

# From Policy Gradient to Actor-Critic methods

## Bias variance trade-off

Olivier Sigaud

Sorbonne Université  
<http://people.isir.upmc.fr/sigaud>



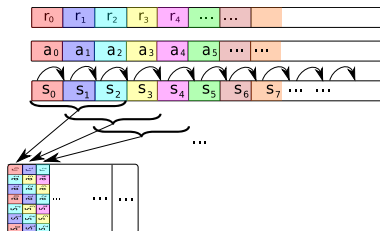
## Bias versus variance

- ▶ If you compute an expectation over infinitely many samples, you get the same expectation each time
- ▶ But if you compute it over a finite set of samples, you get a different expectation each time
- ▶ This is known as **variance**
- ▶ Given a large variance, you need many samples to get an accurate estimate of the mean
- ▶ If you update an expectation based on a previous (wrong) expectation estimate, the expectation estimate you get provided infinitely many samples is wrong
- ▶ This is known as **bias**



Geman, S., Bienenstock, E., & Doursat, R. (1992) Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1-58

## N-step return and replay buffer



- ▶ One-step TD is poor at backpropagating value from the end to the beginning of trajectories
- ▶ N-step TD is better: N steps of backprop per trajectory instead of one
- ▶ Can be implemented efficiently using a replay buffer
- ▶ Various implementations are possible



Lin, L.-J. (1992) Self-Improving Reactive Agents based on Reinforcement Learning, Planning and Teaching. *Machine Learning*, 8(3/4), 293–321

## Generalized Advantage Estimation: $\lambda$ return

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

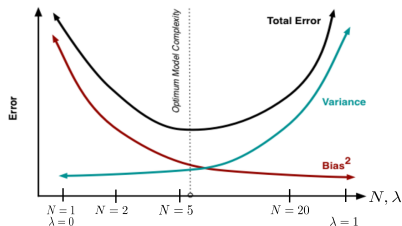
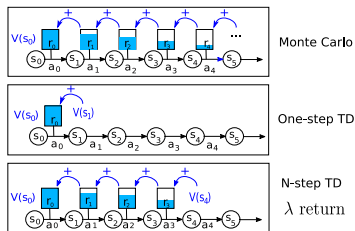


John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015



Sharma, S., Ramesh, S., Ravindran, B., et al. (2017) Learning to mix N-step returns: Generalizing  $\lambda$ -returns for deep reinforcement learning. *arXiv preprint arXiv:1705.07445*

## Bias-variance compromise



- ▶ MC: unbiased estimate of the critic
- ▶ But MC suffers from variance due to exploration (+ stochastic trajectories)
- ▶ MC on-policy  $\rightarrow$  no replay buffer  $\rightarrow$  less sample efficient
- ▶ Bootstrap is sample efficient but suffers from bias and is unstable
- ▶ Total error = bias<sup>2</sup> + variance + irreducible error
- ▶ More complex model (e.g. bigger network): more variance, less bias
- ▶ 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](mailto:Olivier.Sigaud@upmc.fr)



Stuart Geman, Elie Bienenstock, and René Doursat.

Neural networks and the bias/variance dilemma.

*Neural computation*, 4(1):1–58, 1992.



Long-Jin Lin.

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

*Machine Learning*, 8(3/4):293–321, 1992.



Sahil Sharma, Srivatsan Ramesh, Balaraman Ravindran, et al.

Learning to mix n-step returns: Generalizing lambda-returns for deep reinforcement learning.

*arXiv preprint arXiv:1705.07445*, 2017.