Announcements

- None.
- Questions, comments, concerns?
Gradient estimation

- Recall, we aim to design gradient estimators, i.e., $G(\theta)$ such that

$$\mathbb{E}[G(\theta)] = \nabla_{\theta} \mathbb{E}_{X \sim q_{\theta}}[f(X, \theta)]$$

  - Assume it exists.
  - $X$ is a random variable with a prob. density $q_{\theta}$.
  - $f : \mathcal{X} \times \mathbb{R}^D \rightarrow \mathbb{R}$ is a function.

- Will briefly discuss two (pretty distinct) important ideas.
  - Policy gradient theorem.
  - Stochastic computation graphs.
Recall

- Infinite-horizon MDP, finite action space, finite state space. An agent interacts with the environment \( p(s_{t+1}|s_t, a_t) \) using a policy \( \pi_\theta(a_t|s_t) \) for \( T = \infty \) steps.

The agent’s objectives is to maximize its return:

\[
J(\theta) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]
\]

This is finite if \( r \) is bounded, but can we get gradients \( \nabla_\theta J(\theta) \) if the process is actually infinite-horizon???

- As we saw last week in one of the talks, we can simulate episodic MDPs in this framework by introducing absorbing states.
In the finite-horizon setting, i.e., $T$ is finite, we had a simple expression that we’ve seen now a couple times:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p} \left[ \sum_{t=0}^{T} r(\tau) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$$

The policy gradient theorem gives us a very simple and intuitive expression for the policy gradient in the infinite horizon setting.
Policy gradient theorem

- Recall:

\[
Q^{\pi_\theta}(s, a) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right| s_0 = s, a_0 = a \]

\[
V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) Q^{\pi_\theta}(s, a)
\]

- Let’s start by trying to compute the gradient \( \nabla_\theta V^{\pi_\theta}(s) \).
Policy gradient theorem

\[
\nabla_\theta V^{\pi_\theta}(s) = \nabla_\theta \sum_{a_0} Q^{\pi_\theta}(s, a_0) \pi_\theta(a_0|s) \\
= \sum_{a_0} [Q^{\pi_\theta}(s, a_0) \nabla_\theta \pi_\theta(a_0|s) + \pi_\theta(a_0|s) \nabla_\theta Q^{\pi_\theta}(s, a_0)]
\]

define \( g(\theta, s) = \sum_a Q^{\pi_\theta}(s, a) \nabla_\theta \pi_\theta(a|s) \)

\[
= g(\theta, s) + \sum_{a_0} \pi_\theta(a_0|s) \nabla_\theta Q^{\pi_\theta}(s, a_0) \\
= g(\theta, s) + \sum_{a_0} \left[ \pi_\theta(a_0|s) \nabla_\theta \left( r(s, a_0) + \gamma \sum_{s_1} p(s_1|s, a_0) V^{\pi_\theta}(s_1) \right) \right]
\]
Policy gradient theorem

\[
= g(\theta, s) + \sum_{a_0} \left[ \pi_\theta(a_0|s) \nabla_\theta \left( r(s, a_0) + \gamma \sum_{s_1} p(s_1|s, a_0) V^{\pi_\theta}(s_1) \right) \right]
\]

\[
= g(\theta, s) + \gamma \sum_{a_0} \sum_{s_1} \pi_\theta(a_0|s) p(s_1|s, a_0) \nabla_\theta V^{\pi_\theta}(s_1)
\]

\[
= g(\theta, s) + \gamma \sum_{a_0} \sum_{s_1} \pi_\theta(a_0|s) p(s_1|s, a_0) g(\theta, s_1)
\]

\[
+ \gamma^2 \sum_{a_0} \sum_{s_1} \sum_{a_1} \sum_{s_2} \pi_\theta(a_0|s) p(s_1|s, a_0) \pi_\theta(a_1|s_1) p(s_2|s_1, a_1) \nabla_\theta V^{\pi_\theta}(s_2)
\]
Policy gradient theorem

If we keep unrolling we get this:

$$\sum_{k=0}^{\infty} \sum_{s'} g(\theta, s') \left( \sum_{a_0:k-1} \gamma^k \pi_\theta(a_0|s)p(s_1|s, a_0) \cdots \pi_\theta(a_{k-1}|s_{k-1})p(s'|s_{k-1}, a_{k-1}) \right)$$

What the heck is this?

$$\sum_{k=0}^{\infty} \sum_{s'} g(\theta, s') \left( \sum_{a_0:k-1} \gamma^k \pi_\theta(a_0|s)p(s_1|s, a_0) \cdots \pi_\theta(a_{k-1}|s_{k-1})p(s'|s_{k-1}, a_{k-1}) \right)$$
The discounted state visitation distribution

- Define the following distribution:
  
  **Input:** Initial state \( s_0 = s \)

  flip coin with prob. \( \gamma \), init. \( k = 0 \);

  **while coin is heads do**

  \[
  a_k \sim \pi_{\theta}(\cdot | s_k) ;
  
  s_{k+1} \sim p(\cdot | s_k, a_k) ;
  
  \text{flip coin with prob. } \gamma, \text{ increment } k;
  \]

  **end**

  return \( s_k \);

- Start in \( s \), at each iteration flip a coin with \( \mathbb{P}(\text{heads}) = \gamma \), terminate if tails, else continue.
The discounted state visitation distribution

- Start in $s$, at each iteration flip a coin with $\mathbb{P}(\text{heads}) = \gamma$, terminate if tails, else continue.

- What is the probability $\mathbb{P}(\text{returned on iteration } k \text{ and } s_k = s')$?

$$\gamma^k(1 - \gamma) \sum_{a_0:k-1, s_1:k-1} \pi_\theta(a_0|s)p(s_1|s, a_0)\ldots \pi_\theta(a_{k-1}|s_{k-1})p(s'|s_{k-1}, a_{k-1})$$

- The marginal is the discounted state visitation distribution:

$$d_{\pi_\theta}^{\gamma}(s'|s) := \sum_{k=0}^{\infty} \gamma^k(1 - \gamma) \sum_{a_0:k-1, s_1:k-1} \pi_\theta(a_0|s)\ldots \pi_\theta(a_{k-1}|s_{k-1})p(s'|s_{k-1}, a_{k-1})$$
Policy gradient theorem

Let’s get back to business

\[ \nabla_\theta V^{\pi_\theta}(s) \]

\[ = \sum_{k=0}^{\infty} \sum_{s'} g(\theta, s') \left( \sum_{a_0:k-1} \gamma^k \pi_\theta(a_0|s) \pi_\theta(a_{k-1}|s_{k-1}) p(s'|s_{k-1}, a_{k-1}) \right) \]

\[ = \sum_{s'} \frac{g(\theta, s')}{1 - \gamma} d^{\pi_\theta}_\gamma (s'|s) \]

\[ = \sum_{s'} \sum_{a} d^{\pi_\theta}_\gamma (s'|s) \pi_\theta(a|s') \frac{Q^{\pi_\theta}(s', a) \nabla_\theta \log \pi_\theta(a|s')}{1 - \gamma} \]
Policy gradient theorem

- All together, with \( s_0 \sim p(s_0), \ s \sim d_{\gamma}^{\pi_{\theta}}(s|s_0), \ a \sim \pi_{\theta}(a|s) \):

\[
(1 - \gamma) \nabla_{\theta} J(\theta) = \mathbb{E} [Q^{\pi_{\theta}}(s, a) \nabla_{\theta} \log \pi_{\theta}(a|s)]
\]

- Very satisfying form! This is the **policy gradient theorem**.

- Again, we can use control variates:

\[
(1 - \gamma) \nabla_{\theta} J(\theta) = \mathbb{E} [(Q^{\pi_{\theta}}(s, a) - V^{\pi_{\theta}}(s)) \nabla_{\theta} \log \pi_{\theta}(a|s)] \tag{1}
\]

\[
= \mathbb{E} [A^{\pi_{\theta}}(s, a) \nabla_{\theta} \log \pi_{\theta}(a|s)] \tag{2}
\]

- Because of discounting, we can get an unbiased estimator of this infinite-horizon return!
Let's compare the policy gradient theorem in the infinite-horizon

\[(1 - \gamma) \nabla_\theta J(\theta) = \mathbb{E} [Q^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a|s)]\]

with the finite-horizon setting:

\[\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p} \left[ \sum_{t=0}^{T} r(\tau) \nabla_\theta \log \pi_\theta(a_t|s_t) \right]\]

Notice that the policy gradient in the infinite-horizon does not depend on the return that was actually achieved by the agent in its rollout.
This motives so-call actor-critic methods, in which the true $Q^{\pi_\theta}(s, a)$ is replaced by a learned $\hat{Q}(s, a)$.

$$E [Q^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a|s)] \approx E [\hat{Q}(s, a) \nabla_\theta \log \pi_\theta(a|s)]$$

$\hat{Q}(s, a)$ is called the critic. This is a very successful family of methods.
So far we’ve talked about:

▶ Pathwise gradient estimators.
▶ Score function gradient estimators.
▶ Control variates and baselines.
▶ Critics.

These ideas can be mixed-and-matched. How exactly to mix-and-match them is formalized in a framework called **stochastic computation graphs** (SCG).

▶ Gradient Estimation Using Stochastic Computation Graphs (Schulman et al., 2015).
▶ Credit Assignment Techniques in Stochastic Computation Graphs (Weber et al., 2019)

Briefly mention today, more next week.
Stochastic computation graphs

A SCG is a directed, acyclic graph with nodes $\mathcal{V}$ has two types of nodes

- **Stochastic nodes** $\mathcal{S} \subseteq \mathcal{V}$, which are conditionally independent r.v.s given their parents.

- **Deterministic nodes** $\mathcal{D} \subseteq \mathcal{V}$, which are deterministic functions of their parents.
Stochastic computation graphs

Deterministic nodes are further specialized

- **Inputs** are deterministic nodes that have no parents. Includes the parameters $\theta$.

- **Losses** $\mathcal{L} \subseteq \mathcal{V}$ are the deterministic nodes whose average expectation we aim to minimize in $\theta$. 
We say that $w$ descends from $v$, $v \prec w$, if a path from $w$ to $v$ exists.

Can request the value of node $w$.

- Resolve the value of its ancestors $A_w = \{v : v \prec w\}$.
- In particular, all inputs in $A$ need to have their values given by a user or fixed.

Value of a stochastic node is a realization of the random variable.
For $L := \sum_{\ell \in L} \ell$ are interested in:

$$\nabla_\theta J(\theta) = \mathbb{E} [L] = \mathbb{E} \left[ \sum_{\ell \in L} \ell \right]$$

The partial derivative of stochastic nodes w.r.t. their parents is 0 by convention.

Then we have

$$\nabla_\theta J(\theta) = \mathbb{E} \left[ \sum_{v \in S} L \frac{d \log p(v)}{d \theta} + \sum_{\ell \in L} \frac{d \ell}{d \theta} \right]$$
Talks today

- Can we use SCG to compute higher order derivatives?
- Can we derive a policy gradient when our data is not generated with $d_{\gamma}^{\pi_\theta}(s|s_0)$?
- Can we compute gradients when we do not have the density of the random variables?