Reinforcement Learning In The Presence Of Rare Events

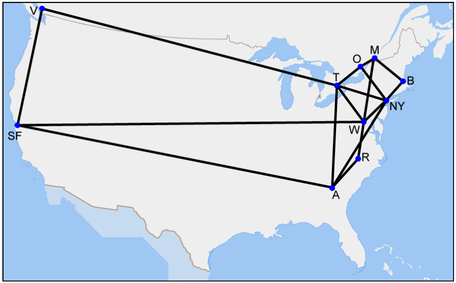
Jordan Frank, Shie Mannor, and Doina Precup

jordan.frank@cs.mcgill.ca
http://www.cs.mcgill.ca/∼jfrank8/

Reasoning and Learning Lab
McGill University
Montreal, Canada

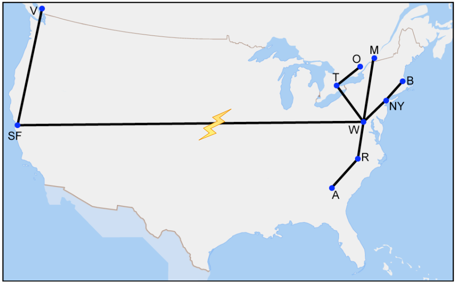
ICML'08
July 6, 2008

Motivation: Network Planning Problem



- Goal: Agent that can build and maintain a network.
- States: Network configuration, traffic demands, etc.
- Actions: Build or upgrade any link.
- Rewards:
  - Revenue for delivering traffic.
  - Significant penalty for undelivered traffic.

Motivation: Network Planning Problem



Problems:

- Large state space, large action space.
- Rare, very significant events (link failures).

Our Approach

- Most RL tasks work with simulators.
- Use techniques from simulation literature for variance reduction and rare event prediction:
  - Adaptive Importance Sampling
- Adapt for on-line RL:
  - Proofs of convergence.
  - Bias-variance results.

Importance Sampling

$$\mathbb{E}_p\left(h(X)\right) = \int h(x)p(x) = \int \frac{h(x)p(x)}{q(x)} q(x)dx = \mathbb{E}_q\left(h(X)w(X)\right)$$

- $w(x) = p(x)/q(x)$ – importance sampling *correction*.
- Consistent, unbiased estimator for $\mathbb{E}_p\left(h(X)\right)$ is:

$$\widehat{I} = \frac{1}{N}\sum_{i=1}^{N} h(X_i)w(X_i).$$

- Used before in RL (Precup et al. ICML'01) for off-policy learning.
- Variance depends on choice of sampling distribution (q).

Minimum-Variance IS Distribution

Theorem (Stats 101)

The choice of q that minimizes the variance of $\widehat{I}$ is

$$q^*(x) = \frac{|h(x)|p(x)}{\int |h(s)|p(s)ds}$$

- Problem: $\int h(s)p(s)ds$ is the quantity we are trying to estimate.
- ASA algorithm (Ahamed et al., 2006) adapted this for estimating expected total cost on discrete Markov chains.
  - Uses stochastic approximation to estimate minimum-variance sampling distribution.

Markov Decision Processes

- Set of states $\mathcal{S}$ and actions $\mathcal{A}$. Agent selects actions according to a policy $\pi(s,a) = \Pr\left(a_t = a | s_t = s\right)$.
- Environment dynamics defined by specifying the *transition probabilities* and the *rewards*

$$\mathcal{P}^a_{ss'} = \Pr\left(s_{t+1} = s' | s_t = s, a_t = a\right),$$
$$\mathcal{R}^a_{ss'} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}.$$

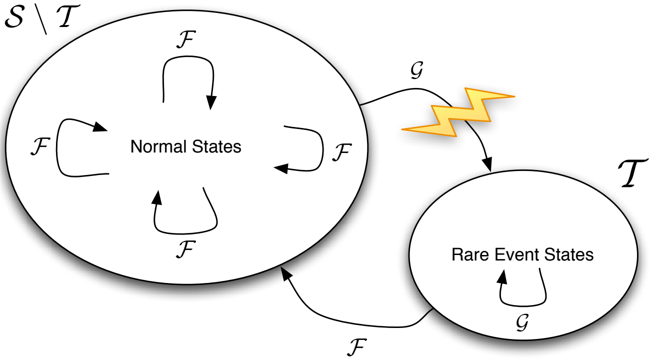
- The *value function* is given by

$$V^\pi(s) = \mathbb{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{k+1} | s_0 = s\right).$$

- The *Bellman equations* for V^π is

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(s,a) \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma V^\pi_{s'}\right].$$

MDPs with Rare Events



- $\varepsilon(s)$ is the true probability of rare event at state s.
- Assume $\mathcal{P}^a_{ss'} = (1 - \varepsilon(s))\mathcal{F}^a_{ss'} + \varepsilon(s)\mathcal{G}_{ss'}$.
- Assume simulator allows ε to be changed.
- $\mathcal{F}$ and $\mathcal{G}$ may not be known.

Rare event state sets

Definition

A subset of state $\mathcal{T} \subset \mathcal{S}$ is called a *rare event state set* for policy π if the following three properties hold:

1. For all $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{T}$, $\mathcal{F}^{a}_{ss'} = 0$.
2. There exists $s \in \mathcal{S}$, $s' \in \mathcal{T}$ such that $\mathcal{G}_{ss'} > 0$.
3. Let $V^{\pi}_{\mathcal{F}}$ denote the value function obtained by using $\mathcal{F}$ for the transition probabilities, then

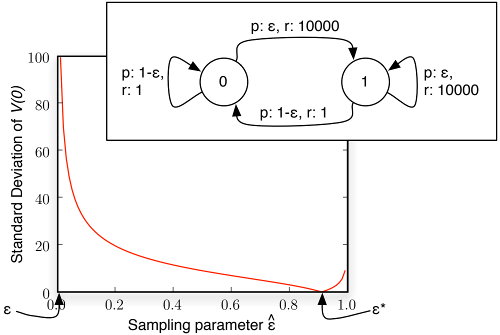
$$\exists s \in \mathcal{S} \text{ s.t. } |V^{\pi}_{\mathcal{F}}(s) - V^{\pi}(s)| > \Delta.$$

Rare Event Adaptive Stochastic Approximation

Optimal rare event parameter $\hat{\varepsilon}$:

$$\varepsilon^*(s) = \varepsilon(s) \frac{\sum_{s' \in \mathcal{T}} \mathcal{G}_{ss'} \sum_{a \in \mathcal{A}} \pi(s,a)[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')]}{V^\pi(s)}.$$

Example ($\gamma = 0$, $\varepsilon = 0.001$):



REASA Algorithm

- Sketch:
  - Use rare event prob. estimate $\hat{\varepsilon}(s)$ to generate rare events.
  - On each transition (s, a, s', r), calculate IS correction

$$w_s = \begin{cases} \varepsilon(s)/\hat{\varepsilon}(s) & \text{if } s \in \mathcal{T}, \\ (1-\varepsilon(s))/(1-\hat{\varepsilon}(s)) & \text{if } s \notin \mathcal{T}, \end{cases}$$

    and update trajectory IS correction W.
  - Update our value function estimate for s:

$$\widehat{V}^{\pi}(s) \leftarrow \widehat{V}^{\pi}(s) + \alpha W[w_s(r + \widehat{V}^{\pi}(s')) - \widehat{V}^{\pi}(s)].$$

  - Update the contribution to V^{π} of the rare events states ($T(s)$) and regular states ($U(s)$).
  - Update rare event parameter estimate $\hat{\varepsilon}$ (keep bounded using parameter $\delta \in (0, 1)$).

$$\hat{\varepsilon}(s) = \min\left(\max\left(\frac{\varepsilon(s)|T(s)|}{\varepsilon(s)|T(s)| + (1-\varepsilon(s))|U(s)|}, \delta\right), 1-\delta\right)$$

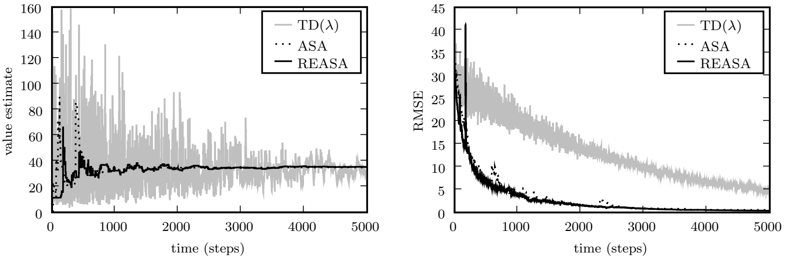
Theoretical Results: Convergence

- Convergence shown for both tabular and linear function approx. cases.
  - Tabular: Convergence of value function estimate to true value function.
  - FA: Convergence of value function to same value as TD with no IS would converge to.
- If $\forall s \in \mathcal{S}$, $\delta \leq \varepsilon^* \leq 1 - \delta$, then $\hat{\varepsilon} \rightarrow \varepsilon^*$.

Theoretical Results: Bias-variance

- Mannor et al. (2007) derive equations for bias and variance of temporal difference algorithms.
- Relies on count of minimally observed transitions. Rare events lead to loose bounds.
- We can split value function into two parts, $V^\pi_{\mathcal{F}}$ and $V^\pi_{\mathcal{G}}$, and consider each independently, leading to tighter bounds.
- Consequence of oversampling rare events is increased errors in estimates of $V^\pi_{\mathcal{F}}$, but improvements in estimates for $V^\pi_{\mathcal{G}}$ are much greater.

Random MDPs: Value estimate for State 0

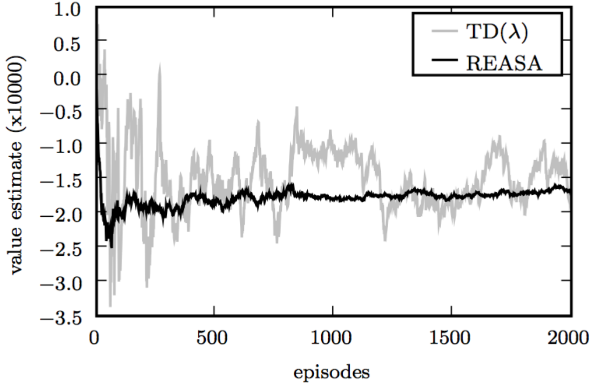


- One step in REASA and ASA is a single transition; one step in TD(λ) is 2300 transitions.
- Results are averaged over 70 runs.
- TD(λ) exhibits high variance.
- REASA and ASA converge quickly.
- ASA knows and can manipulate entire trans. prob. dist.

Network Planning Problem: Policy Evaluation

- Large state space: linear function approximation for value function (271 binary features).
- Tree network, policy upgrades links with over 90% utilization.
- Links go down approx. once per 4 years.
- $\varepsilon \approx 0.00896$.
- Compare TD(λ) and REASA.

Network Planning Problem: Result



- TD(λ) has much higher variance.
- REASA finds optimal $\hat{\varepsilon} \approx 0.155$, or one failure every 54 days.

Conclusions

- Incorporated variance reduction techniques from simulation literature into on-line RL algorithm.
- By not treating simulator as a “black box”, we can obtain significant improvements in performance.
- Large variance reduction with modest assumptions on simulator.
- Validated empirically on large real-world problem.
- Convergence guarantees and bias-variance analysis.

Future Work

- Consider variance in $\mathcal{F}$ and $\mathcal{G}$. Add UCB-like exploration.
- Incorporate REASA into policy optimization algorithm (eg. Sarsa).
- Make network planning task more interesting (bigger network, incorporate node failures, etc.).
- Better bias and variance results.
- Apply to more problems. Any suggestions are appreciated.
- Consider other types of parameterized transition probability distributions.

References

- Ahamed, T. P. I., Borkar, V. S. & Juneja, S. (2006).
  Adaptive importance sampling technique for Markov chains using stochastic approximation. *Oper. Res.*,*54*, 489–504.
- Bucklew, J. (2004).
  Introduction to Rare Event Simulation. Springer.
- Frank, J., Mannor, S. & Precup, D. (2008).
  Reinforcement learning in the presence of rare events. In *Proc. ICML'08*.
- Mannor, S., Simester, D., Sun, P., & Tsitsiklis, J. (2007).
  Bias and variance approximation in value function estimates. *Management Science*, *53*, 308.
- Precup, D., Sutton, R., & Dasgupta, S. (2001).
  Off-policy temporal-difference learning with function approximation. In *Proc. ICML'01*.
- Sutton, R. S. & Barto, A. G. (1998).
  Reinforcement Learning. The MIT Press.

Thank You

Questions?