Abstract

- Regret analysis for two online RL algorithms
- Theoretical implications
  - TD makes more accurate predictions
  - RG makes more consistent predictions
- Suggests usefulness of online-learning techniques in RL analysis

Sequential Prediction in RL

- A.k.a. “policy evaluation”
- Critical in approximate policy iteration (e.g., Lspi (Lagoudakis & Parr, 03))
- Markovian assumption is common

Notation

- Trajectory: \( x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \cdots \rightarrow x_{t+1} \)
- Value (return) at time \( t \): \( y_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \)
- “Bellman equation”: \( y_t = r_t + \gamma y_{t+1} \)
- Prediction: \( \hat{y}_t = w_t \cdot x_t \)
- Prediction error: \( e_t = y_t - \hat{y}_t \)
- Temporal difference: \( d_t = r_t + \gamma x_t \cdot x_{t+1} - w_t \cdot x_t \)
- Total “losses”: \( \ell_p = \sum d_t^2 \) and \( \ell_p = \sum d_t^2 \)

We compare an algorithm’s loss to that of the best weight vector \( u \)

- \( \ell_p^* = \sum (y_t - u \cdot x_t)^2 \) and \( \ell_p^* = \sum (r_t + \gamma u \cdot x_{t+1} - u \cdot x_t)^2 \)
- Goal: upper bound \( \ell_p / \ell_p^* \) and \( \ell_p / \ell_p^* \)

Algorithm

- TD(0) (Sutton, 88)
  \[ w_{t+1} \leftarrow w_t + \eta d_t x_t \]
- TD*(0) (Schapire & Warmuth, 96)
  \[ w_{t+1} \leftarrow w_t + \frac{\eta}{1 - \gamma \eta x_t \cdot x_{t+1}} d_t x_t \]
- RG (Baird, 95)
  \[ w_{t+1} \leftarrow w_t + \eta d_t (x_t - y x_{t+1}) \]

Regret Analysis Summary

- TD(0)
  \[ \ell_p^* / \ell_p \]
  \[ TD^*(0) + o(1) \]
  \[ 1 + o(1) \]
- RG
  \[ \ell_p^* / \ell_p \]
  \[ 1 + o(1) \]

Proof Idea

- Use \( ||w_t - u||^2 \) as the potential function to measure progress of learning in each step
- Lower bound \( ||w_t - u||^2 \) in terms of \( \eta \), \( r_t \), \( d_t \), \( e_t \), and \( u \)
- Upper bound \( d_t^2 \) and \( e_t^2 \) in terms of \( \eta \), \( r_t \), and \( u \)
- Upper bound \( \sum d_t^2 \) and \( \sum e_t^2 \) by algebraic manipulations
- Optimize \( \eta \) to minimize these upper bounds

Related Work

- The model was proposed by Schapire & Warmuth (96)
- Prediction error of TD*(\( \lambda \)) was analyzed
- Related problem in online learning
  - (Cesa-Bianchi, Long, & Warmuth, 96)
- Effectively, \( \gamma = 0 \)
- TD(0) was observed to converge faster than RG empirically
  - (Baird, 95)
- TD(0) was provably faster when (Schoknecht & Merke, 03)
- Updates are synchronous
- No function approximation is used
- A certain matrix has real eigenvalues only
- In approximate policy iteration
- TD solutions seem more desirable (Lagoudakis & Parr, 03)
- RG solutions are more stable (Munos, 03)

Future Work

- Extend analysis to TD(\( \lambda \)) and TD*(\( \lambda \))
- Consider multiplicative updates
  - (Kivinen & Warmuth, 97) and (Precup & Sutton, 97)
- Find (matching) lower bounds