Sorta Optimal Kinda Reinforcement Learning

Sham Kakade and John Langford

Approximately Optimal Approximate Reinforcement Learning

Sham Kakade and John Langford

Approximate methods for RL are crucial

1. Approximate value function methods have weak guarantees

2. Policy gradient methods fail (no exploration)

Another Approximate Algorithm?

1. Assume access to an ϵ-good 'supervised learning algorithm'

2. Present Conservative Policy Iteration (CPI)

3. CPI ceases 'quickly' and returns an 'approximately optimal' policy

4. guarantees depend on ϵ but not explicitly on the state-space size

Standard Definitions

- The value function:

$$V_\pi(s) \equiv (1-\gamma)E\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t)|\pi, s\right]$$

- The state-action value

$$Q_\pi(s,a) \equiv (1-\gamma)\mathcal{R}(s) + \gamma E_{s' \sim \text{next state}}\left[V_\pi(s')\right]$$

- Given start-state s_0, optimize $V_\pi(s_0)$

Less Standard Definitions

- The γ-discounted future state distribution for π,μ:

$$d_{\pi,\mu}(s) \propto \sum_t \gamma^t \mathsf{P}(s_t = s)$$

- Note that

$$V_\pi(s) = E_{s' \sim d_{\pi,s}} \left[\mathcal{R}(s') \right]$$

- The advantage:

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$$

- Actions with positive advantages are good

Approximate Value Function Methods

- use greedy updates $\pi \leftarrow \pi_{\text{greedy}}$

- with error $\varepsilon = ||\widetilde{V}_\pi - V_\pi||_\infty$, policy can degrade by

$$\frac{2\varepsilon\gamma}{1-\gamma}$$

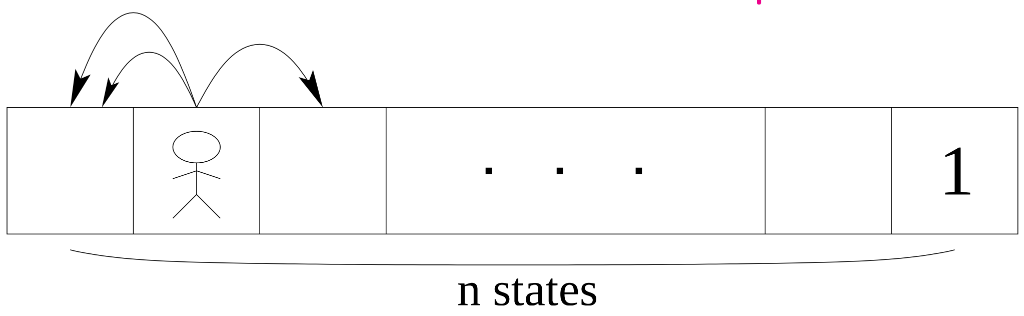
- this error is too large in practice

- average errors often used in practice

Policy Gradient Methods

$$\nabla V_\pi(s_0) = E_{s \sim d_{\pi,s_0}} [\nabla \pi(a|s) A_\pi(s,a)]$$

- suffer from random walk exploration



- sensible gradient estimate is 0
- exponentially large variance in gradient direction
- stuck far from optimal policy

Our Demands

1. verifiable policy improvement

2. the algorithm must halt 'quickly'

3. after halting, require a 'good' policy

A New Performance Measure

consider $E_{s\sim\mu}\left[V_\pi(s)\right]$

- use a more uniform distribution μ
- Make two assumptions

Assumption 1

$$E_{s\sim\mu}\left[V_\pi(s)\right]$$

- We assume the ability to 'restart' to $s \sim \mu$.

- Gradient methods assume ability to 'reset' to s_0

- much weaker than knowing the transition model

- substitute for exploration

Assumption 2

Assume we can do supervised learning with ε average (l_1) error

- supervised learning under $d_{\pi,\mu} \rightarrow \pi_{\text{greedy}}$

The Policy Update Rule

$$\pi \leftarrow (1-\alpha)\pi + \alpha\pi_{\text{greedy}}$$

- greedy update rule for $\alpha = 1$

- want improvement for 'small' α

The Policy Advantage

$$\pi \leftarrow (1-\alpha)\pi + \alpha\pi_{\text{greedy}}$$

- the gradient with respect to α is

$$E_{s\sim d_{\pi,\mu}}\left[E_{a\sim\pi_{\text{greedy}}}[A_\pi]\right]$$

- define this to be the Policy Advantage of $\pi_{\text{greedy}} = \mathcal{A}_{\pi,\mu}(\pi_{\text{greedy}})$ (under π, μ)

- measures degree to which π_{greedy} chooses better actions

- If $\mathcal{A} > 0$, then policy improvement for 'small' alpha

ϵ-Greedy Policies

- Define $\mathrm{OPT}\mathcal{A}_{\pi,\mu}$ to be the largest policy advantage

- The policy advantage of π_{greedy} is ε-close to optimal

- If $\mathrm{OPT}\mathcal{A}_{\pi,\mu}$ is large, then improvement possible

$$\mathrm{OPT}\mathcal{A}_{\pi,\mu} > \varepsilon \Rightarrow \mathcal{A}_{\pi,\mu}(\pi_{\text{greedy}}) > 0$$

- If $\mathrm{OPT}\mathcal{A}_{\pi,\mu}$ is small, improvement not guaranteed

Theorem 1: How Much Policy Improvement?

If $\mathcal{A} > 0$, using $\alpha = \frac{(1-\gamma)\mathcal{A}}{8} \Rightarrow$ improvement of $\frac{\mathcal{A}^2}{8}$

- proof considers the change in $d_{\pi,\mu}$
- justification for using ϵ-good π_{greedy}
- verification easy (just estimate $\mathcal{A}$)

Conservative Policy Iteration (CPI)

1. supervised learning under $d_{\pi,\mu} \rightarrow \pi_{\text{greedy}}$

2. Estimate the policy advantage of π_{greedy}

3. Is the policy advantage large (greater than ε)?

   - yes $\rightarrow$ update the policy

     $$\pi \leftarrow (1-\alpha)\pi + \alpha\pi_{\text{greedy}}$$

     with $\alpha = \frac{(1-\gamma)\mathcal{A}}{8}$ and go to (1).

   - no $\rightarrow$ HALT and return π.

Theorem 2: CPI's guarantees

1. The performance measure under μ improves

2. Halts in at most $O\left(\frac{1}{\varepsilon^2}\right)$ steps

3. The output π has small $\mathrm{OPT}\mathcal{A}_{\pi,\mu}$.

Theorem 3: The Quality of the Output

- Assume π has small optimal policy advantage.

- For an optimal policy π^*,

$$V_{\pi^*}(s_0) - V_\pi(s_0) \leq \frac{\varepsilon}{(1-\gamma)^2} \left\| \frac{d_{\pi^*,s_o}}{\mu} \right\|_\infty$$

- $\left\| \frac{d_{\pi^*,s_0}}{\mu} \right\|_\infty$ is a mismatch between where an optimal policy visits and where we give optimization weight.

The Mismatch Factor

- $\left\| \frac{d_{\pi^*,s_0}}{\mu} \right\|_\infty$ suggests choosing μ to match optimal

- Uniform μ most robust

- Some policies are 'learnable' using supervised learning with average ε error

- Weights also important for approximate LP

CPI's Ingredients

- Uses μ for 'restarts'
- Supervised learning under 'changing' $d_{\pi,\mu}$
- 'conservative' updates

Further Directions

- Practical variants?
- Weaken Assumption 1: Approximate Exploration?