STA 4273: Minimizing Expectations

Lecture 9 - Policy Optimization II

Chris J. Maddison

University of Toronto

Announcements

- Questions, comments, concerns?

Today

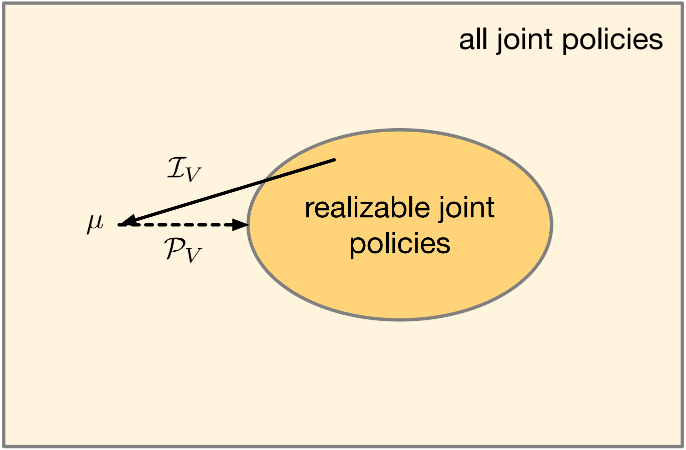
- Last week we heard from George Tucker about offline reinforcement learning.
- Today I'll wrap up some of the introductory ideas from policy optimization (still mostly online).
- Student presentations will continue on the theme of offline policy optimization.

Recall: 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 is the composition $\mathcal{P}_V \circ \mathcal{I}_V$ of operators



Policy improvement $\mathcal{I}_V$ and projection $\mathcal{P}_V$.

Standard policy gradient

- The composition of the operators is

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

- 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.

Other returns, other improvement operators

- Using improvement operators based on non-linear transformations of the return might lead to speed ups.
- Intuitively, policies at the beginning of training are so bad, that over-emphasizing high reward trajectories might lead to faster learning.

Polynomial returns

- Ghosh et al. (2020) study the case of polynomial returns, i.e., the improvement operator for $\alpha > 0$:

$$\mu(s, a) = \mathcal{I}_V^\alpha d^\pi(s, a) \propto d^\pi(s, a)(Q^\pi(s, a))^{\frac{1}{\alpha}}$$

  In this case, we have

$$\mu(a|s) \propto \pi(a|s)(Q^\pi(s, a))^{\frac{1}{\alpha}}$$

- They show that an optimal policy π^* is the fixed point of $\mathcal{P}_V^\alpha \circ \mathcal{I}_V^\alpha$, where $\mathcal{P}_V^\alpha$ is a projection operator based on the α-divergence.

Polynomial returns

- The polynomial return improved policy is

$$\mu(a|s) \propto \pi(a|s)(Q^\pi(s,a))^{\frac{1}{\alpha}}$$

- $\alpha = 1$ is standard policy gradient improvement operator that we just discussed.
- As $\alpha \to 0$, we get

$$\mu(a|s) = \begin{cases} 1 & \text{if } a \in \arg\max_a Q^\pi(s,a) \\ 0 & \text{o.w.} \end{cases}$$

  i.e., the greedy policy improvement operator!
- This shows that policy gradient and policy iteration are on a continuum, with the main difference being how aggressively they use the value function to determine the next policy.

Other state of the art methods

- This framework can be used to describe state-of-the-art methods, like PPO and Maximum a Posterior Policy Optimization (MPO, Abdomaleki et al., 2018).
- MPO, in particular, is easy to describe and motivated by similar intuitions as the polynomial returns.

MPO (Abdomaleki et al., 2018)

- MPO's improvement operator is

$$\mathcal{I}_V^{MPO} d^\pi(s,a) = \arg\max_{\mu(a|s)} \sum_s d^\pi(s) \left[\beta \mathbb{E}_{a\sim\mu(a|s)}\left[Q^\pi(s,a)\right] - KL(\mu||\pi)\right]$$
$$= d^\pi(s) \frac{\pi(a|s)\exp(\beta Q^\pi(s,a))}{Z_\beta^\pi(s)}$$

  - $Z_\beta^\pi(s) = \sum_{a'} \pi(a'|s)\exp(\beta Q^\pi(s,a'))$

- MPO improvement operator optimizes an ELBO in terms of μ, i.e., it finds a posterior.

MPO (Abdomaleki et al., 2018)

- In general, we use parameteric policies, and $\mu(a|s) \propto \pi(a|s) \exp(\beta Q^{\pi}(s,a))$ may not be realizable.
- The MPO projection operator is basically the same as the standard policy gradient, i.e., it projects μ back to the set of parameteric policies:

$$\mathcal{P}_V^{MPO} \mu(s,a) = \arg\max_{\theta \in \mathbb{R}^n} \sum_{s,a} \mu(s,a) \log \pi_\theta(a|s) p(\theta)$$

  - $p(\theta)$ is a user-specified, parameter prior.
- Projection operator solves a maximum-likelihood problem.

MPO (Abdomaleki et al., 2018)

- MPO is directly analogous to the EM algorithm in statistics.
  - Improvement operator $\rightarrow$ E-step.
  - Projection operator $\rightarrow$ M-step.
- Why is it sensible?
  - Similar intuition as polynomial returns, β can balance the preference for high-return actions with the need to explore.
  - Is it the optimal balance? As we'll hear from Brendan O'Donoghue, methods in this space are closely related to efficient exploration (although not necessarily for MPO).

Other perspectives, other methods

- MPO is not the only algorithm of this type, there's a whole zoo.
  - REPS (Peters et al., 2010)
  - TRPO (Schulman et al., 2015)
  - PPO (Schulman et al., 2017)
- This operator view is not the only view.
  - Mirror descent (Neu at al., 2017)

Today's talks

Hear more about offline policy optimization.

- MOPO (model based offline policy optimization)
- Policy distillation
- Unified view of RL via Fenchel-Rockafellar Duality