STA 4273: Minimizing Expectations

Lecture 7 - Policy Optimization I

Chris J. Maddison

University of Toronto

Announcements

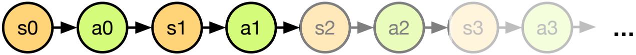
- A few proposal feedbacks left to send out. Great job!
- Questions, comments, concerns?

Today

- Switching gears to reinforcement learning.
- We will discuss the basic structure of so-called policy optimization algorithms.
- There are (roughly speaking) two perspectives on reinforcement learning.
  - Value-based methods, like Q-learning, which we discussed in Lec. 2.
  - Policy-gradient methods, which we will discuss today.

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 $\pi(a_t|s_t)$ for $T = \infty$ steps.



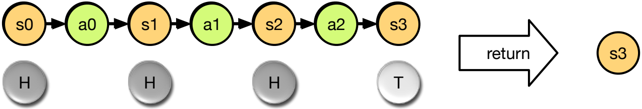
- The agent's objectives is to maximize its return:

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

- Today we are assuming $r(s, a) \geq 0$.

Recall: discounted state visitation distribution

- 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 is the marginal:

$$d^\pi(s) := \sum_{k=0}^{\infty} \gamma^k (1-\gamma) \sum_{\substack{a_{0:k-1} \\ s_{0:k-1}}} p(s_0)\pi_\theta(a_0|s_0)..p(s|s_{k-1}, a_{k-1})$$

- Also define the joint:

$$d^\pi(s, a) = d^\pi(s)\pi(a|s)$$

Recall: the policy gradient theorem

- Define:

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

$$V^\pi(s) = \sum_a \pi(a|s) Q^\pi(s,a)$$

- The policy gradient theorem tells us

$$(1-\gamma)\nabla_\pi J(\pi) = \mathbb{E}_{s,a\sim d^\pi}\left[Q^\pi(s,a)\nabla_\pi \log \pi(a|s)\right]$$

- It's also not too hard to derive:

$$(1-\gamma)J(\pi) = \mathbb{E}_{s,a\sim d^\pi}\left[r(s,a)\right]$$

- So we can think of this as single-step MDP with a certain environment.

An operator view (Ghosh et al., 2020)

- Policy-gradient(PG) methods use $\nabla_\pi J(\pi)$ (or an estimator of it) to find better policies.
- Suppose our policies are in some parametric family with parameters θ. We could always do gradient descent (or SGD),

$$\theta_{t+1} = \theta_t + \epsilon \mathbb{E}_{s,a \sim d^\pi} \left[Q^\pi(s,a) \nabla_{\theta_t} \log \pi(a|s) \right] \qquad \text{GD}$$
$$\theta_{t+1} = \theta_t + \epsilon Q^\pi(s,a) \nabla_{\theta_t} \log \pi(a|s) \text{ where } s, a \sim d^\pi \qquad \text{SGD},$$

  and understand its convergence via smoothness of J and stochastic approximation theory. Can we do better?
- There are many policy-gradient methods, but a unified view is not yet fully realized. Today we will discuss an operator view on PG methods.

An operator view (Ghosh et al., 2020)

- Value-based methods $\leftrightarrow$ apply operators on the value function.
  - Bellman optimality operator.
  - Bellman policy operator.
  - Greedy policy improvement operator.
- Policy-gradient methods $\leftrightarrow$ apply operators on the policy.
  - Dibya Ghosh, Marlos C. Machado, Nicolas Le Roux. An operator view of gradient methods. NeurIPS 2020.
- Ghosh et al. (2020) have the trajectory and state-action formulation. We will emphasize state-action.
- Focus on noiseless case (i.e. no stochasticity).

An operator view (Ghosh et al., 2020)

An update to the policy is the composition $\mathcal{P}_V \circ \mathcal{I}_V$ of operators

- Informally, a joint policy is a joint distribution $\mu(s, a)$ over states and actions that achieves return $\mathbb{E}_{s,a\sim\mu}[r(s, a)]$. A policy is realizable if $\mu(s, a) = d^{\pi'}(s)\pi'(a|s)$ for some agent's policy π'.
- Improvement operator. Maps a joint policy to another joint policy (sometimes) improves the return.

$$\mu(s, a) = \mathcal{I}_V d^{\pi}(s, a)$$

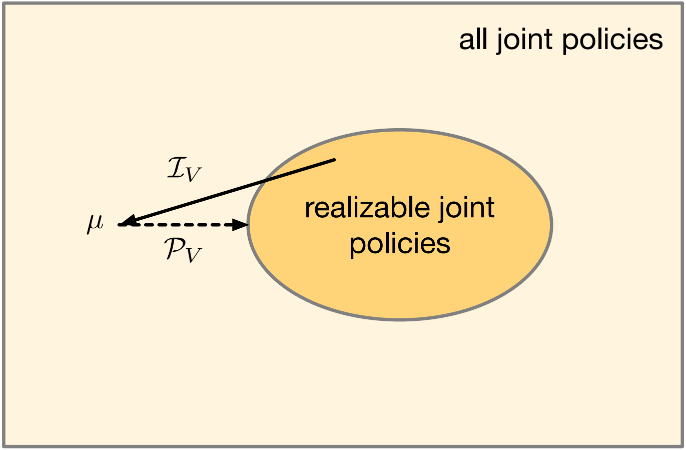
- Projection operator. Maps a distribution over states and actions into a realizable policy by minimizing some divergence:

$$z(a|s) = \mathcal{P}_V \mu(s, a) = \arg\min_{z\in\Pi} D_{\mu}(\mu||z)$$

and using $d^z(s)z(a|s)$ as the joint. Often, a gradient step is taken instead of a full minimization.

An operator view (Ghosh et al., 2020)

An update to the policy is the composition $\mathcal{P}_V \circ \mathcal{I}_V$ of operators



Let's see how this works for standard policy gradient.

Policy gradient improvement operator

Policy gradient improvement operator is:

$$\mu(s, a) = \mathcal{I}_V d^\pi(s, a) = \frac{d^\pi(s, a) Q^\pi(s, a)}{\mathbb{E}_{s,a\sim d^\pi}[Q^\pi(s, a)]}$$

So, reweight state-action pairs by the Q^π function. The new reward is,

$$\begin{aligned}
\mathbb{E}_{s,a\sim\mu}[r(s, a)] &= \frac{\mathbb{E}_{s,a\sim d^\pi}[Q^\pi(s, a) r(s, a)]}{\mathbb{E}_{s,a\sim d^\pi}[Q^\pi(s, a)]} \\
&= J(\pi) \frac{\mathbb{E}_{s,a\sim d^\pi}[Q^\pi(s, a) r(s, a)]}{\mathbb{E}_{s,a\sim d^\pi}[Q^\pi(s, a)] J(\pi)} \\
&= J(\pi) \left(1 + \frac{\mathrm{Cov}_{s,a\sim d^\pi}(Q^\pi(s, a), r(s, a))}{\mathbb{E}_{s,a\sim d^\pi}[Q^\pi(s, a)] J(\pi)} \right)
\end{aligned}$$

If $\mathrm{Cov}_{s,a\sim d^\pi}(Q^\pi(s, a), r(s, a)) \geq 0$, this is an improvement.

Policy gradient projection operator

Policy gradient projection operator is computed using:

$$z(a|s) = \mathcal{P}_V \mu(s,a) = \arg\min_{z \in \Pi} \sum_s \mu(s) KL(\mu(a|s) || z(a|s))$$

Using $\mu = \mathcal{I}_V d^\pi$, this resolves to

$$\begin{aligned} z(a|s) &= (\mathcal{P}_V \circ \mathcal{I}_V)(d^\pi) \\ &= \arg\min_{z \in \Pi} -\sum_{s,a} d^\pi(s,a) Q^\pi(s,a) \log z(a|s) \end{aligned}$$

where we dropped a bunch of constants in z (OK, since it's an argmin).

Policy gradient

- An optimal policy π^* is a fixed point of $\mathcal{P}_V \circ \mathcal{I}_V$ (Ghosh et al., 2020).
- But, how does $\mathcal{P}_V \circ \mathcal{I}_V$ relate to a policy gradient step?

Policy gradient step

- Suppose $\Pi = \{\pi_\theta(a|s) : \theta \in \mathbb{R}^n\}$ is a set of parametric policies.
- In this case,

$$(\mathcal{P}_V \circ \mathcal{I}_V)(d^{\pi_{\theta_t}}) = \arg\min_{\theta \in \mathbb{R}^n} -\sum_{s,a} d^{\pi_{\theta_t}}(s,a) Q^{\pi_{\theta_t}}(s,a) \log \pi_\theta(a|s)$$

- Instead of a full minimization, what if we took one step of gradient descent? $\mathcal{P}_V \circ \mathcal{I}_V(d^{\pi_{\theta_t}}) \approx \pi_{\theta_{t+1}}$ where

$$\theta_{t+1} = \theta_t + \epsilon \sum_{s,a} d^{\pi_{\theta_t}}(s,a) Q^{\pi_{\theta_t}}(s,a) \nabla_\theta \log \pi_\theta(a|s)$$

- This is the standard policy gradient step.

Policy gradient step

Why does the policy gradient step work?

- Policy gradient step:

$$(\mathcal{P}_V \circ \mathcal{I}_V)(d^{\pi_{\theta_t}}) \approx \arg\min_{\theta \in \mathbb{R}^n} - \sum_{s,a} d^{\pi_{\theta_t}}(s,a) Q^{\pi_{\theta_t}}(s,a) \log \pi_\theta(a|s)$$

- Ghosh et al. (2020) show that for any two $\pi(a|s)$ and $\mu(a|s)$

$$J(\pi) \geq J(\mu) + \sum_{s,a} d^{\mu}(s,a) Q^{\mu}(s,a) \log \frac{\pi(a|s)}{\mu(a|s)}$$

- The standard policy gradient iteratively maximizes a local approximation around π_{θ_t}, which is a global lower bound.
- Didn't prove conditions under which this converges, but should be pretty mild.