Approximate Value Iteration and Variants

Chen-Yu Wei

Value Iteration

$$V^{(k)}(s) \leftarrow \max_a \Big\{ \underbrace{R(s,a) + \gamma \sum_{s'} P(s'|s,a) \underbrace{V^{(k-1)}(s')}_{\max_{a'} Q^{(k-1)}(s',a')}}_{Q^{(k)}(s,a)} \Big\}$$

For $k = 1,\ 2, \ldots$

$$\forall s, a, \qquad Q^{(k)}(s,a) \leftarrow \underbrace{\boxed{R(s,a)}}_{\text{unknown}} + \gamma \sum_{s'} \underbrace{\boxed{P(s'|s,a)}}_{\text{unknown}} \max_{a'} Q^{(k-1)}(s',a')$$

Idea: In each iteration, use multiple samples to estimate the right-hand side.

Least-Square Value Iteration (LSVI)

For $k = 1,\ 2, \ldots$

Obtain n samples $\mathcal{D}^{(k)} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{n}$ where $\mathbb{E}[r_i] = R(s_i, a_i),\ s_i' \sim P(\cdot\,|s_i, a_i)$

We want these samples to be "exploratory"

Perform **regression** on $\mathcal{D}^{(k)}$ to find $Q^{(k)}$ such that

$$Q^{(k)}(s,a) \approx R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} Q^{(k-1)}(s', a')\right]$$

Tabular $\quad \forall s, a, \quad Q^{(k)}(s,a) = \dfrac{\sum_{i=1}^{n} \mathbb{I}\{(s_i, a_i) = (s,a)\}\left(r_i + \gamma \max_{a'} Q^{(k-1)}(s_i', a')\right)}{\sum_{i=1}^{n} \mathbb{I}\{(s_i, a_i) = (s,a)\}}$

$\overset{\mathbb{E}}{=} R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} Q^{(k-1)}(s', a')\right]$

General function approximation $\quad \theta_k = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \left(Q_\theta(s_i, a_i) - r_i - \gamma \max_{a'} Q_{\theta_{k-1}}(s_i', a')\right)^2$

Linear function approximation $\quad \theta_k = \left(\lambda I + \sum_{i=1}^{(n^k)} \phi(s_i, a_i)\phi(s_i, a_i)^\top\right)^{-1} \left(\sum_{i=1}^{(n_k)} \phi(s_i, a_i)\left(r_i + \gamma \max_{a'} \phi(s_i', a')^\top \theta_{k-1}\right)\right)$

Comparison with Contextual Bandits

Env

$\xrightarrow{x_t}$

$\xleftarrow{a_t}$

$\xrightarrow{r_t}$

Exploration

$$p_t(a) \propto e^{\lambda\, \hat{R}(x_t, a)}$$

$$a_t = \operatorname*{argmax}_a \left(\hat{R}(x_t, a) + b_t(a)\right)$$

$$\ldots$$

Regression

Fit $\hat{R}(x_i, a_i) \approx r_i$

Env

$\xrightarrow{s_t}$

$\xleftarrow{a_t}$

$\xrightarrow{r_t}$

Exploration

$$p_t(a) \propto e^{\lambda\, Q^{(k)}(s_t, a)}$$

$$a_t = \operatorname*{argmax}_a \left(Q^{(k)}(s_t, a) + b_t(a)\right)$$

$$\ldots$$

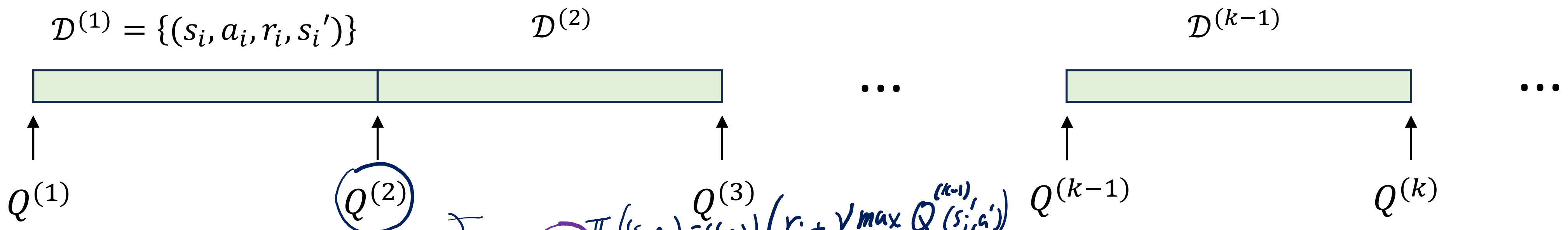
Value Iteration + Regression

For $k = 1, 2, \ldots$

Fit $Q^{(k)}(s_i, a_i) \approx r_i + \gamma \max_{a'} Q^{(k-1)}(s_i', a')$

It is Valid to Reuse Samples

(e.g., using ϵ-greedy)



$$Q^{(k)}(s,a) = \frac{\sum_{(s_i,a_i,r_i,s_i')\in D^{(k-1)}} \mathbb{I}\big((s_i,a_i)=(s,a)\big)\big(r_i+\gamma \max_{a'} Q^{(k-1)}(s_i',a')\big)}{\sum_{(s_i,a_i,r_i,s_i')\in D^{(k-1)}} \mathbb{I}\big((s_i,a_i)=(s,a)\big)}$$

$D^{(1)} \cup D^{(2)} \cup \dots \cup D^{(k-1)}$

LSVI that Reuses All Previous Samples

For $k = 1,\ 2, \ldots$

Obtain n samples $\mathcal{D}^{(k)} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{n}$ where $\mathbb{E}[r_i] = R(s_i, a_i),\ s_i' \sim P(\cdot\,|s_i, a_i)$

Perform **regression** on $\mathcal{D}^{(1)} \cup \mathcal{D}^{(2)} \cup \cdots \cup \mathcal{D}^{(k)}$ to find $Q^{(k)}$ such that

$$Q^{(k)}(s,a) \approx R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} Q^{(k-1)}(s',a')\right]$$

In practice, we reuse "recent" data but not all previous data (discussed later).

Analysis of LSVI under Certain Assumptions

To theoretically show that LSVI converges to the optimal value function, we will make some assumptions to ensure the following holds for all iteration k:

$$Q^{(k)}(s,a) \approx R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q^{(k-1)}(s',a') \right]$$

Linear case:

$$\phi(s,a)^\top \theta_k \approx R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} \phi(s',a')^\top \theta_{k-1} \right]$$

Analysis of LSVI under Certain Assumptions

$d = S \cdot A$

$\phi(s,a) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow (s,a)\text{-th entry}$

1. Bellman Completeness Assumption: For any $\theta \in \mathbb{R}^d$, there exists a $\theta' \in \mathbb{R}^d$ such that

$$\phi(s,a)^\top \theta' = R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} \phi(s',a')^\top \theta\right] \qquad \forall s,a$$

This ensures that no matter what θ_{k-1} is, there always exists a $\theta_k^\star$ such that

$\forall s,a \quad \theta^\star_{k,s,a} \leftarrow \boxed{R(s,a) + \cdots}$

$$\underbrace{\phi(s,a)^\top}_{\text{one-hot at }(s,a)\text{ entry}} \theta_k^\star = R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} \phi(s',a')^\top \theta_{k-1}\right]$$

$\boxed{\theta^\star_{k,s,a}}$

This is similar to the linear assumption $\phi(s,a)^\top \theta^\star = R(s,a)$ in contextual bandits, but is qualitatively stronger because the assumption require "for any θ".

Analysis of LSVI under Certain Assumptions

$D^{(1)} \cup \dots \cup D^{(K)}$

2. Coverage Assumption: The dataset $\cancel{\mathcal{D}^{(k)}}$ collected up to k-th iteration allows us to find θ_k so that for any s, a,

$$\left|\phi(s,a)^\top \theta_k - \phi(s,a)^\top \theta_k^\star\right| \le \epsilon_{\text{stat}}$$

(Similar to linear contextual bandits analysis) With

$$\theta_k = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \Big(\phi_i^\top \theta - \underbrace{\big(r_i + \gamma \max_{a'} \phi(s_i', a')^\top \theta_{k-1}\big)}_{\text{Expectation} = \phi_i^\top \theta_k^\star}\Big)^2 + \lambda \|\theta\|^2$$

we have $\left|\phi(s,a)^\top(\theta_k - \theta_k^\star)\right| \lesssim \sqrt{\beta}\|\phi(s,a)\|_{\Lambda^{-1}}$ where $\Lambda = \lambda I + \sum_{i=1}^{n} \phi_i \phi_i^\top$

In linear CB, we did not make such an assumption. What we did there is adding $\sqrt{\beta}\|\phi(s,a)\|_{\Lambda^{-1}}$ as **exploration bonus**, which encourages exploration and aims to make $\sqrt{\beta}\|\phi(s,a)\|_{\Lambda^{-1}}$ small for all s, a.

Analysis of LSVI under Certain Assumptions (Recap)

1. Bellman Completeness (i.e., function approximation is sufficiently expressive)

$$\forall \theta_{k-1}, \exists \theta_k^\star \qquad \phi(s,a)^\top \theta_k^\star = R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} \phi(s',a')^\top \theta_{k-1} \right] \quad \forall s, a$$

$$\left(\forall \theta_{k-1}, \exists \theta_k^\star \qquad Q_{\theta_k^\star}(s,a) = R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q_{\theta_{k-1}}(s',a') \right] \quad \forall s, a \right)$$

2. Coverage Assumption (i.e., the collected data is sufficient and explores the state-action space) Regression over $\mathcal{D}^{(k)}$ allows us to find θ_k such that

$$\left| \phi(s,a)^\top \theta_k - \phi(s,a)^\top \theta_k^\star \right| \le \epsilon_{\text{stat}} \quad \forall s, a$$

$$\left(\left| Q_{\theta_k}(s,a) - Q_{\theta_k^\star}(s,a) \right| \le \epsilon_{\text{stat}} \quad \forall s, a \right)$$

The two assumptions jointly imply $Q_{\theta_k}(s,a) \approx R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q_{\theta_{k-1}}(s,a) \right]$

Analysis of LSVI under Certain Assumptions

Under Bellman completeness and coverage assumptions, LSVI ensures

$$\left\|Q^{(k)} - Q^\star\right\|_\infty \leq O\left(\gamma^k \left\|Q^{(0)} - Q^\star\right\|_\infty + \frac{\epsilon_{\text{stat}}}{1-\gamma}\right)$$

where $\left\|Q^{(k)} - Q^\star\right\|_\infty \coloneqq \max_{s,a} \left|Q^{(k)}(s,a) - Q^\star(s,a)\right|$

Also, the greedy policy $\pi^{(k)}(s) = \operatorname{argmax}_a Q^{(k)}(s,a)$ satisfies for all s,

$$V^\star(s) - V^{\pi^{(k)}}(s) \leq O\left(\gamma^k \left\|Q^{(0)} - Q^\star\right\|_\infty + \frac{\epsilon_{\text{stat}}}{1-\gamma}\right)$$

$$\left|Q^{(k)}(s,a) - Q^*(s,a)\right| \le \left|\underline{r(s,a) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a)}\left(\max_{a'} Q^{(k-1)}(s',a')\right)} - \underbrace{r(s,a) - \gamma \mathbb{E}_{s'\sim p(\cdot|s,a)}\left(\max_{a'} Q^*(s',a')\right)}_{-Q^*(s,a)}\right| + \varepsilon_{stat} \qquad Q^{(k)}(s,a) = \phi(s,a)^T\theta_k$$

Assumption 2 : $\left|Q^{(k)}(s,a) - r(s,a) - \gamma \mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\max_{a'} Q^{(k-1)}(s',a')\right]\right| \le \varepsilon_{stat}$

Bellman opt. eq. $Q^*(s,a) - r(s,a) - \gamma \mathbb{E}_{s'\sim p(\cdot|s,a)}\left(\max_{a'} Q^*(s',a')\right) = 0$

$$\le \gamma \left|\mathbb{E}_{s'\sim p(\cdot|s,a)}\left(\max_{a'} Q^{(k-1)}(s',a') - \max_{a'} Q^*(s',a')\right)\right| + \varepsilon_{stat}$$

$$\le \gamma \left|\mathbb{E}_{s'\sim p(\cdot|s,a)} \max_{a'} \left|Q^{(k-1)}(s',a') - Q^*(s',a')\right|\right| + \varepsilon_{stat}$$

$$\left|\max_a f(a) - \max_a g(a)\right| \le \max_a |f(a) - g(a)|$$

$$\le \gamma \max_{s',a'} \left|Q^{(k-1)}(s',a') - Q^*(s',a')\right| + \varepsilon_{stat}$$

$$\Rightarrow \max_{s,a} \left|Q^{(k)}(s,a) - Q^*(s,a)\right| \le \gamma \max_{s,a}\left|Q^{(k-1)}(s,a) - Q^*(s,a)\right| + \varepsilon_{stat}$$

$$\le \gamma\left(\gamma \max_{s,a}\left|Q^{(k-2)}(s,a) - Q^*(s,a)\right| + \varepsilon_{stat}\right) + \varepsilon_{stat}$$

$$\le \dots \le \gamma^k \max_{s,a}\left|Q^{(0)}(s,a) - Q^*(s,a)\right| + \varepsilon_{stat}\left(1 + \gamma + \gamma^2 + \dots \gamma^k\right) \le \frac{1}{1-\gamma}$$

Notes on Exploration in MDPs

The Coverage Assumption

$$\left|\phi(s,a)^\top \theta_k - \phi(s,a)^\top \theta_k^\star\right| \le \epsilon_{\text{stat}} \quad \forall s, a$$

θ_k: our regression solution

$\theta_k^\star$: ground truth

- Requires the state-action space to be explored
  - **Tabular case**: every state-action pair needs to be visited many times
  - **Linear case**: the feature space $\{\phi(s,a)\}_{s,a}$ needs to be explored in all directions

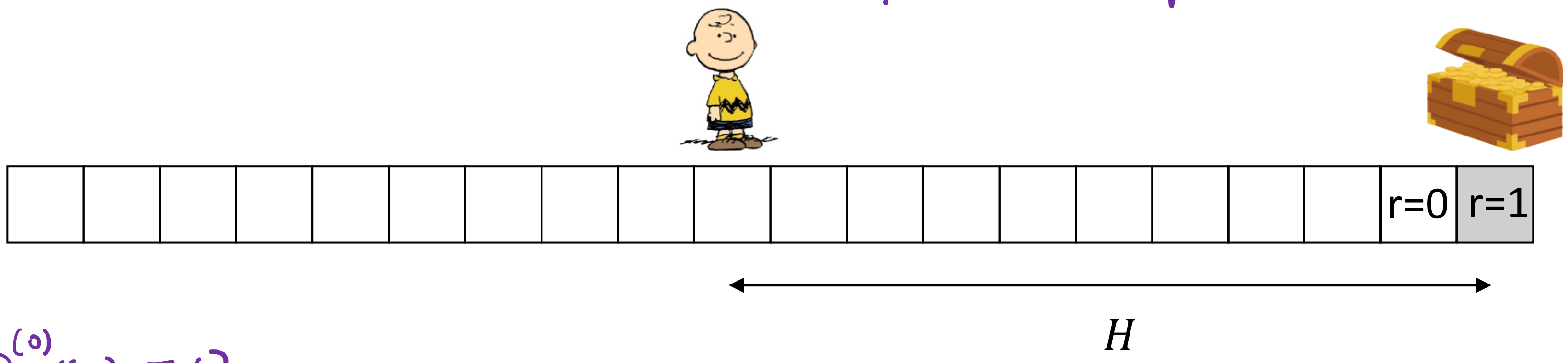
- In bandits, we focus on "action-space" exploration
  - Exploration bonus (UCB, Thompson Sampling) $a_t = \arg\max_a \{ \underline{\hat{R}(a)} + b_t(a) \}$
  - Randomization (ϵ-greedy, Boltzmann exploration, inverse-gap weighting)

$p_t(a) \propto \exp(\lambda \hat{R}(a))$

- In MDPs, we further need "state-space" exploration

$$\begin{cases} a_1: \text{go right} \\ a_2: \text{go left} \end{cases}$$

Each episode has H steps to execute



$Q^{(0)}(s,a) = 0$

If we do randomized exploration e.g. $p_t(a) \propto \exp(\lambda Q^{(k)}(s,a))$ $\longrightarrow$ Prob(reaching the r=1 state) $\simeq \frac{1}{2^H}$

ε-greedy

\# episodes needed to see signal $\gtrsim 2^H$

Removing the Coverage Assumption

Use exploration bonus in LSVI:

Tabular Case: $\tilde{R}(s,a) = \hat{R}(s,a) + \frac{\text{const}}{\sqrt{n(s,a)}}$

Linear MDP (a class of MDPs that satisfies linear Bellman completeness)**:**

$\tilde{R}(s,a) = \phi(s,a)^\top \hat{\theta} + \text{const}\, \|\phi(s,a)\|_{\Lambda^{-1}}$ where $\Lambda = I + \sum_{i=1}^{t-1} \phi(s_i,a_i)\phi(s_i,a_i)^\top$

UCB in tabular MDP: Minimax regret bounds for reinforcement learning. 2017.

UCB in linear MDP: Provably efficient reinforcement learning with linear function approximation. 2019.

TS in tabular MDP: Near-optimal randomized exploration for tabular Markov decision processes. 2021.

TS in linear MDP: Frequentist regret bounds for randomized least-squares value iteration. 2020.

Summary for LSVI

Env $\xrightarrow{s}$ / $\xleftarrow{a}$ / $\xrightarrow{r}$

Exploration Mechanism

Value Iteration + Regression

Value Iteration + Regression

$\mathcal{D}^{(1)} = \{(s,a,r,s')\}$ $\quad \mathcal{D}^{(2)}$ $\quad \cdots \quad \mathcal{D}^{(k-1)}$ $\quad \cdots$

$Q^{(1)} \quad Q^{(2)} \quad Q^{(3)} \quad Q^{(k-1)} \quad Q^{(k)}$

$(\mathcal{D}^{(k-1)} \cup \dots \cup \mathcal{D}^{(1)}),\ Q^{(k-1)}$

$$\theta_k = \underset{\theta}{\operatorname{argmin}} \sum_{(s_i, a_i, r_i, s_i') \sim \mathcal{D}} \left(Q_\theta(s_i, a_i) - r_i - \gamma \max_{a'} Q_{\theta_{k-1}}(s_i', a') \right)^2$$

not reuse sample (use $\mathcal{D}^{(k-1)}$) or
reuse sample (use $\mathcal{D}^{(1)} \cup \cdots \cup \mathcal{D}^{(k-1)}$)

cf. Contextual bandits (only regression)

$$\theta_k = \underset{\theta}{\operatorname{argmin}} \sum_{(x_i, a_i, r_i)} \left(R_\theta(x_i, a_i) - r_i \right)^2$$

Summary for LSVI

Exploration Mechanism

1. Randomized policies (ϵ-Greedy, Boltzmann exploration, inverse-gap weighting)
   – usually used in practice

2. Exploration bonus (UCB) / Randomized values (TS)
   – can give rigorous regret bounds for tabular MDPs and MDPs with linear Bellman completeness

Other Names for LSVI

- Fitted Q Iteration (FQI)
- Least Square Q Iteration (LSQI)

Q-Learning

Q-Learning (Watkins, 1992)

For $i = 1,\ 2, \ldots$

Obtain sample (s_i, a_i, r_i, s_i')

$$Q^{(i)}(s_i, a_i) \leftarrow (1-\alpha)Q^{(i-1)}(s_i, a_i) + \alpha\left(r_i + \gamma \max_a Q^{(i-1)}(s_i', a)\right)$$

$$Q^{(i)}(s, a) \leftarrow Q^{(i-1)}(s, a) \qquad \forall (s, a) \neq (s_i, a_i)$$

cf. LSVI:

$$\forall s, a, \qquad Q^{(k)}(s, a) = \frac{\sum_{i=1}^{n_k} \mathbb{I}\{(s_i, a_i) = (s, a)\}\left(r_i + \gamma \max_{a'} Q^{(k-1)}(s_i', a')\right)}{\sum_{i=1}^{n_k} \mathbb{I}\{(s_i, a_i) = (s, a)\}}$$

Q-Learning (Watkins, 1992)

Watkin's Q-Learning + Linear Function Approximation

For $i = 1,\ 2, \ldots$

Obtain sample (s_i, a_i, r_i, s_i')

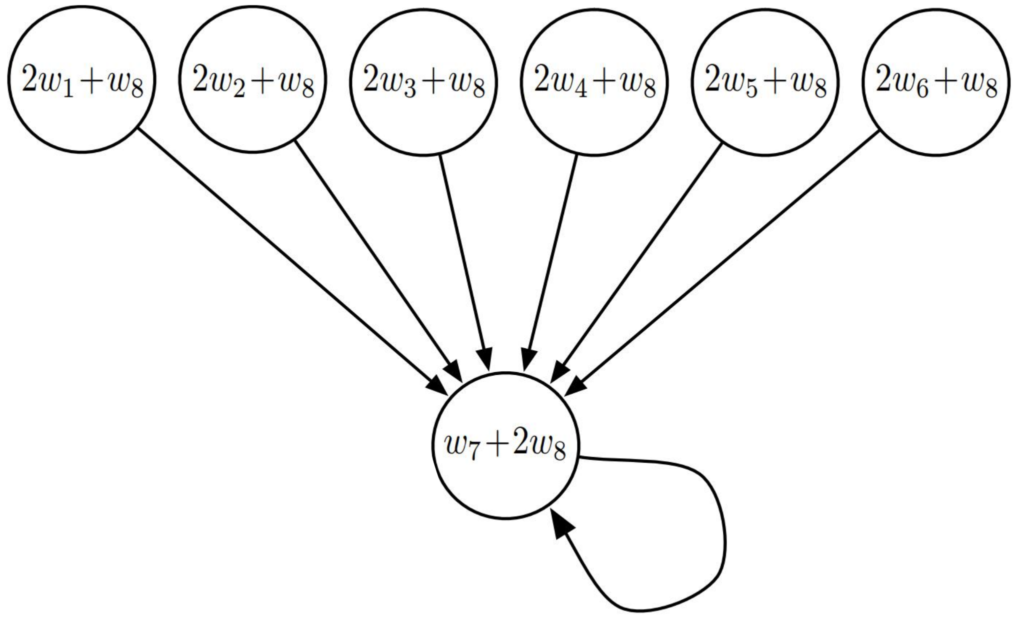
$$\theta_i \leftarrow \theta_{i-1} - \alpha \nabla_\theta \left(\phi(s_i, a_i)^\top \theta - r_i - \gamma \max_a \phi(s_i', a)^\top \theta_{i-1} \right)^2 \Bigg|_{\theta = \theta_{i-1}}$$

$$= \theta_{i-1} - 2\alpha \left(\phi(s_i, a_i)^\top \theta_{i-1} - r_i - \gamma \max_a \phi(s_i', a)^\top \theta_{i-1} \right) \phi(s_i, a_i)$$

c.f. LSVI: $\theta_k = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n_k} \left(\phi(s_i, a_i)^\top \theta - r_i - \gamma \max_{a'} \phi(s_i', a')^\top \theta_{k-1} \right)^2$

Watkin’s Q-Learning + LFA Does Not Converge

Even when Bellman completeness and coverage assumption hold



Baird’s example