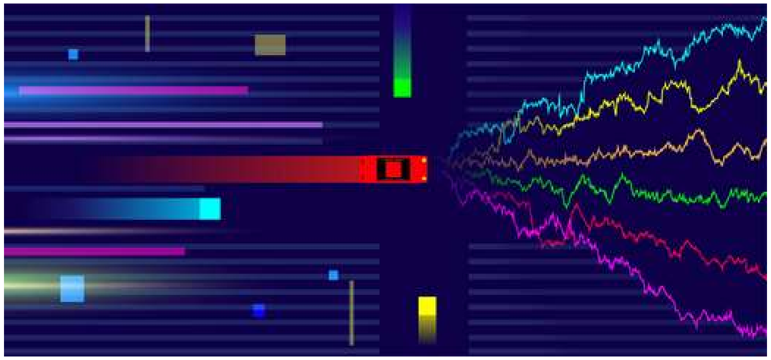
RL for Operations, 2022

Online Tabular Algorithms

Sean Sinclair
Cornell University



Finite Horizon

A MDP is defined by: $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, r, T, s_0, H\}$

$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$r_h : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$	Reward
$T_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$	Transitions
H	Time horizon
$\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})$	Policy

Bellman Equation

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(S_{h'}, A_{h'}) \mid S_h = s, A_{h'} \sim \pi(S_{h'}), S_{h'+1} \sim T_{h'}(\cdot \mid S_{h'}, A_{h'})\right]$$

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(S_{h'}, A_{h'}) \mid (S_h, A_h) = (s,a), A_{h'} \sim \pi(S_{h'}), S_{h'+1} \sim T_{h'}(\cdot \mid S_{h'}, A_{h'})\right]$$

The Bellman Equations note that:

$$V_h^\pi(s) = \mathbb{E}_{A \sim \pi_h(s)}[r_h(s,A) + \mathbb{E}_{S' \sim T_h(\cdot|s,A)}[V_{h+1}^\pi(S')]]$$

$$Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}^\pi(S')]$$

Main Question

Given a MDP, how do we find the optimal policy?

Fully Known Model

- Reward function, transition distribution fully known
- Understand computational complexity to scale to large problems

Generative Model

- Sample from reward function / transition distribution from arbitrary (state,action)
- Understand statistical complexity to scale to large problems
- No issue of dynamic environment

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Main Question

Given a MDP, how do we find the optimal policy?

Fully Known Model

- Reward function, transition distribution fully known
- Understand computational complexity to scale to large problems

Generative Model

- Sample from reward function / transition distribution from arbitrary (state,action)
- Understand statistical complexity to scale to large problems
- No issue of dynamic environment

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Main Question

Given a MDP, how do we find the optimal policy?

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Main Question

Given a MDP, how do we find the optimal policy?

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Unknown transition + reward

Over sequence of episodes:

Main Question

Given a MDP, how do we find the optimal policy?

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Unknown transition + reward

Over sequence of episodes:
- Pick current policy π^k

Main Question

Given a MDP, how do we find the optimal policy?

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Unknown transition + reward

Over sequence of episodes:

- Pick current policy π^k
- Execute over H steps in MDP

Main Question

Given a MDP, how do we find the optimal policy?

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Unknown transition + reward

Over sequence of episodes:

- Pick current policy π^k
- Execute over H steps in MDP
- Collect dataset and update policy

$$\{(S_1^k, A_1^k, R_1^k), \ldots, (S_H^k, A_H^k, R_H^k)\}$$

Main Question

Unknown transition + reward

Over sequence of episodes:

- Pick current policy π^k
- Execute over H steps in MDP
- Collect dataset and update policy

$$\{(S_1^k, A_1^k, R_1^k), \ldots, (S_H^k, A_H^k, R_H^k)\}$$

Goal: Minimize regret:

$$\textsc{Regret}(K) = \sum_{k=1}^{K} V_1^*(s_0) - V_1^{\pi^k}(s_0)$$

Recall................

Given a MDP, how do we find the optimal policy?

Fully Known Model

- Reward function, transition distribution fully known
- Understand computational complexity to scale to large problems

Two Approaches

Value Iteration
Policy Iteration

Recall……………

Given a MDP, how do we find the optimal policy?

Fully Known Model

- Reward function, transition distribution fully known
- Understand computational complexity to scale to large problems

Two Approaches

[redacted]

Main Question

Given a MDP, how do we find the optimal policy?

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Two Approaches

Value Based
Policy Based

Main Question

Given a MDP, how do we find the optimal policy?

Online Model

- Sample trajectory under current policy, update policy, repeat
- Understand statistical complexity
- *"Most complex"*, additional correlations in estimates

Two Approaches

Value Based
Policy Based

Typically done with function approximation, will discuss later

Value Based

The Bellman Optimality Equations note that:

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$$
$$Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}^*(S')]$$

Value Based

The Bellman Optimality Equations note that:

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$$

$$Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}^*(S')]$$

Model-Based

- Maintain estimates of reward and transition
- Plug estimates into Bellman equations for estimated V^*, Q^*
- Play greedy w.r.t. Q^*
- Time complexity / storage scales S^2A

Model Free

- Only maintain estimates of V^*, Q^* using fixed point
- Play greedy w.r.t. Q^*
- Better time complexity / storage (only SA)

Value Based

The Bellman Optimality Equations note that:

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$$
$$Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}^*(S')]$$

Model-Based

- Maintain estimates of reward and transition
- Plug estimates into Bellman equations for estimated V^*, Q^*
- Play greedy w.r.t. Q^*
- Time complexity / storage scales S^2A

Model Free

- Only maintain estimates of V^*, Q^* using fixed point
- Play greedy w.r.t. Q^*
- Better time complexity / storage (only SA)

Model Based

The Bellman Optimality Equations note that:

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$$
$$Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}^*(S')]$$

[Azar2017]

Model Based

The Bellman Optimality Equations note that:

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$$
$$Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}^*(S')]$$

At start of episode k, have collected data: $\mathcal{D}^k$

[Azar2017]

Model Based

The Bellman Optimality Equations note that:

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$$
$$Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}^*(S')]$$

At start of episode k, have collected data: $\mathcal{D}^k$

Estimate reward and transition via empirical:

$\overline{r}_h^k(s,a) = \frac{1}{n_h(s,a)} \sum_{(s,a)\in\mathcal{D}^k} R_h^k$ $\overline{T}_h^k(\cdot \mid s,a) = \frac{1}{n_h(x,a)} \sum_{(s,a,S_{h+1}^{k'})\in\mathcal{D}^k} \delta_{S_{h+1}^{k'}}$

$n_h(s,a)$ Number of times (s,a) visited

[Azar2017]

Model Based

The Bellman Optimality Equations note that:

$$V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a)$$
$$Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}^*(S')]$$

At start of episode k, have collected data: $\mathcal{D}^k$

Estimate reward and transition via empirical:

$$\overline{r}_h^k(s,a) = \frac{1}{n_h(s,a)} \sum_{(s,a) \in \mathcal{D}^k} R_h^k \qquad \overline{T}_h^k(\cdot \mid s,a) = \frac{1}{n_h(x,a)} \sum_{(s,a,S_{h+1}^{k'}) \in \mathcal{D}^k} \delta_{S_{h+1}^{k'}}$$

$n_h(s,a)$ Number of times (s,a) visited

Plug estimates into Bellman Optimality Equations

[Azar2017]

Model Based

Estimate reward and transition via empirical:

$$\overline{r}^k_h(s,a) = \frac{1}{n_h(s,a)} \sum_{(s,a)\in\mathcal{D}^k} R^k_h \qquad \overline{T}^k_h(\cdot \mid s,a) = \frac{1}{n_h(x,a)} \sum_{(s,a,S^{k'}_{h+1})\in\mathcal{D}^k} \delta_{S^{k'}_{h+1}}$$

$n_h(s,a)$ Number of times (s,a) visited

Plug estimates into Bellman Optimality Equations

$$\overline{V}^k_h(s) = \max_{a\in\mathcal{A}} \overline{Q}^k_h(s,a)$$

$$\overline{Q}^k_h(s,a) = \overline{r}^k_h(s,a) + \mathbb{E}_{S'\sim\overline{T}^k_h(\cdot|s,a)}[\overline{V}^k_{h+1}(S')]$$

$$\pi^k_h(s) = \operatorname*{argmax}_{a\in\mathcal{A}} \overline{Q}^k_h(s,a)$$

[Azar2017]

Model Based

Estimate reward and transition via empirical:

$\overline{r}_h^k(s,a) = \frac{1}{n_h(s,a)} \sum_{(s,a)\in\mathcal{D}^k} R_h^k$ $\quad\quad$ $\overline{T}_h^k(\cdot \mid s,a) = \frac{1}{n_h(x,a)} \sum_{(s,a,S_{h+1}^{k'})\in\mathcal{D}^k} \delta_{S_{h+1}^{k'}}$

$n_h(s,a)$ Number of times (s,a) visited

Plug estimates into Bellman Optimality Equations

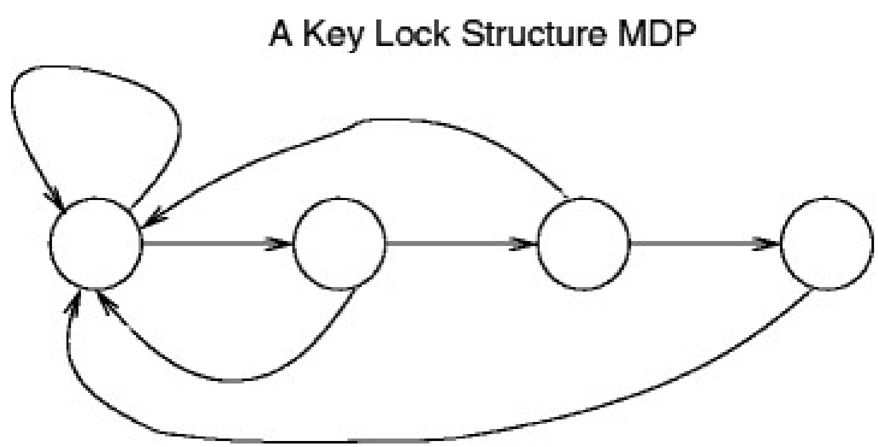
$$\overline{V}_h^k(s) = \max_{a\in\mathcal{A}} \overline{Q}_h^k(s,a)$$
$$\overline{Q}_h^k(s,a) = \overline{r}_h^k(s,a) + \mathbb{E}_{S'\sim\overline{T}_h^k(\cdot|s,a)}[\overline{V}_{h+1}^k(S')]$$
$$\pi_h^k(s) = \operatorname*{argmax}_{a\in\mathcal{A}} \overline{Q}_h^k(s,a)$$

Empirical value iteration with reward and transition estimates

[Azar2017]

Exploration

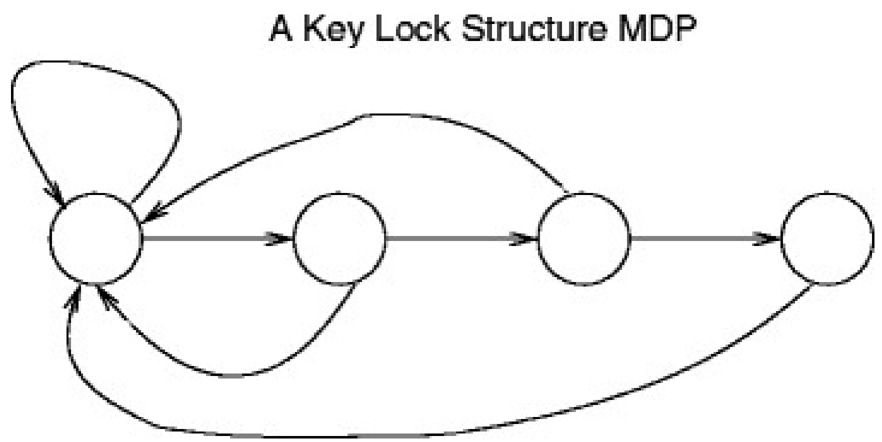
Unfortunately this algorithm is missing one key ingredient



[Azar2017]

Exploration

Unfortunately this algorithm is missing one key ingredient

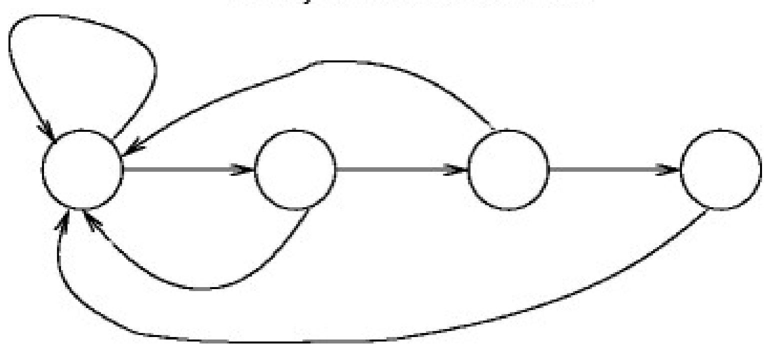


Without exploration, no reason for algorithm to explore to unobserved (s,a) pairs

[Azar2017]

Model Based

A Key Lock Structure MDP



$$\overline{V}_h^k(s) = \max_{a \in \mathcal{A}} \overline{Q}_h^k(s,a)$$

$$\overline{Q}_h^k(s,a) = \overline{r}_h^k(s,a) + \mathbb{E}_{S' \sim \overline{T}_h^k(\cdot|s,a)}[\overline{V}_{h+1}^k(S')] + \boxed{\iota \frac{1}{\sqrt{n_h(s,a)}}}$$

$$\pi_h^k(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \overline{Q}_h^k(s,a)$$

Empirical value iteration with reward and transition estimates and exploration bonuses

[Azar2017]

Model Based

$$\overline{V}_h^k(s) = \max_{a \in \mathcal{A}} \overline{Q}_h^k(s,a)$$

$$\overline{Q}_h^k(s,a) = \overline{r}_h^k(s,a) + \mathbb{E}_{S' \sim \overline{T}_h^k(\cdot|s,a)}[\overline{V}_{h+1}^k(S')] + \iota \frac{1}{\sqrt{n_h(s,a)}}$$

$$\pi_h^k(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \overline{Q}_h^k(s,a)$$

Informal Theorem: In a H-step MDP we have that:

$$\textsc{Regret}(K) \leq H^2 \sqrt{S^2 A K}$$

- Optimal dependence on K
- Suboptimal time + space complexity
- Dependence on H still current research

[Azar2017]

Model Based

$$\overline{V}_h^k(s) = \max_{a \in \mathcal{A}} \overline{Q}_h^k(s,a)$$

$$\overline{Q}_h^k(s,a) = \overline{r}_h^k(s,a) + \mathbb{E}_{S' \sim \overline{T}_h^k(\cdot|s,a)}[\overline{V}_{h+1}^k(S')] + \boxed{\iota \frac{1}{\sqrt{n_h(s,a)}}}$$

$$\pi_h^k(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \overline{Q}_h^k(s,a)$$

Regret guarantees are worst case, don't capture specific problem structure

In practice: exploration is done via ϵ exploration or bonus terms are tuned for performance

[Azar2017]

Model Free

If $V_h(s) = \max_{a \in \mathcal{A}} r_h(s,a) + \mathbb{E}_{S' \sim T_h(\cdot|s,a)}[V_{h+1}(S')]$

then $V(s) = V^*(s) \forall s$

[Robbins-Monro]

Model Free

If $V_h(s) = \max_{a\in\mathcal{A}} r_h(s,a) + \mathbb{E}_{S'\sim T_h(\cdot|s,a)}[V_{h+1}(S')]$

then $V(s) = V^*(s) \forall s$

From stochastic approximation:

Can use “stochastic” root finding algorithms to find the zero of:

$$\Delta_n(s,a) = Q_n(x,a) - Q^*(x,a)$$

[Robbins-Monro]

Stochastic Approximation

Let $\Delta_{n+1}(x) = (1 - \alpha_n(x))\Delta_n(x) + \beta_n(x)F_n(x)$

Converges to zero almost surely if:

- State space finite
- $\|\mathbb{E}[F_n(x)]\| \leq \gamma\|\Delta_n(x)\|$

From stochastic approximation:

Can use "stochastic" root finding algorithms to find the zero of:

$$\Delta_n(s,a) = Q_n(x,a) - Q^*(x,a)$$

[Robbins-Monro]

Model Free

Results in following update procedure:

$$\overline{V}_h^k(s) = \max_{a \in \mathcal{A}} \overline{Q}_h^k(s, a)$$

$$\overline{Q}_h^{k+1}(S_h^k, A_h^k) = (1 - \alpha_t)\overline{Q}_h^k(S_h^k, A_h^k) + \alpha_t(R_h^k + \overline{V}_h^k(S_{h+1}^k) + \boxed{\iota \frac{1}{\sqrt{t}}})$$

$$\pi_h^k(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \overline{Q}_h^k(s, a)$$

Empirical fixed point iteration with exploration bonuses

[Jin2018]

Model Free

$$\overline{V}_h^k(s) = \max_{a \in \mathcal{A}} \overline{Q}_h^k(s,a)$$

$$\overline{Q}_h^{k+1}(S_h^k, A_h^k) = (1-\alpha_t)\overline{Q}_h^k(s,a) + \alpha_t(R_h^k + \overline{V}_h^k(S_{h+1}^k) + \iota\frac{1}{\sqrt{t}})$$

$$\pi_h^k(s) = \operatorname{argmax}_{a \in \mathcal{A}} \overline{Q}_h^k(s,a)$$

Informal Theorem: In a H-step MDP we have that:

$$\textsc{Regret}(K) \leq H^{3/2}\sqrt{SAK}$$

- Strong relation to theory of Stochastic Approximation (Robbins Munro)
- Optimal dependence on K
- Better time + space complexity than model-based algorithms
- Dependence on H still current research

[Jin2018]

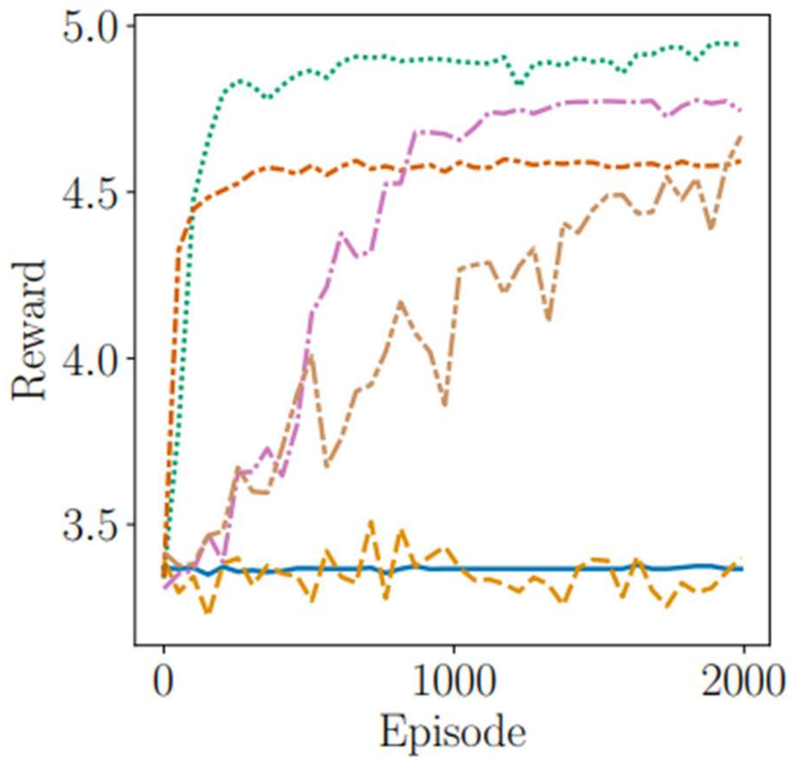
Model Free vs Model Based

Some folklore comparisons:

- Model based has better performance than model free due to how information is “propagated” across (s,a) pairs
- Increased computational overhead for model based algorithms
- Model based easier to implement, can use existing code for solving Bellman equations

Model Free vs Model Based

We will see a comparison this afternoon!



References

[Jin2018] Chi Jin, Zeyuan Allen-Zhu, Sebastian Bubeck, Michael Jordan. "Is Q Learning Provably Efficient?" *NeurIPS*, 2018.
[Sutton2018] Richard Sutton. "Reinforcement Learning: An Introduction." *MIT Press*, 2018.
[Azar2017] Mohammad Gheshlaghi Azar, Ian Osband, Rémi Munos. "Minimax Regret Bounds for Reinforcement Learning". *ICML*, 2017.