0}^{\infty} \mathbb{1}[X_k = x] | X_0 = x_0]$.

- We also have a version that's well-suited to episodic REINFORCE:

$$\nabla J(\theta) = \mathbb{E}_\theta \left[\sum_{t=0}^{T_0} \sum_a [q_\theta(X_t, a) \nabla_\theta \pi_\theta(a | X_t)] | X_0 = x_0 \right],$$

where the expectation is over a single episode X_1, \dots, X_T played according to π_θ .

References

- The development of Markov decision processes (MDPs) is based on [SB18, Ch 3].
- The proof for the policy gradient theorem is based on [SMSM00], which is essentially the same as the proof in [SB18, p. 325]. We deviated in making an “episodic” version.
- The presentation of the recurrence part of the policy gradient theorem proof is based on Lilian Weng’s blog, which is a good source for additional detail and discussion [Wen18].

References I

- [SB18] Richard S. Sutton and Andrew G. Barto, *Reinforcement learning: An introduction*, A Bradford Book, Cambridge, MA, USA, 2018.
- [SMSM00] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour, *Policy gradient methods for reinforcement learning with function approximation*, Advances in Neural Information Processing Systems (S. Solla, T. Leen, and K. Müller, eds.), vol. 12, MIT Press, 2000.
- [Wen18] Lilian Weng, *Policy gradient algorithms*, Apr 2018,
[https://lilianweng.github.io/lil-log/2018/04/08/
policy-gradient-algorithms.html#proof-of-policy-gradient-theorem](https://lilianweng.github.io/lil-log/2018/04/08/policy-gradient-algorithms.html#proof-of-policy-gradient-theorem).