STUDY WEEK Report #2

20161106 Jerry

Done List

- Propose VCQL and VCSARSA
- Go through 2 Papers

Unrelated Stuff:

1. Learn how to use Haskell : A functional programming language To Chapter 12
2. Revise for Physics Mid-Term
3. restore from sick

Paper Note

Mean-Variance Optimization in Markov Decision Processes

2011 posted 2013 accepted , John N. Tsitsiklis , MIT , Computer Science Field

Content

1. Author define a computer science problem for computing a fully observable mdp whose $Variance < C_1$ and $Q(s,a) > C_2$.
2. Author prove that, By **Constructing Method**, the specific subproblem can be reducted to **Subset Sum Problem** , Which indicates the subproblem is **NP-Hard**.
3. Author gives a **useless** approximate algorithm which make reductions of this problem into Interger Linear Programming.

Thoughts

1. We can also show one specific subproblem is actually P-Hard, Thus **whether the total problem is NP-Hard is under doubt**.
2. This point hasn't be applied into Reinforcement Learning yet.

3. The technique of appoximation to converse the limit equations into a **polyhedron** and then apply **Interger Linear Programming**. However this method is totally useless for any problem. Not only because it requires every information like probability transition, but also the complexity is actually exponential for real problem due to its unrealistic constraints in this paper.

Quantile Reinforcement Learning

2016 Hugo Gilber, Sorbonnes Universit´es; Paul Weng , CMU , Machine Learning Field

Content

1. The author propose a new criterior for MDP that based on **order** of result.
2. They design a new problem which maimize the worst case with hisgh probalbility getting an better endstate. 3. They only consider the speical MDP, only with last-step reward.
3. They transverse the solution to a constucted MDP with Q-Learning given the worst case.
4. They design a new algorithm , simultaneously optimize the worst case and the best policy under that worst case based on paper 1.

Thoughts

1. Their proposed algorithm fully relies on the **last-step reward**, which actually makes no contributions to the real problem. Because for non-order reward this technique also works and is trivial.
2. The 2 Timescale technique is useful. Supposing we have problem $max(u(v(x)))\forall u, v$ we now only need to make sure $|log(\frac{u_t+1}{u_t})| < |log(\frac{v_t+1}{v_t})|$ and the convergence of $max(u(constant))$ and we can prove simultaneous optimizing these 2 function will eventually works.
3. The idea is interesting because we now only consider the order of the reward.
4. Robustness idea contains in their work.