Disclaimer

- The slides below are excerpts from Prof. Rich Sutton's NIPS2015 tutorial "Introduction to Reinforcement Learning with Function Approximation", with page numbered 1,48,52.
- • For full slides, please check here: http://www.incompleteideas.net/Talks/RLtutorialNIPS2015.pdf

Introduction to Reinforcement Learning with Function Approximation

Rich Sutton

Reinforcement Learning and Artificial Intelligence Laboratory
Alberta Centre for Machine Learning
Department of Computing Science
University of Alberta
Canada

(with thanks to David Silver and Michael Littman for some slides and ideas)

The deadly triad

- The risk of divergence arises whenever we combine three things:

  1. Function approximation

     significantly generalizing from large numbers of examples

  2. Bootstrapping

     learning value estimates from other value estimates,
     as in dynamic programming and temporal-difference learning

  3. Off-policy learning (Why is dynamic programming off-policy?)

     learning about a policy from data not due to that policy,
     as in Q-learning, where we learn about the greedy policy from
     data with a necessarily more exploratory policy

- Any two without the third is ok

Other ways to survive the deadly triad

- Use high λ. Use good features
- Recent results suggest that replay and more stable targets (e.g., Double Q-learning, van Hasselt 2010) may help, but it is too soon to be sure
- Use least-squares methods like off-policy LSTD(λ) (Yu 2010, Mahmood et al. 2015). Such methods (Bradtke & Barto 1996, Boyan 2000) easily survive the triad, but their computational costs scale with the *square* of the number of parameters
- Try the new true-gradient RL methods (Gradient-TD and proximal-gradient-TD) developed by Maei (2011) and Mahadevan (2015) et al. These seem to me to be the best attempts to make TD methods with the robust convergence properties of stochastic gradient descent. Residual gradient methods (Baird 1999) are also true gradient methods, but optimize a poor objective, or can't learn purely from data (double sampling). These and other methods based on the Bellman error/residual are not recommended
- Try the even newer Emphatic-TD methods (Sutton, White & Mahmood 2015, Yu 2015). These semi-gradient methods attain stability through an extension of the early on-policy theorems