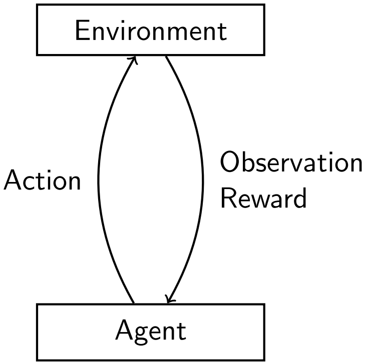
Bayesian Reinforcement Learning with Exploration

Tor Lattimore[1] & Marcus Hutter[2]

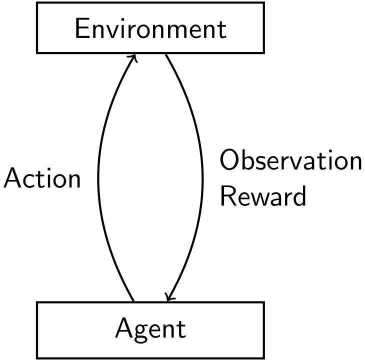
[1]University of Alberta [2]Australian National University

History-Based Reinforcement Learning



- Take actions
- Receive observations and rewards
- World dynamics are unknown
- Maximise rewards

History-Based Reinforcement Learning



- Take actions
- Receive observations and rewards
- World dynamics are unknown
- Maximise rewards
- No i.i.d. assumption
- No Markov assumption
- No state is ever seen more than once

Notation

- History $\equiv$ sequence of action/observation/reward tuples
- Policy π : History $\rightarrow$ Action
- Environment μ : History $\times$ Action $\rightsquigarrow$ Reward $\times$ Observation
- Policy and environment interact to generate random history sequence

$$a_1 o_1 r_1, \ldots, a_t, o_t, r_t$$

Notation

- History $\equiv$ sequence of action/observation/reward tuples
- Policy π : History $\rightarrow$ Action
- Environment μ : History $\times$ Action $\rightsquigarrow$ Reward $\times$ Observation
- Policy and environment interact to generate random history sequence

$$a_1 o_1 r_1, \ldots, a_t, o_t, r_t$$

- $\gamma \in [0, 1)$ is discount factor
- $V_\mu^\pi(x)$ is value given history sequence $x = a_1 o_1 r_1, \ldots, a_{t-1}, o_{t-1}, r_{t-1}$

$$V_\mu^\pi(x) = \mathbf{E}_\mu^\pi \left[\sum_{s=t}^{\infty} \gamma^{s-t} r_s \middle| x \right]$$

- π_μ^* is the optimal policy (maximising $V_\mu^{\pi^*}$) and V_μ^* is its value

Objective – Minimise Sample-Complexity

Given:

- Set of environments $\mathcal{M}$
- Accuracy $\varepsilon > 0$ and confidence $\delta > 0$

Goal: Find π that minimises sample-complexity $N = N(\mathcal{M}, \pi, \delta, \varepsilon)$

$$\forall \mu \in \mathcal{M}, \qquad P_\mu^\pi \left\{ \sum_{t=1}^{\infty} \mathbb{1}\left\{ V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon \right\} > N \right\} \leq \delta$$

Objective – Minimise Sample-Complexity

Given:

- Set of environments $\mathcal{M}$
- Accuracy $\varepsilon > 0$ and confidence $\delta > 0$

Goal: Find π that minimises sample-complexity $N = N(\mathcal{M}, \pi, \delta, \varepsilon)$

$$\forall \mu \in \mathcal{M}, \qquad P_\mu^\pi \left\{ \sum_{t=1}^{\infty} \mathbb{1}\left\{ V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon \right\} > N \right\} \leq \delta$$

- Uniform criterion
- Unobtainable (in general) unless $\mathcal{M}$ is finite (or compact)

Objective – Minimise Sample-Complexity

Given:

- Set of environments $\mathcal{M}$
- Accuracy $\varepsilon > 0$ and confidence $\delta > 0$

Goal: Find π that minimises sample-complexity $N = N(\mathcal{M}, \pi, \delta, \varepsilon)$

$$\forall \mu \in \mathcal{M}, \qquad P_\mu^\pi \left\{ \sum_{t=1}^{\infty} \mathbb{1}\left\{ V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon \right\} > N \right\} \leq \delta$$

- Uniform criterion
- Unobtainable (in general) unless $\mathcal{M}$ is finite (or compact)

Assume from now: $|\mathcal{M}| = K$

Bayesian Prediction

- Briefly forget control – no policy
- Bayesian mixture:

$$(\xi\text{-probability of observing } x) \equiv P_\xi(x) = \sum_{\nu \in \mathcal{M}} w_\nu P_\nu(x)$$

Bayesian Prediction

- Briefly forget control – no policy
- Bayesian mixture:

$$(\xi\text{-probability of observing } x) \equiv P_\xi(x) = \sum_{\nu \in \mathcal{M}} w_\nu P_\nu(x)$$

- d-step total variation distance given history x

$$\delta_d(\mu, \xi | x) = \frac{1}{2} \sum_{y \in \mathcal{H}^d} |P_\mu(y|x) - P_\xi(y|x)|$$

Bayesian Prediction

- Briefly forget control – no policy
- Bayesian mixture:

$$(\xi\text{-probability of observing } x) \equiv P_\xi(x) = \sum_{\nu \in \mathcal{M}} w_\nu P_\nu(x)$$

- d-step total variation distance given history x

$$\delta_d(\mu, \xi | x) = \frac{1}{2} \sum_{y \in \mathcal{H}^d} |P_\mu(y|x) - P_\xi(y|x)|$$

Theorem: Let x be the infinite history generated by μ and $t_1, t_2, \ldots$ a sequence of stopping times with $t_{k+1} \geq t_k + d_k$ almost surely with d_k measurable at time-step t_k. Then

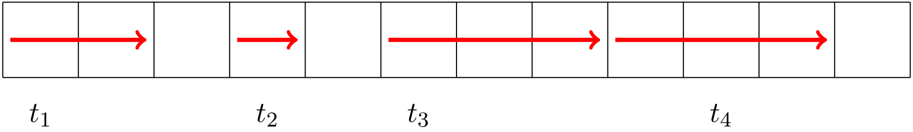
$$P_\mu \left\{ \sum_{k=1}^{\infty} \delta_{d_k}^2(\mu, \xi | x_{<t_k}) \leq \log \frac{1}{w_\mu \delta^2} \right\} \leq \delta$$

Bayesian Prediction

Theorem: Let x be the infinite history generated by μ and $t_1, t_2, \ldots$ a sequence of stopping times with $t_{k+1} \geq t_k + d_k$ almost surely with d_k measurable at time-step t_k. Then

$$P_\mu \left\{ \sum_{k=1}^{\infty} \delta^2_{d_k}(\mu, \xi | x_{<t_k}) \leq \log \frac{1}{w_\mu \delta^2} \right\} \leq \delta$$

Example:



From Prediction to Confidence Sets

- Define confidence set
- Choose $w_\nu = 1/K$ (uniform prior)

$$\mathcal{M}_k := \left\{ \nu : \sum_{j=1}^{k} \delta^2_{d_k}(\mu, \xi | x_{<t_k}) \leq \log \frac{K}{\delta^2} \right\}$$

"$\mathcal{M}_k$ is the set of environments for which the prediction has so far been acceptable"

From Prediction to Confidence Sets

- Define confidence set
- Choose $w_\nu = 1/K$ (uniform prior)

$$\mathcal{M}_k := \left\{ \nu : \sum_{j=1}^{k} \delta_{d_k}^2(\mu, \xi | x_{<t_k}) \leq \log \frac{K}{\delta^2} \right\}$$

"$\mathcal{M}_k$ is the set of environments for which the prediction has so far been acceptable"

- $\mu \in \mathcal{M}_k$ for all k with probability at least $1 - \delta$
- Similar idea and benefits as "Online-to-Confidence" by Abbasi-Yadkori et. al. (2012)

How to Act

If confident, then Bayes, else explore

How to Act

If confident, then Bayes, else explore

Proposition: If $d \approx \frac{1}{1-\gamma} \log \frac{1}{\varepsilon(1-\gamma)}$ and $\delta_d(\mu^\pi, \xi^\pi | x) \leq \varepsilon(1-\gamma)$, then

$$V_\mu^\pi(x) - V_\xi^\pi(x) \leq \varepsilon$$

($d \equiv$ effective horizon, $\mu^\pi \equiv$ measure on histories induced by μ and π)

How to Act

If confident, then Bayes, else explore

Proposition: If $d \approx \frac{1}{1-\gamma} \log \frac{1}{\varepsilon(1-\gamma)}$ and $\delta_d(\mu^\pi, \xi^\pi | x) \leq \varepsilon(1-\gamma)$, then

$$V_\mu^\pi(x) - V_\xi^\pi(x) \leq \varepsilon$$

($d \equiv$ effective horizon, $\mu^\pi \equiv$ measure on histories induced by μ and π)

Corollary: If $\delta_d(\mu^\pi, \xi^\pi | x) \leq \varepsilon(1-\gamma)$ for $\pi \in \{\pi_\mu^*, \pi_\xi^*\}$, then

$$V_\mu^*(x) - V_\mu^{\pi_\xi^*}(x) \lesssim \varepsilon \qquad \text{(Bayes is nearly optimal)}$$

Algorithm

1. **Input:** $\mathcal{M} = \{\nu_i\}_{i=1}^K$, discount γ, accuracy ε, confidence δ
   $d \leftarrow \frac{1}{1-\gamma} \log \frac{1}{\varepsilon(1-\gamma)}$ and $k \leftarrow 0$
2. Compute differences in policies:

$$\Pi^* = \{\pi^*_\nu : \nu \in \mathcal{M}\} \cup \left\{\pi^*_\xi\right\}$$

$$\pi = \arg\max_{\pi \in \Pi^*} \max_{\nu \in \mathcal{M}} \delta^d_x(\nu^\pi, \xi^\pi)$$

$$\Delta = \max_{\pi \in \Pi^*, \nu \in \mathcal{M}} \delta^d_x(\nu^\pi, \xi^\pi)$$

3. **If** $\Delta \gtrsim \varepsilon(1-\gamma)$
   - $k \leftarrow k + 1$ and $t_k =$ current time-step and $d_k = d$
   - Follow policy π for d time-steps
4. **Else**
   - $k \leftarrow k + 1$ and $t_k =$ current time-step and $d_k = 1$
   - Follow policy π^*_ξ for 1 time-step
5. Update plausible environments and **Goto** 2

Why it Works

Either

Algorithm is confident, when it is nearly optimal

Or

Algorithm is exploring, when it is gaining information

Theorems

Assume rewards are in $[0, 1]$

Theorem

If $|\mathcal{M}| = K$, *then* $N(\varepsilon, \delta) \in O\left(\frac{K}{\varepsilon^2(1-\gamma)^3}\left(\log\frac{K}{\delta}\right)\left(\log\frac{1}{\varepsilon(1-\gamma)}\right)\right)$.

Theorems

Assume rewards are in $[0, 1]$

Theorem

If $|\mathcal{M}| = K$*, then* $N(\varepsilon, \delta) \in O\left(\frac{K}{\varepsilon^2(1-\gamma)^3}\left(\log\frac{K}{\delta}\right)\left(\log\frac{1}{\varepsilon(1-\gamma)}\right)\right)$.

Theorem

For every policy and sufficiently small ε, δ

$$N(\varepsilon, \delta) \in \Omega\left(\frac{K}{\varepsilon^2(1-\gamma)^3}\log\frac{K}{\delta}\right).$$

- Shaves numerous logarithmic factors from previous work (L & Hutter, ICML 2013)

No Optimism?

Standard approach: Optimisim in the face of uncertainty

$$\text{(Optimistic Policy)} \qquad \pi = \arg\max_{\pi} \max_{\nu \in \mathcal{M}_t} V_{\nu}^{\pi}(x_{<t})$$

where $\mathcal{M}_t \equiv$ set of plausible environments

Very successful: Bandits, Linear Bandits, MDPs, and many problems in online learning

No Optimism?

Standard approach: Optimisim in the face of uncertainty

$$\text{(Optimistic Policy)} \qquad \pi = \arg\max_{\pi} \max_{\nu \in \mathcal{M}_t} V^{\pi}_{\nu}(x_{<t})$$

where $\mathcal{M}_t \equiv$ set of plausible environments

Very successful: Bandits, Linear Bandits, MDPs, and many problems in online learning

Does not work (easily) for general RL

No Optimism?

Standard approach: Optimisim in the face of uncertainty

$$\text{(Optimistic Policy)} \qquad \pi = \arg\max_{\pi} \max_{\nu \in \mathcal{M}_t} V^{\pi}_{\nu}(x_{<t})$$

where $\mathcal{M}_t \equiv$ set of plausible environments

Very successful: Bandits, Linear Bandits, MDPs, and many problems in online learning

Does not work (easily) for general RL

Even if $\mathcal{M}_t = \mathcal{M}_{t+1}$, it is **not true** that

$$\arg\max_{\pi} \max_{\nu \in \mathcal{M}_t} V^{\pi}_{\nu}(x_{<t}) = \arg\max_{\pi} \max_{\nu \in \mathcal{M}_{t+1}} V^{\pi}_{\nu}(x_{<t+1})$$

Sequence of optimistic policies are not compatible

Bandit Connection ($\gamma = 0$)

- $\gamma = 0 \implies d = 1$
- Still not i.i.d.
- Can use optimism
- Sample-complexity becomes

$$\frac{K}{\varepsilon^2} \log \frac{K}{\delta}$$

- For i.i.d. bandits the sample-complexity is optimised by the median elimination algorithm (Even-Dar et. al. 2006)

$$\frac{K}{\varepsilon^2} \log \frac{1}{\delta}$$

- So non-i.i.d. really is harder

Other Results

- Dependence on K can be significantly reduced if environments share structure (eg., MDPs)
- Dependence on ε can be reduced if environments are well separated
- Asymptotic results possible for countable classes

Some Downsides

- Computationally expensive unless $\gamma = 0$
- Finite $\mathcal{M}$
- Algorithm fails if $\mu \notin \mathcal{M}$
- Worst-case linear dependence on K is pretty bad

Conclusions

Summary

- Improved algorithm that optimises sample-complexity for general RL
- Nearly matching upper/lower bounds
- Algorithm is adaptive to easier environments
- Bounds on sample-complexity improve on previously known

Future

- Explore non-uniform bounds
- Explore possibility of regret bounds (assumptions necessary)
- Investigate computational issues in specific environments (eg., MDPs)
- Extensions to continuous/compact classes