UMassAmherst

Regularized Off-Policy TD-Learning

Bo Liu, Sridhar Mahadevan, Ji Liu

Poster ID: W88

University of Massachusetts Amherst

AUTONOMOUS LEARNING LABORATORY

THE UNIVERSITY of WISCONSIN MADISON

Problem Setting

- **Off-Policy training** is training on data from one policy in order to learn the value of another policy
- **TD Learning** algorithm diverges in off-policy training.
- **TD with Gradient Correction (TDC)** algorithm is an off-policy convergent RL algorithm [Sutton et. al, 2009]
- Regularization helps improve stability of TD methods
- **RO-TD** algorithm: First ***Regularized Off-Policy convergent*** TD algorithm with ***Linear Computation Complexity***

Essence Of RO-TD Algorithm

Objective Function:

$\boldsymbol{l}_1$-regularized approximate solution of linear equation formulation of TDC

Convex-concave Formulation:

Saddle-point bilinear representation enables stochastic regularization

Linear Computation:

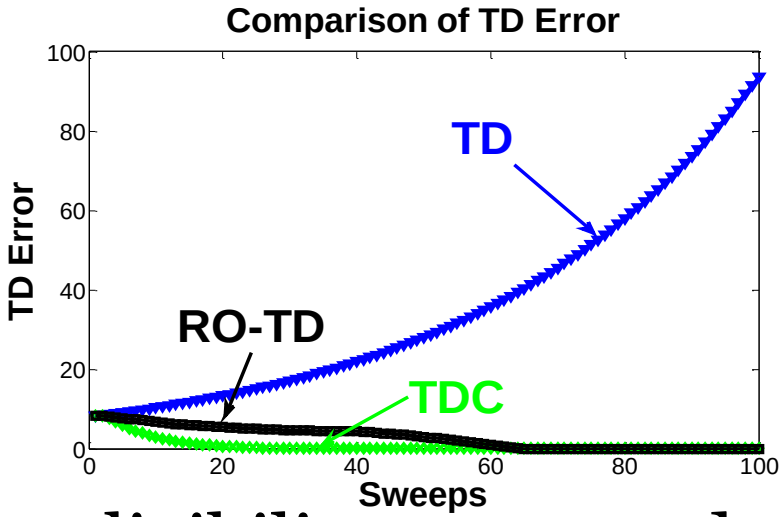
Linear complexity w.r.t sample and feature size *O(Nd)*

Control Learning Extension:

RO-GQ(λ)

Performance of RO-TD Algorithm

- Off-Policy Convergence



- Feature Selection
- Control Learning including eligibility traces and temporal abstraction prediction
- Visit our poster W88!