Exploring Policy Gradients

Chintak Sheth

Slides adapted from John Schulman's slides from MLSS '16

Notation

- Markov Decision Process is defined as $(\mathcal{S}, \mathcal{A}, \mathcal{P})$ where
  - $\mathcal{S}$: state space
  - $\mathcal{A}$: action space
  - $P(r, s' \mid s,\ a)$: transition probability distribution

- Some more definition for describing the problem
  - μ: Initial state distribution
  - γ: discount factor – rewards - how far into the future do you care about?

Episodic Setting

- In each episode, the initial state is sampled from μ, and the process proceeds until the terminal state is reached.
- Goal: maximize the expected reward per episode
- Type of Policies:
  - Deterministic policies: $a = \pi(s)$
  - Stochastic policies: $a \sim \pi(a \mid s)$
  - Parametric policies: π_θ
- Objective:

$$\text{maximize}\, \eta(\pi), \text{ where}$$
$$\eta(\pi) = E[r_0 + r_1 + \cdots + r_{T-1} \mid \pi]$$

$$\begin{aligned}
s_0 &\sim \mu(s_0) \\
a_0 &\sim \pi(a_0 \mid s_0) \\
s_1, r_0 &\sim P(s_1, r_0 \mid s_0, a_0) \\
a_1 &\sim \pi(a_1 \mid s_1) \\
s_2, r_1 &\sim P(s_2, r_1 \mid s_1, a_1) \\
&\cdots \\
a_{T-1} &\sim \pi(a_{T-1} \mid s_{T-1}) \\
s_T, r_{T-1} &\sim P(s_T \mid s_{T-1}, a_{T-1})
\end{aligned}$$

Policy Gradient Methods

- Problem:

$$\text{maximize } E[R \mid \pi_\theta]$$

- Main idea: collect a bunch of trajectories, push the gradient of the policy model in the direction which encourages high rewarding trajectories and push them away such that they discourage trajectories with low reward.

Score Function Gradient Estimator

- Consider an expectation $E_{x \sim p(x \mid \theta)}[f(x)]$
- We are interested in computing gradient wrt θ

$$
\begin{aligned}
\nabla_\theta E_x[f(x)] &= \nabla_\theta \int \mathrm{d}x \; p(x \mid \theta) f(x) \\
&= \int \mathrm{d}x \; \nabla_\theta p(x \mid \theta) f(x) \\
&= \int \mathrm{d}x \; p(x \mid \theta) \frac{\nabla_\theta p(x \mid \theta)}{p(x \mid \theta)} f(x) \\
&= \int \mathrm{d}x \; p(x \mid \theta) \nabla_\theta \log p(x \mid \theta) f(x) \\
&= E_x[f(x) \nabla_\theta \log p(x \mid \theta)].
\end{aligned}
$$

Score Function Gradient Estimator

$$E_x[f(x)\nabla_\theta \log p(x \mid \theta)]$$

- This gives us an unbiased gradient estimator. Just sample $x_i \sim p(x \mid \theta)$, and compute $\hat{g}_i = f(x_i)\nabla_\theta \log p(x_i \mid \theta)$

- The only requirement is that we are able to compute and differentiate density $p(x \mid \theta)$

Score Gradient Estimator for Policies

- In this case, RV x is the entire trajectory for an episode

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$$

$$\nabla_\theta E_\tau[R(\tau)] = E_\tau[\nabla_\theta \log p(\tau \mid \theta) R(\tau)]$$

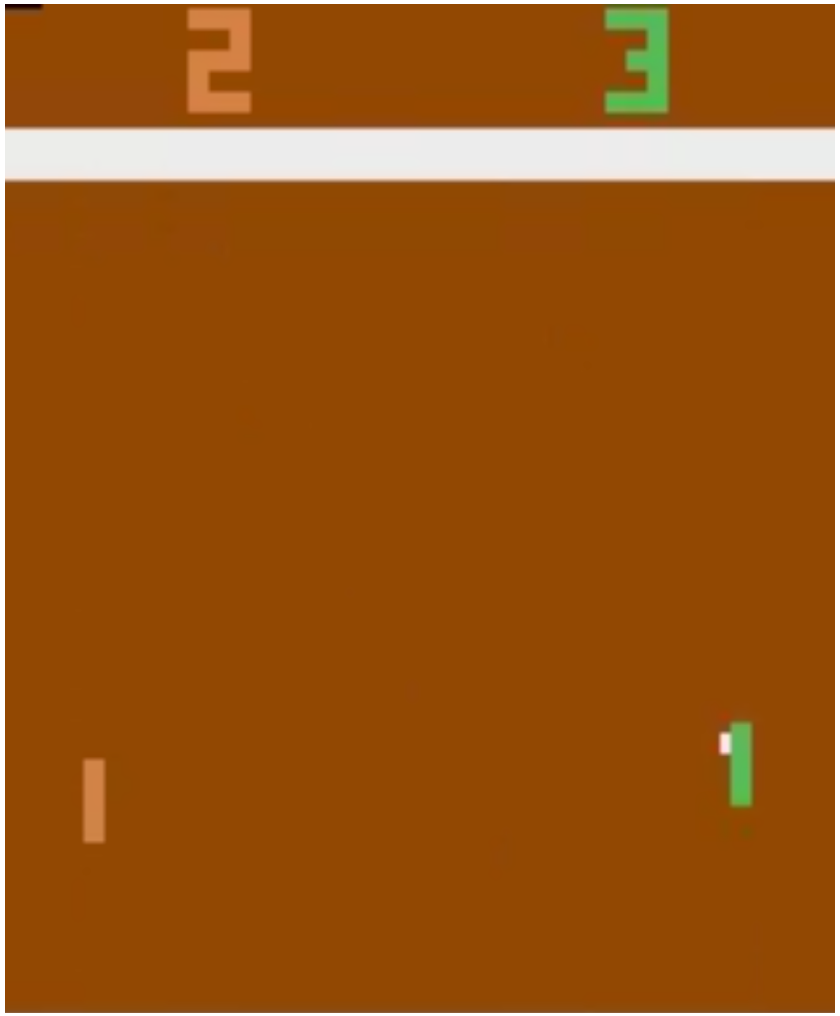
$$p(\tau \mid \theta) = \mu(s_0) \prod_{t=0}^{T-1} [\pi(a_t \mid s_t, \theta) P(s_{t+1}, r_t \mid s_t, a_t)]$$

$$\log p(\tau \mid \theta) = \log \mu(s_0) + \sum_{t=0}^{T-1} [\log \pi(a_t \mid s_t, \theta) + \log P(s_{t+1}, r_t \mid s_t, a_t)]$$

$$\nabla_\theta \log p(\tau \mid \theta) = \nabla_\theta \sum_{t=0}^{T-1} \log \pi(a_t \mid s_t, \theta)$$

$$\nabla_\theta \mathbb{E}_\tau [R] = \mathbb{E}_\tau \left[R \nabla_\theta \sum_{t=0}^{T-1} \log \pi(a_t \mid s_t, \theta) \right]$$

Performance on Atari's Pong Environment



check out the video