IEOR 8100: Reinforcement learning

Lecture 6: Actor-critic methods

The methods we have seen so far can be divided into two categories [Konda and Tsitsiklis, 2003]

- Actor-only methods (vanilla policy gradient) work with a parameterized family of policies. The gradient of the performance, with respect to the actor parameters, is directly estimated by simulation, and the parameters are updated in a direction of improvement. A possible drawback of such methods is that the gradient estimators may have a large variance. Furthermore, as the policy changes, a new gradient is estimated independently of past estimates (by sampling trajectories). Hence, there is no "learning", in the sense of accumulation and consolidation of older information.
- Critic-only methods (e.g., Q-learning, TD-learning) rely exclusively on value function approximation and aim at learning an approximate solution to the Bellman equation, which will then hopefully prescribe a near-optimal policy. Such methods are indirect in the sense that they do not try to optimize directly over a policy space. A method of this type may succeed in constructing a "good" approximation of the value function, yet lack reliable guarantees in terms of near-optimality of the resulting policy.

Actor-critic methods aim at combining the strong points of actor-only and critic-only methods, by incorporating value function approximation in the policy gradient methods. We already saw the potential of using value function approximation for picking baseline for variance reduction. Another more obvious place to incorporate Q-value approximation is for approximating Q-function in the policy gradient expression (refer to Policy gradient theorem in the last lecture). Recall, by policy gradient theorem:

$$\nabla_\theta \rho(\pi_\theta) = \sum_s d^{\pi_\theta}(s) \mathbb{E}_{a\sim\pi(s)} \left[(Q^{\pi_\theta}(s,a) - b^{\pi_\theta}(s)) \nabla_\theta \log(\pi_\theta(s,a)) \right]$$

for any baseline $b^{\pi_\theta}(\cdot)$.

In the vanilla policy gradient algorithm, we essentially approximated the Q-value function by Monte-Carlo estimation. In every iteration, we sampled multiple independent trajectory to implicitly do a Q-value estimation. Thus, as the policy changes, a new gradient is estimated independently of past estimates. In the actor-critic method, we use Q-function approximation in the gradient estimate. The actor-critic algorithm simultaneously/alternatively updates the Q-function approximation as well as policy parameters, as it sees more samples. Over iterations, as policy changes, the Q-function approximations also improves with more sample observations.

Before describing the algorithm, let's first understand the requirements from such an approximation – what kind of Q-function approximations are desirable (at least in ideal conditions)?

1 Policy gradient theorem with Q-function approximation.

For a scalable estimation of Q-value function, one may want to use function approximation for Q-value approximation as well. Let $f_\omega(s,a)$ be approximation of Q^{π_θ}. The next theorem shows that certain types of Q-function approximation are more 'compatible' with policy gradient approach.

Theorem 1 (Sutton et al. [1999]). *If function f_ω is compatible with policy parametrization θ in the sense that for every s, a,*

$$\nabla_\omega f_\omega(s,a) = \frac{1}{\pi_\theta(s,a)} \nabla_\theta \pi_\theta(s,a) = \nabla_\theta \log(\pi_\theta(s,a))$$

And, further we are given parameter ω which is a stationary point of the following least squares problem:

$$\min_\omega \mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(s,\cdot)} \left[(Q^{\pi_\theta}(s,a) - b(s;\theta) - f_\omega(s,a))^2 \right]$$

where $b(\cdot;\theta)$ any baseline, which may depend on the current policy π_θ. Then,

$$\nabla_\theta \rho(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_\theta}} \mathbb{E}_{a\in\pi_\theta(s)} \left[f_\omega(s,a) \nabla_\theta \log(\pi_\theta(s,a)) \right]$$

That is, function approximation f_ω can be used in place of Q-function to obtain gradient with respect to θ.

(Here, we abuse the notation and use $\mathbb{E}_{s\sim d^{\pi_\theta}}[x]$ as a shorthand for $\sum_s d^{\pi_\theta}(s)x$. This is not technically correct in the discounted case since in that case $d^{\pi_\theta}(s) = \mathbb{E}_{s_1}[\sum_{t=1}^{\infty} \Pr(s_t = s; \pi, s_1)\gamma^{t-1}]$, which is not a distribution. In fact in discounted case, $(1-\gamma)d^{\pi_\theta}$ is a distribution.

Proof. Given θ, for stationary point ω of the least squares problem:

$$\mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(s,\cdot)}[(Q^{\pi_\theta}(s,a) - b(s;\theta) - f_\omega(s,a))\nabla_\omega f_\omega(s,a)] = 0$$

Substituting the compatibility condition:

$$\mathbb{E}_{s\sim d^{\pi_\theta}, a\sim\pi_\theta(s,\cdot)}[(Q^{\pi_\theta}(s,a) - b(s;\theta) - f_\omega(s,a))\nabla_\theta \pi_\theta(s,a)\frac{1}{\pi_\theta(s,a)}] = 0$$

Or,

$$\sum_s d^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(s,a)(Q^{\pi_\theta}(s,a) - b(s;\theta) - f_\omega(s,a)) = 0$$

Since $b(s;\theta)\sum_a \nabla_\theta \pi_\theta(s,a) = 0$,

$$\sum_s d^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(s,a)(Q^{\pi_\theta}(s,a) - f_\omega(s,a)) = 0$$

using this with the policy gradient theorem, we get

$$\nabla_\theta \rho(\pi_\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(s,a) f_\omega(s,a)$$

□

1.1 Example: softmax policy

Consider policy set parameterized by θ such that given $s \in S$, probability of picking action $a \in A$ is given by:

$$\pi_\theta(s,a) = \frac{e^{\theta^\top \phi_{sa}}}{\sum_{a'\in A} e^{\theta^\top \phi_{sa'}}}$$

where each ϕ_{sa} is an ℓ-dimensional feature vector characterizing state-action pair s, a. This is a popular form of parameterization. Here,

$$\nabla_\theta \pi_\theta(s,a) = \phi_{sa}\pi_\theta(s,a) - \left(\sum_{a'\in A} \phi_{sa'}\pi_\theta(s,a') \right) \pi_\theta(s,a)$$

Meeting the compatibility condition in Theorem 1 requires that

$$\nabla_\omega f_\omega(s,a) = \frac{1}{\pi_\theta(s,a)} \nabla_\theta \pi_\theta(s,a) = \phi_{sa} - \sum_{a'\in A} \phi_{sa'}\pi_\theta(s,a')$$

A natural form of $f_\omega(s,a)$ satisfying this condition is:

$$f_\omega(s,a) = \omega^\top (\phi_{sa} - \sum_{b\in A} \phi_{sb}\ \pi_\theta(s,b))$$

Thus f_w must be linear in the same features as the policy, except normalized to be mean zero for each state. In this sense it is better to think of f_w as an approximation of the advantage function, $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$, rather than Q^π.

1.2 Example: Gaussian policy for continuous action spaces

In continuous action spaces, it is natural to use Gaussian policy. Given state s, the probability of action a is given as:

$$\pi_\theta(s,a) = \mathcal{N}(\phi(s)^T\theta, \sigma^2)$$

for some constant σ. Here $\phi(s)$ is a feature representation of s. Then, compatibility condition for $f_\omega(s,a)$:

$$\nabla_\omega f_\omega(s,a) = \nabla_\theta \log(\pi_\theta(s,a)) = \nabla_\theta \frac{-(a-\theta^\top\phi(s))^2}{2\sigma^2} = \frac{(\theta^\top\phi(s)-a)}{\sigma^2}\phi(s)$$

For f_ω to satisfy this, it must be linear in ω, e.g., $f_\omega(s,a) = \frac{(\theta^\top\phi(s)-a)}{\sigma^2}\phi(s)^\top\omega$

2 Policy iteration algorithm with function approximation [Sutton et al., 1999]

Let $f_\omega(\cdot,\cdot)$ be such that $\nabla f_\omega(s,a) = \nabla_\theta \log \pi_\theta(s,a)$ for all ω, θ, s, a. Initialize $\theta_1, \pi_1 := \pi_{\theta_1}$. Pick step sizes $\alpha_1, \alpha_2, \ldots,$.

In iteration $k = 1, 2, 3, \ldots,$

- **Policy evaluation:** Find $w_k = w$ such that

  $$\mathbb{E}_{s\sim d^{\pi_k}}\mathbb{E}_{a\sim\pi_k(s)}\left[(Q^{\pi_k}(s,a) - b_k(s) - f_\omega(s,a))\nabla_\theta \log(\pi_k(s,a))\right] = 0$$

  (Here, d^{π_k} is not normalized to 1, and sums to $1/(1-\gamma)$.)

- **Policy improvement:**

  $$\theta_{k+1} \leftarrow \theta_k + \alpha_k\ \mathbb{E}_{s\sim d^{\pi_k}}\mathbb{E}_{a\sim\pi_k(s)}\left[f_\omega(s,a)\nabla_\theta \log(\pi_k(s,a))\right]$$

A similar algorithm appears in Konda and Tsitsiklis [1999].

2.1 Convergence Guarantees

Following version of convergence guarantees were provided by Sutton et al. [1999] for infinite horizon MDPs (average or discounted).

Theorem 2 (Sutton et al. [1999]). *Given $\alpha_1, \alpha_2, \ldots,$ such that*

$$\lim_{T\to\infty}\sum_{k=1}^{T}\alpha_k = \infty, \quad \lim_{T\to\infty}\sum_{k=1}^{T}\alpha_k^2 < \infty,$$

and $\max_{\theta,s,a,i,j}\frac{\partial^2\pi_\theta(s,a)}{\partial\theta_i\theta_j} < \infty$. *Then, for $\theta_1, \theta_2, \ldots,$ obtained by the above algorithm,*

$$\lim_{k\to\infty}\nabla_\theta\rho(\theta)|_{\theta_k} = 0$$

The proof of the above theorem can be obtained by viewing the algorithm as a stochastic approximation method with $h(\theta) = \nabla_\theta\rho(\pi_\theta)$. Then, under the assumptions stated in the theorem, the convergence to $h(\theta) = 0$ can be obtained using Proposition 3.5 of Bertsekas and Tsitsiklis [1996].

Note that above theorem needs several desirable conditions in order to achieve convergence. Notably (a) second derivative of the policy function approximation is small, (b) in every iteration, good Q-function approximation parameters ω are found (specifically, ω is a stationary point of the squared loss function $\min_\omega(f_\omega(s,a) - Q^{\pi_\theta}(s,a))^2$), and (c) policy gradient is estimated accurately. In practice, Q-function approximation would only be approximately

satisfy the condition. Further, the policy gradient estimation through sampling will be approximate. In the next lecture, we study the work by Kakade and Langford [2002], which highlight the instability of this algorithm under such approximation, through examples where this algorithm would require exponential number of iterations. They also propose approximate algorithms that are guaranteed to terminate in a small number of steps.

The work by Kakade and Langford [2002] also form the basis of the recent 'Trust Region Policy Optimization' (TRPO) algorithm Schulman et al. [2015a]. Other recent actor-critic algorithms based on advantage estimation include 'Asynchronous Advantage Actor Critic Algorithm (A3C)' [Mnih et al., 2016] and 'Generalized Advantage Estimation (GAE)' [Schulman et al., 2015b].

References

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pages 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1-55860-873-7. URL `http://dl.acm.org/citation.cfm?id=645531.656005`.

Vijay R. Konda and John N. Tsitsiklis. On actor-critic algorithms. *SIAM J. Control Optim.*, 42(4):1143–1166, April 2003. ISSN 0363-0129. doi: 10.1137/S0363012901385691. URL `https://doi.org/10.1137/S0363012901385691`.

Vijaymohan R. Konda and John N. Tsitsiklis. Actor-critic algorithms. *Proceedings of the 12th International Conference on Neural Information Processing Systems*, 1999.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016. URL `http://arxiv.org/abs/1602.01783`.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1889–1897. JMLR.org, 2015a. URL `http://dl.acm.org/citation.cfm?id=3045118.3045319`.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *CoRR*, abs/1506.02438, 2015b. URL `http://arxiv.org/abs/1506.02438`.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 1057–1063, 1999.