From Policy Gradient to Actor-Critic methods

Bias variance trade-off

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud
Bias versus variance

- PG methods estimate an expectation from a finite state of trajectories
- If you estimate an expectation over a finite set of samples, you get a different number each time
- This is known as variance
- Given a large variance, you need many samples to get an accurate estimate of the mean
- That's the issue with MC methods
- If you update an expectation estimate based on a previous (wrong) expectation estimate, the estimate you get even from infinitely many samples is wrong
- This is known as bias
- This is what bootstrap methods do

Bias variance trade-off

- More complex model (e.g. bigger network): more variance, less bias
- Total error = $\text{bias}^2 + \text{variance} + \text{irreducible error}$
- There exists an optimum complexity to minimize total error
From Policy Gradient to Actor-Critic methods

Using the N-step return

1-step TD

3-step TD

- 1-step TD is poor at backpropagating values along trajectories
- N-step TD is better: N steps of backprop per trajectory instead of one
N-step return and replay buffer

- N-step TD can be implemented efficiently using a replay buffer
- A sample contains several steps
- Various implementations are possible

Generalized Advantage Estimation: \( \lambda \) return

- The N-step return can be reformulated using a continuous parameter \( \lambda \)
- \( \hat{A}_\phi^{(\gamma, \lambda)} = \sum_{l=0}^{H} (\gamma \lambda)^l \delta_{t+l} \)
- \( \hat{A}_\phi^{(\gamma, 0)} = \delta_t = \text{one-step return} \)
- \( \hat{A}_\phi^{(\gamma, 1)} = \sum_{l=0}^{H} (\gamma)^l \delta_{t+l} = \text{MC estimate} \)
- The \( \lambda \) return comes from eligibility trace methods
- Provides a continuous grip on the bias-variance trade-off


Bias-variance compromise

- **MC**: unbiased estimate of the critic
- **But MC suffers from variance due to exploration (+ stochastic trajectories)**
- **MC on-policy → no replay buffer → less sample efficient**
- **Bootstrap is sample efficient but suffers from bias and is unstable**
- **N-step TD or $\lambda$ return: control the bias-variance compromise**
- **Acts on critic, indirect effect on performance**
- **Next lesson: on-policy vs off-policy**
Any question?

Send mail to: Olivier.Sigaud@upmc.fr
Stuart Geman, Elie Bienenstock, and René Doursat.
Neural networks and the bias/variance dilemma.

Long-Jin Lin.
Self-Improving Reactive Agents based on Reinforcement Learning, Planning and Teaching.

Sahil Sharma, Srivatsan Ramesh, Balaraman Ravindran, et al.
Learning to mix n-step returns: Generalizing lambda-returns for deep reinforcement learning.