Approximate Value Iteration and Variants

Chen-Yu Wei

Value Iteration

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

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 $k = 1,\ 2, \ldots$

We want these samples to be "exploratory"

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

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

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

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

Comparison with Contextual Bandits

Env $\xrightarrow{x_t}$ / $\xleftarrow{a_t}$ / $\xrightarrow{r_t}$

Exploration

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

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

$$\ldots$$

Regression

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

Env $\xrightarrow{s_t}$ / $\xleftarrow{a_t}$ / $\xrightarrow{r_t}$

Exploration

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

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

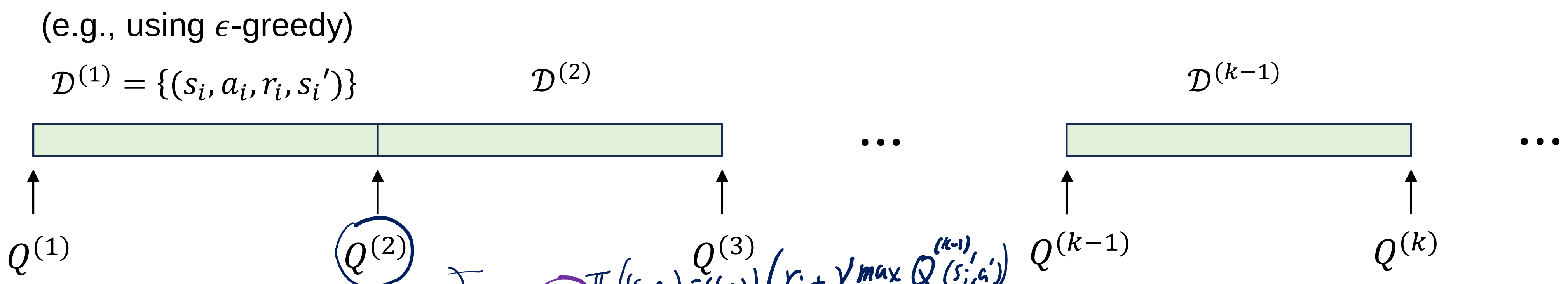
$$\ldots$$

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



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

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

LSVI that Reuses All Previous Samples

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

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

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

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

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

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

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

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,

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

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

$$\underline{Q^{(k)}(s,a) = \phi(s,a)^T \theta_k}$$

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

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

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

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

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

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

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

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

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

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)
  - Randomization (ϵ-greedy, Boltzmann exploration, inverse-gap weighting)

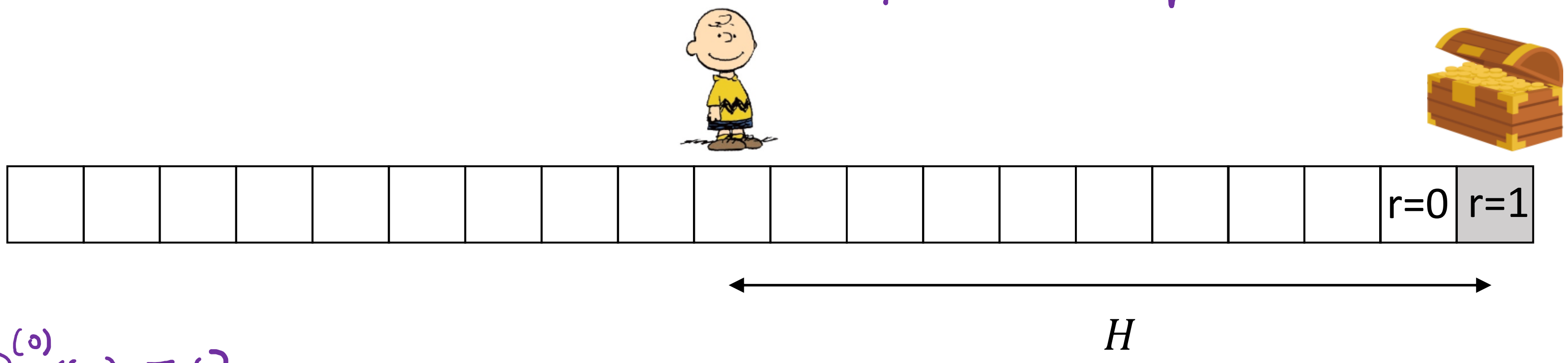
$a_t = \arg\max_a \{\underline{\hat{R}(a)} + b_t(a)\}$

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

- 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.

Exploration bonus for general function approximation (deep learning):

Unifying Count-Based Exploration and Intrinsic Motivation

Curiosity-driven Exploration by Self-supervised Prediction

Exploration by Random Network Distillation

Summary for LSVI

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

Exploration Mechanism

Value Iteration + Regression

Value Iteration + Regression

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

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

$\left(\mathcal{D}^{(k-1)} \cup \cdots \cup \mathcal{D}^{(1)}\right),\ Q^{(k-1)}$

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

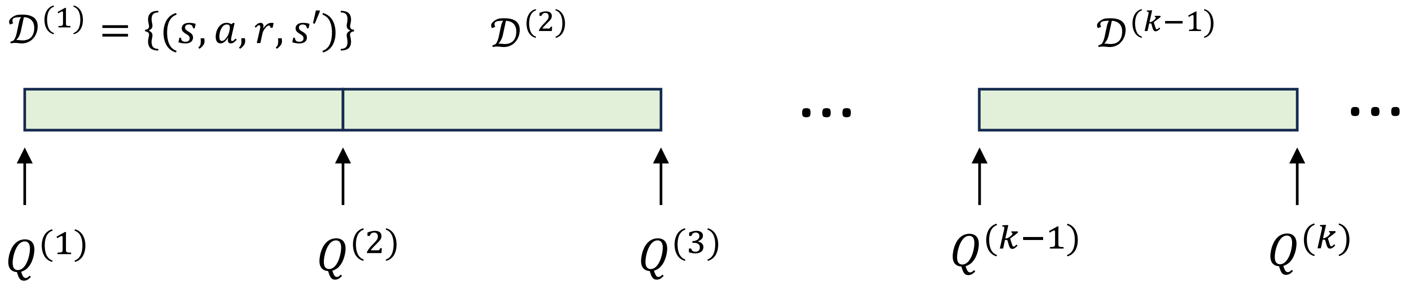
not reuse sample (use $\mathcal{D}^{(k-1)}$) or
reuse sample (use $\mathcal{D}^{(1)} \cup \cdots \cup \mathcal{D}^{(k-1)}$)

cf. Contextual bandits (only regression)

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

Summary for LSVI

Value Iteration + Regression



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

Coverage assumption $\Rightarrow$ $\forall s, a, \quad \left|Q_{\theta_k}(s,a) - Q_{\theta_k^\star}(s,a)\right| \leq \epsilon_{\text{stat}}$
(exploration assumption)

Summary for LSVI

Exploration Mechanism

1. Randomized policies (ϵ-Greedy, Boltzmann exploration, inverse-gap weighting)
   – perform local exploration

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

Other names for LSVI: Fitted Q Iteration, Least-square Q Iteration

Q-Learning

Q-Learning (Watkins, 1992)

$$\hat{R}^{(i)}(a) = (1-\alpha)\hat{R}^{(i-1)}(a) + \alpha\, r_i(a)$$
$$= (1-\alpha)\left((1-\alpha)\hat{R}^{(i-2)}(a) + \alpha r_{i-1}(a)\right) + \alpha r_i(a)$$
$$\Rightarrow \hat{R}^{(i)}(a) = \sum_{j=1}^{i} \alpha(1-\alpha)^{i-j} r_j(a)$$

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

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

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

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

Function approximation : $Q_\theta(s,a)$

cf. LSVI:

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

Q-Learning (Watkins, 1992)

Fixed an (s,a). Let's see what $\underline{Q^{(k)}(s,a)}$

Assume that before iteration k, (s,a) has been visited in iteration $j_1, j_2, \cdots, j_\tau < k$

$$Q^{(k)}(s,a) = \sum_{i=1}^{\tau} \alpha(1-\alpha)^{\tau-i} \left(\underset{\substack{\uparrow \\ R(s,a)}}{r_{j_i}} + \gamma \max_{a'} Q^{(j_i)}(s'_{j_i}, a') \right)$$

Q-Learning (Watkins, 1992)

Suppose that $\alpha_i = \frac{1}{i^\beta}$ for some $\frac{1}{2} < \beta \leq 1$, and every state-action pair is visited infinitely often. Then

$$Q^{(i)}(s,a) \to Q^\star(s,a) \quad \forall s,a.$$

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, Yuxin Chen. Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction. 2020.

Watkins's Q-Learning + Linear Function Approximation

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

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

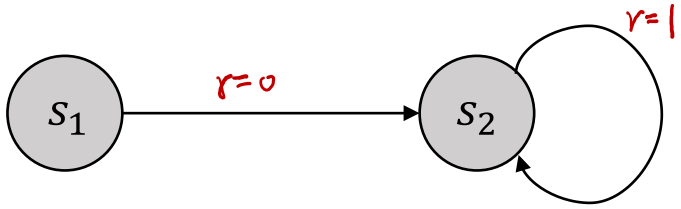
$$\theta_i \leftarrow \theta_{i-1} - \alpha \nabla_\theta \left(\phi(s_i, a_i)^\top \theta - r_i - \gamma \max_a \phi(s_i', a)^\top \theta_{i-1} \right)^2 \Bigg|_{\theta = \theta_{i-1}}$$

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

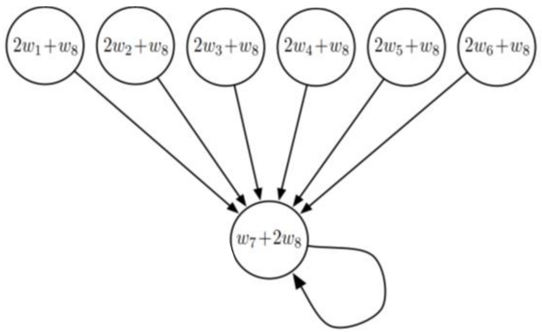
c.f. LSVI: $$\theta_k = \operatorname*{argmin}_\theta \sum_{i=1}^{n_k} \left(\underbrace{\phi(s_i, a_i)^\top \theta}_{Q_\theta(s_i, a_i)} - r_i - \gamma \max_{a'} \phi(s_i', a')^\top \theta_{k-1} \right)^2$$

Watkins's Q-Learning + LFA Does Not Converge

Even when Bellman completeness and coverage assumptions hold



$\phi(s_1, a) = (1,0)$ $\phi(s_2, a) = (2,1)$



Simplified from the "Baird's counterexample"
(see Sutton and Barto Section 11.2)

Bellman completeness assumption

For any $\theta' \in \mathbb{R}^2$, there exists a $\theta \in \mathbb{R}^2$ such that

$$\phi(s,a)^T \theta = R(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} \phi(s',a')^T \theta' \right] \quad \forall s,a.$$

$$\begin{cases} \phi(s_1,a)^T \theta = R(s_1,a) + \gamma \phi(s_2,a)^T \theta' \\ \phi(s_2,a)\, \theta = R(s_2,a) + \gamma \phi(s_2,a)^T \theta' \end{cases}$$

two variables (θ_1, θ_2) with two linearly independent constraints

$Kn = 100000$

The Effect of Fixing the Target

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

$\theta_{k-1} \leftarrow \theta$

when n is large $\Rightarrow \theta \approx \arg\min_{\theta} \left\{ \frac{1}{2}\left(\phi(s_1,a)^\top\theta - 1 - \gamma\phi(s_2,a)^\top\theta_{k-1}\right)^2 + \frac{1}{2}\left(\phi(s_2,a)^\top\theta - 0 - \gamma\phi(s_2,a)^\top\theta_{k-1}\right)^2 \right\}$

For $i = 1, \ldots, n$:

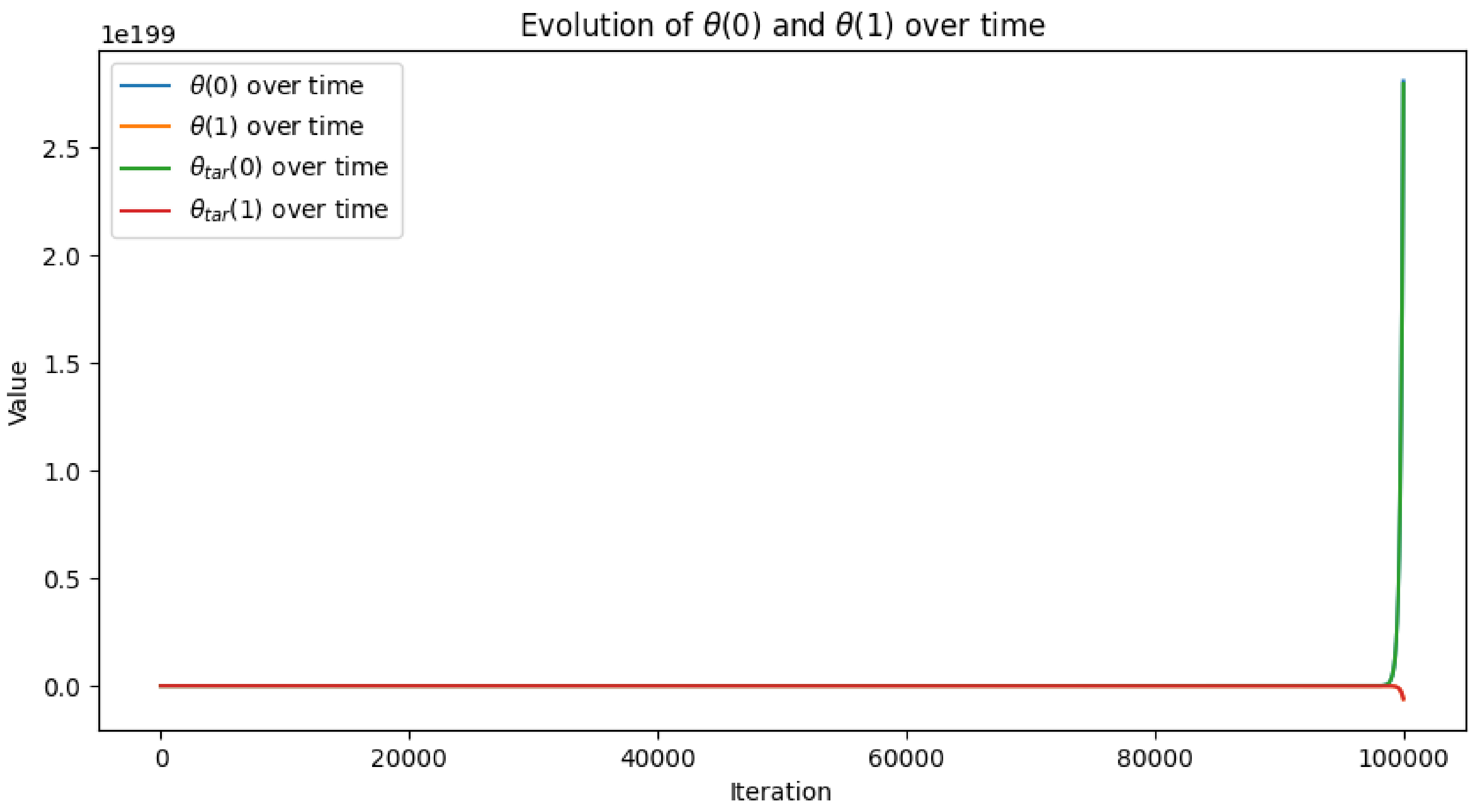
Sample $(s, a, r, s') \sim \mathrm{Uniform}\{(s_1, a, 1, s_2),\ (s_2, a, 0, s_2)\}$

$$\theta \leftarrow \theta - \alpha\left(\phi(s,a)^\top\theta - r - \gamma\phi\left(s',a\right)^\top\theta_{k-1}\right)\phi(s,a)$$

$\theta_k \leftarrow \theta$

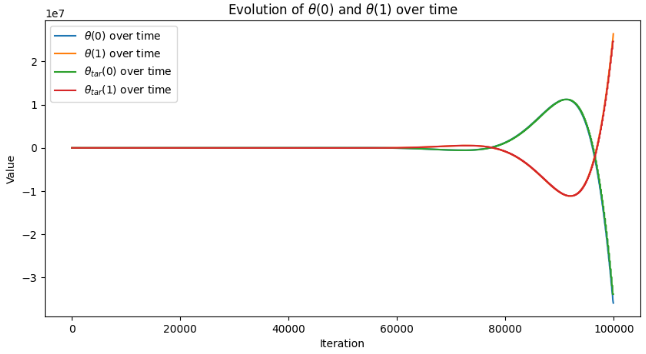
The Effect of Fixing the Target

$n=1$

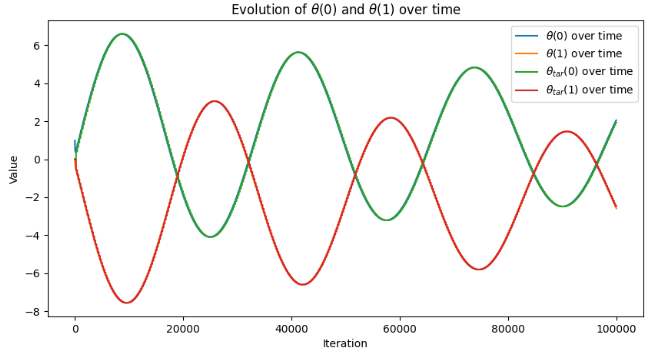


# of iterations in outer loop

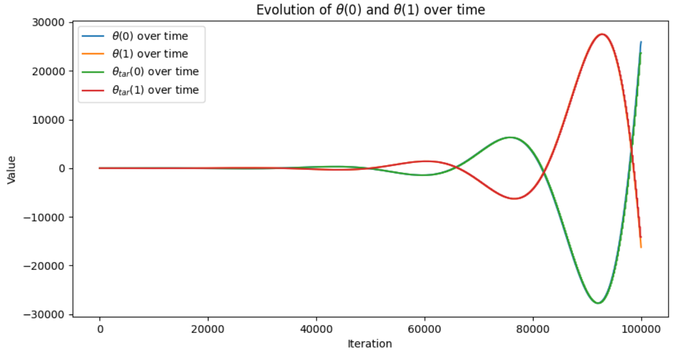
$$K = \frac{100000}{n}$$



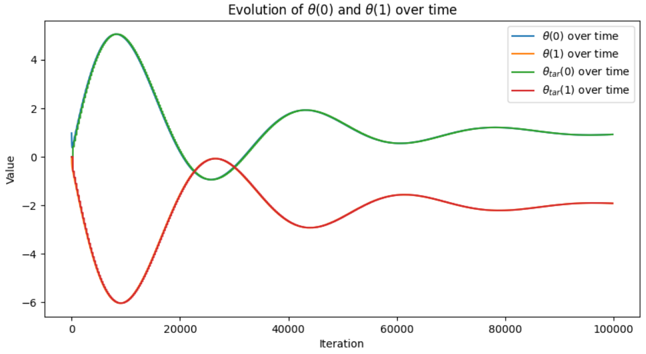
n=150



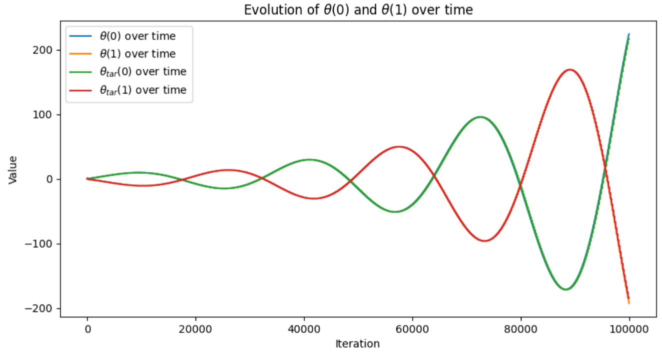
n=210



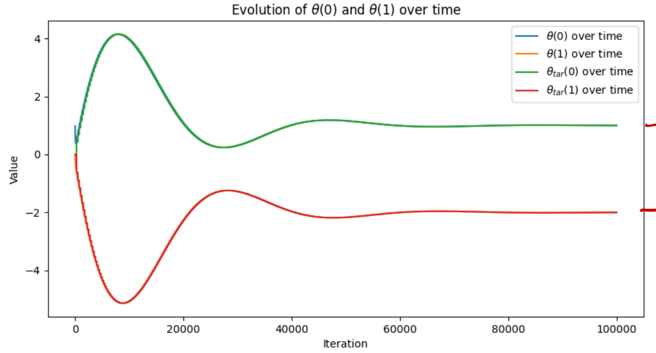
n=170



n=230



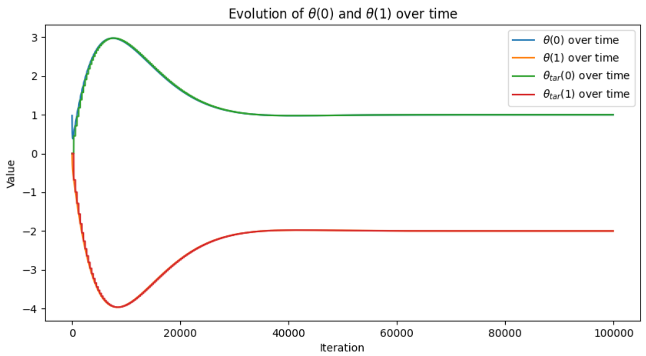
n=190



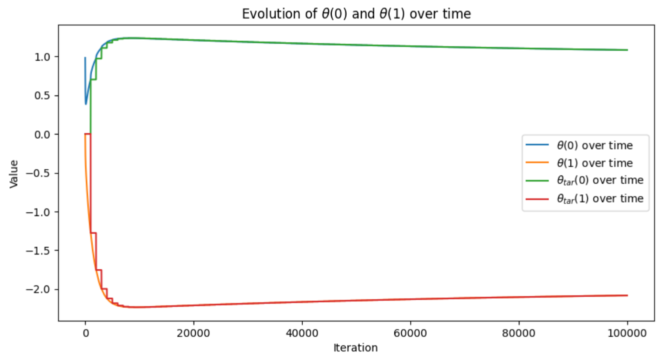
$\rightarrow 1$

$\rightarrow -2$

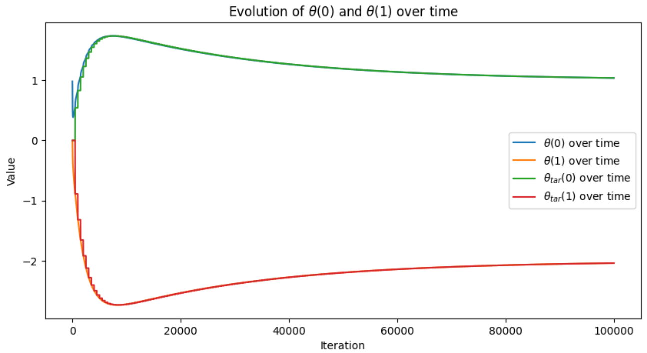
n=250



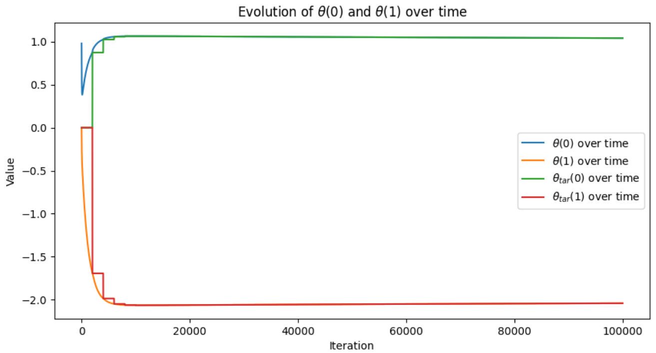
n=300



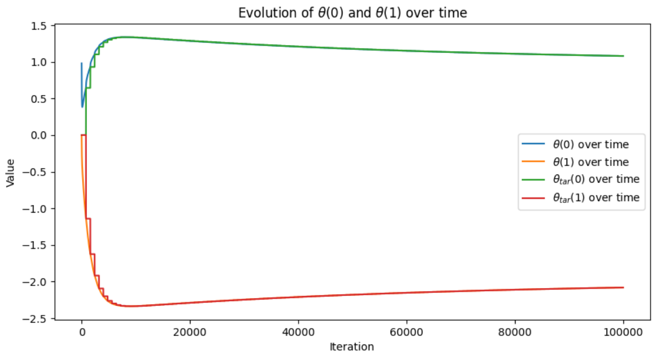
n=1000



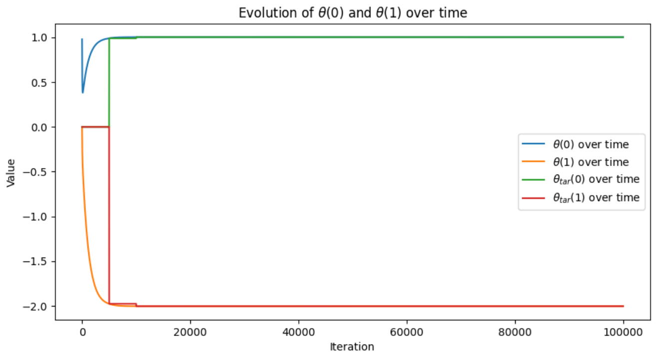
n=500



n=2000



n=800



n=5000

Watkins's Q-Learning vs. LSVI

Under coverage assumption
(i.e., the data $\{(s_i, a_i, r_i, s_i')\}$ sufficiently cover every state-action pair / feature space)

	LSVI	**Watkins's Q-Learning**
Convergence in the tabular case	$Q^{(k)} \to Q^\star$	$Q^{(k)} \to Q^\star$
Convergence under function approximation	$Q^{(k)} \to Q^\star$ under BC	Diverges even with BC
Update style	Two time-scale	Single time-scale

Techniques for Function Approximation (Deep Q-Learning)

Use LSVI Updates

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

Collect samples $\mathcal{D}^{(k)}$ (consisting of (s, a, r, s') tuples) using exploratory policy

Perform regression over dataset $\mathcal{D}^{(1)} \cup \mathcal{D}^{(2)} \cup \cdots \cup \mathcal{D}^{(k)}$:

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

Regression

Implement Regression with SGD

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

Collect samples $\mathcal{D}^{(k)}$ (consisting of (s, a, r, s') tuples) using exploratory policy

$\theta_{k-1} \leftarrow \theta$

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

Randomly draw a minibatch $\{(s_i, a_i, r_i, s_i')\}_{i=1}^{b}$ from $\mathcal{D}^{(1)} \cup \mathcal{D}^{(2)} \cup \cdots \cup \mathcal{D}^{(k)}$

$$\theta \leftarrow \theta - \alpha \sum_{i=1}^{b} \nabla_\theta \left(Q_\theta(s_i, a_i) - r_i - \gamma \max_{a'} Q_{\theta_{k-1}}(s_i', a') \right)^2$$

Typical Implementation of Deep Q-Learning

Interleaving data collection and SGD

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

Obtain a new sample (s, a, r, s') and insert it to a **replay buffer** $\mathcal{B}$

Randomly draw a minibatch $\{(s_i, a_i, r_i, s_i')\}_{i=1}^{b}$ from $\mathcal{B}$ and perform

$$\theta \leftarrow \theta - \alpha \sum_{i=1}^{b} \nabla_\theta \left(Q_\theta(s_i, a_i) - r_i - \gamma \max_{a'} Q_{\theta_{\text{tar}}}(s_i', a')\right)^2$$

// Option 1

If $i \bmod n = 0$:

$\theta_{\text{tar}} \leftarrow \theta$

// Option 2 $\tau = 0.999$

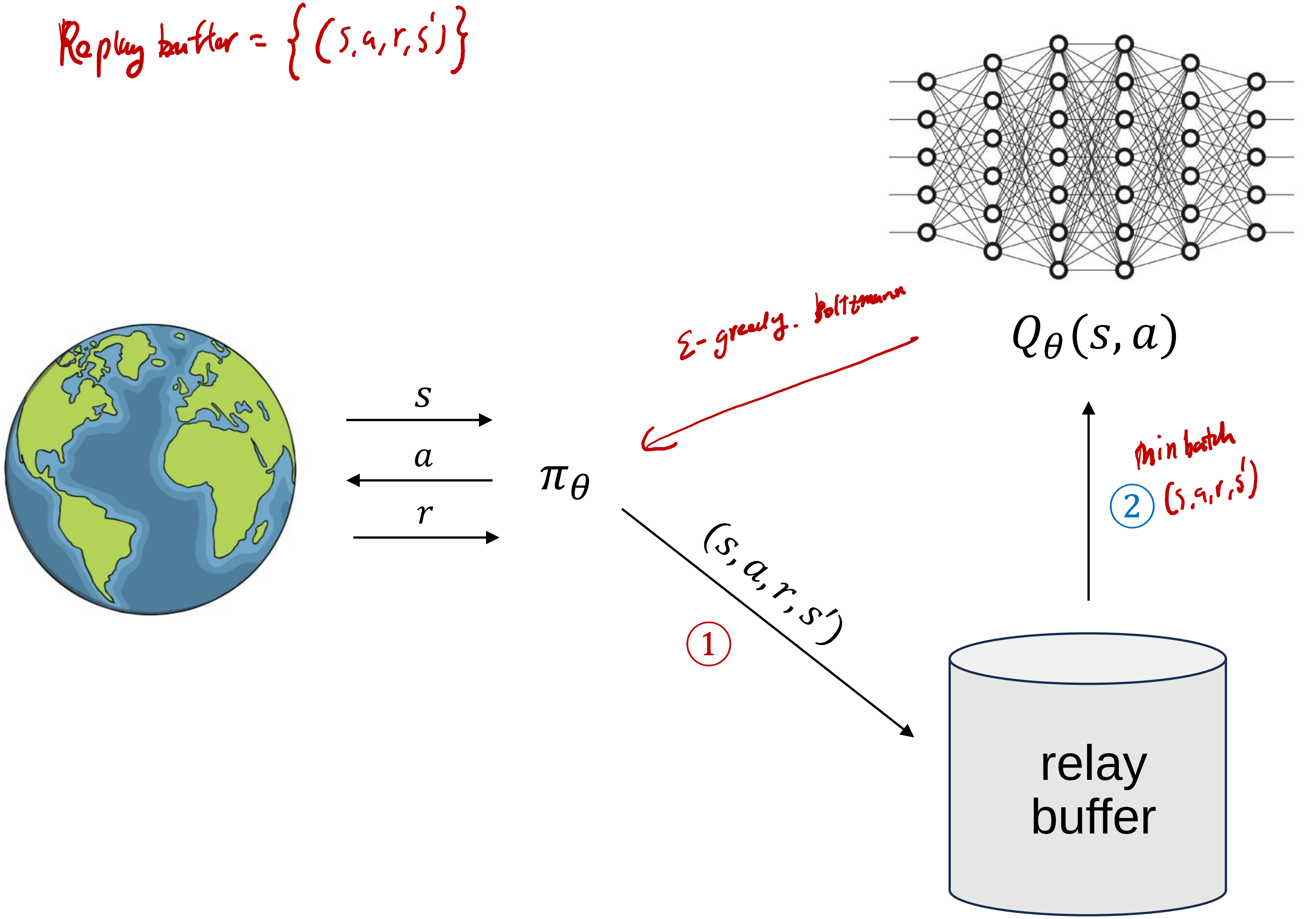
$\theta_{\text{tar}} \leftarrow \tau\theta_{\text{tar}} + (1 - \tau)\theta$

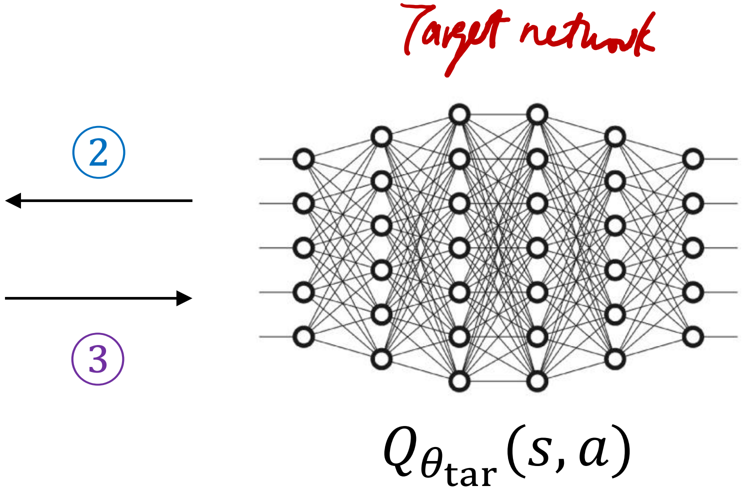
The following update converges but to the wrong solution when the transition is non-deterministic:

$$\theta \leftarrow \theta - \alpha \sum_{i=1}^{b} \nabla_\theta \left(Q_\theta(s_i, a_i) - r_i - \gamma \max_{a'} Q_\theta(s_i', a')\right)^2$$

See Sutton & Barto Section 11.5 or Nan Jiang's lecture note (P.17 bellman error minimization)

Target Network and Replay Buffer





1 collect new samples

2 perform SGD with fixed θ_{tar}

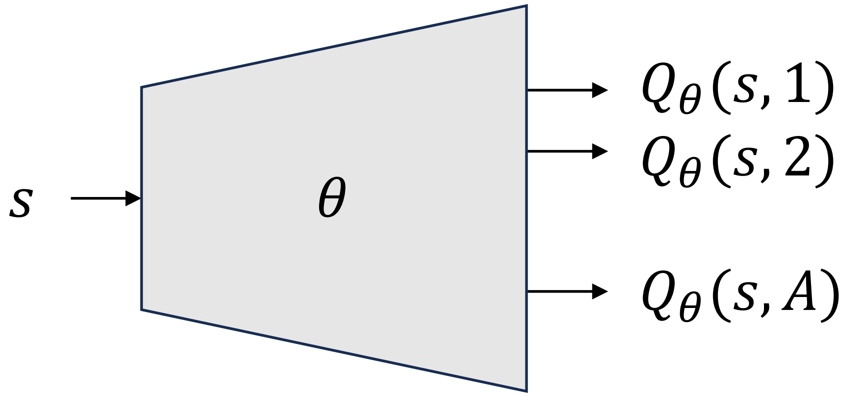
3 update θ_{tar}

Key: 3 is much slower or much more sporadically than 2

1 can be decoupled from 2 and 3

Q-Network Design

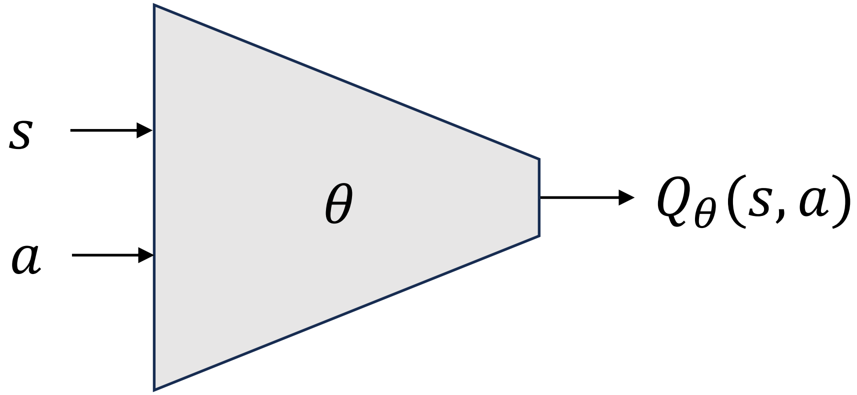
$Q_\theta(s,a)$



$\max_a Q_\theta(s,a)$ is straightforward

only handles discrete action case

Deep Q-Network



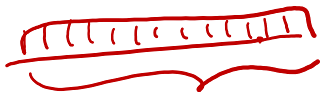
$\max_a Q_\theta(s,a)$ requires run $a \leftarrow a + \eta \frac{\partial\, Q_\theta(s,a)}{\partial a}$ iteratively

can handle continuous action case

Deep Deterministic Policy Gradient

(covered later in the semester)

Replay Buffer and Sampling



Standard implementation: First-in-first-out queue + Uniform sampling

- The data collected from π_θ is not i.i.d.
- Uniform sampling from a large pool makes the data more similar to i.i.d. – the convergence of SGD requires samples to be i.i.d.

Prioritized replay: priority queue + prioritized sampling + importance weight

- Priority queue with priority proportional to $|\delta_i|$, where $\delta_i = Q_\theta(s_i, a_i) - r_i - \gamma \max_{a'} Q_{\theta_{\text{tar}}}(s_i', a')$
- Sample from the buffer with probability $P_i \propto |\delta_i|^\alpha$
- Perform SGD with importance weight $w_i = \left(\frac{P_i}{\max_j P_j}\right)^{-\beta}$, i.e.,

$$\theta \leftarrow \theta - \alpha\, w_i \nabla_\theta \left(Q_\theta(s_i, a_i) - r_i - \gamma \max_{a'} Q_{\theta_{\text{tar}}}(s_i', a')\right)^2$$

Schaul, Quan, Antonoglou, Silver. Prioritized Experience Replay. 2015.

More on DQN

Recall Our Theoretical Analysis for LSVI

We made two assumptions:

- Bellman completeness (the expressiveness of function approximation)
- State-action space / feature space is sufficiently explored

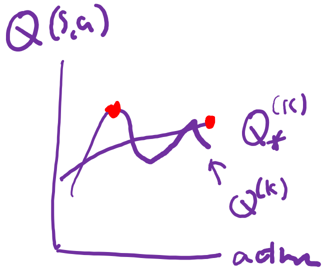
Then we argued that with these assumptions, we can ensure

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

However, these strong assumptions rarely hold.

$\boxed{\text{What happens if they do not hold?}}$

Over-estimation Bias



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

goal: to approx this

zero mean but has some variance

$$Q^{(k)}(s,a) = Q_*^{(k)}(s,a) + b(s,a)$$

next iteration

$$\mathbb{E}\left[\max_a Q^{(k)}(s,a)\right] = \mathbb{E}\left[\max_a \left\{ Q_*^{(k)}(s,a) + b(s,a) \right\}\right] \geq \max_a \mathbb{E}\left[Q_*^{(k)}(s,a) + b(s,a) \right] = \max_a Q_*^{(k)}(s,a)$$

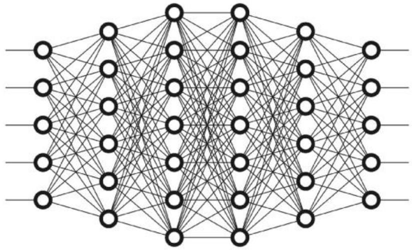
$$\mathbb{E}\left[\max_a f(a)\right] \geq \max_a \mathbb{E}\left[f(a)\right]$$

$$\max_{a'} Q^{(k-1)}(s',a')$$

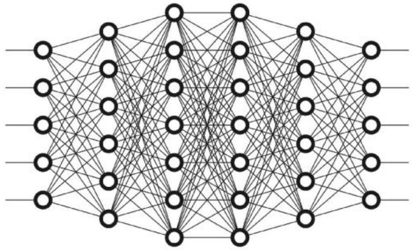
$$Q^{(k-1)}(s',\tilde{a})$$

$$\tilde{a} = \arg\max_{\tilde{a}} \tilde{Q}^{(k-1)}(s',\tilde{a})$$

Mitigating the over-estimation bias of DQN

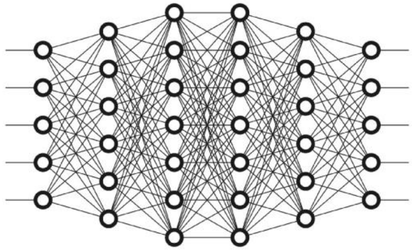


θ_1

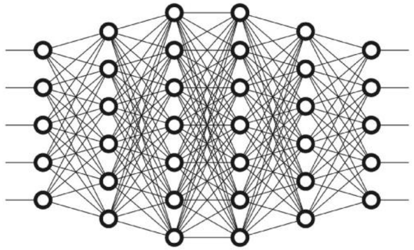


θ_{tar1}

$$\text{loss} = \left(Q_{\theta_1}(s,a) - r - \gamma \max_{a'} Q_{\theta_{\text{tar1}}}(s',a')\right)^2$$



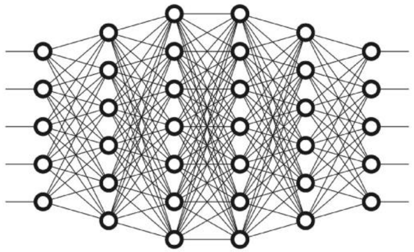
θ_2



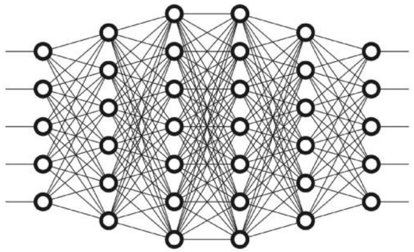
θ_{tar2}

$$\text{loss} = \left(Q_{\theta_2}(s,a) - r - \gamma \max_{a'} Q_{\theta_{\text{tar2}}}(s',a')\right)^2$$

Mitigating the over-estimation bias of DQN



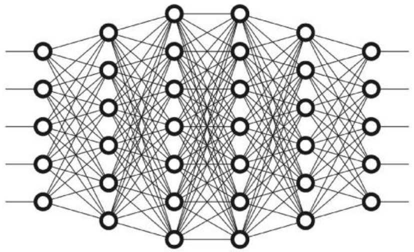
θ_1



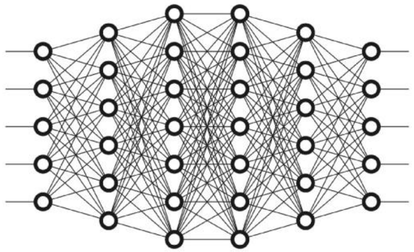
θ_{tar1}

$$\text{loss} = \left(Q_{\theta_1}(s,a) - r - \gamma \qquad Q_{\theta_{\text{tar1}}}(s', a')\right)^2$$

$$\underset{a'}{\text{argmax}}\, Q_{\theta_{\text{tar2}}}(s', a')$$



θ_2

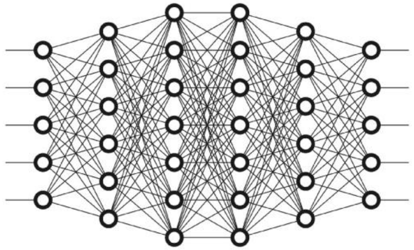


θ_{tar2}

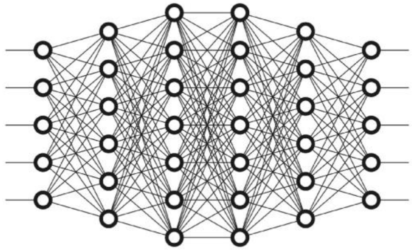
$$\text{loss} = \left(Q_{\theta_2}(s,a) - r - \gamma \qquad Q_{\theta_{\text{tar2}}}(s', a')\right)^2$$

$$\underset{a'}{\text{argmax}}\, Q_{\theta_{\text{tar1}}}(s', a')$$

A More Practical Solution



θ



θ_{tar}

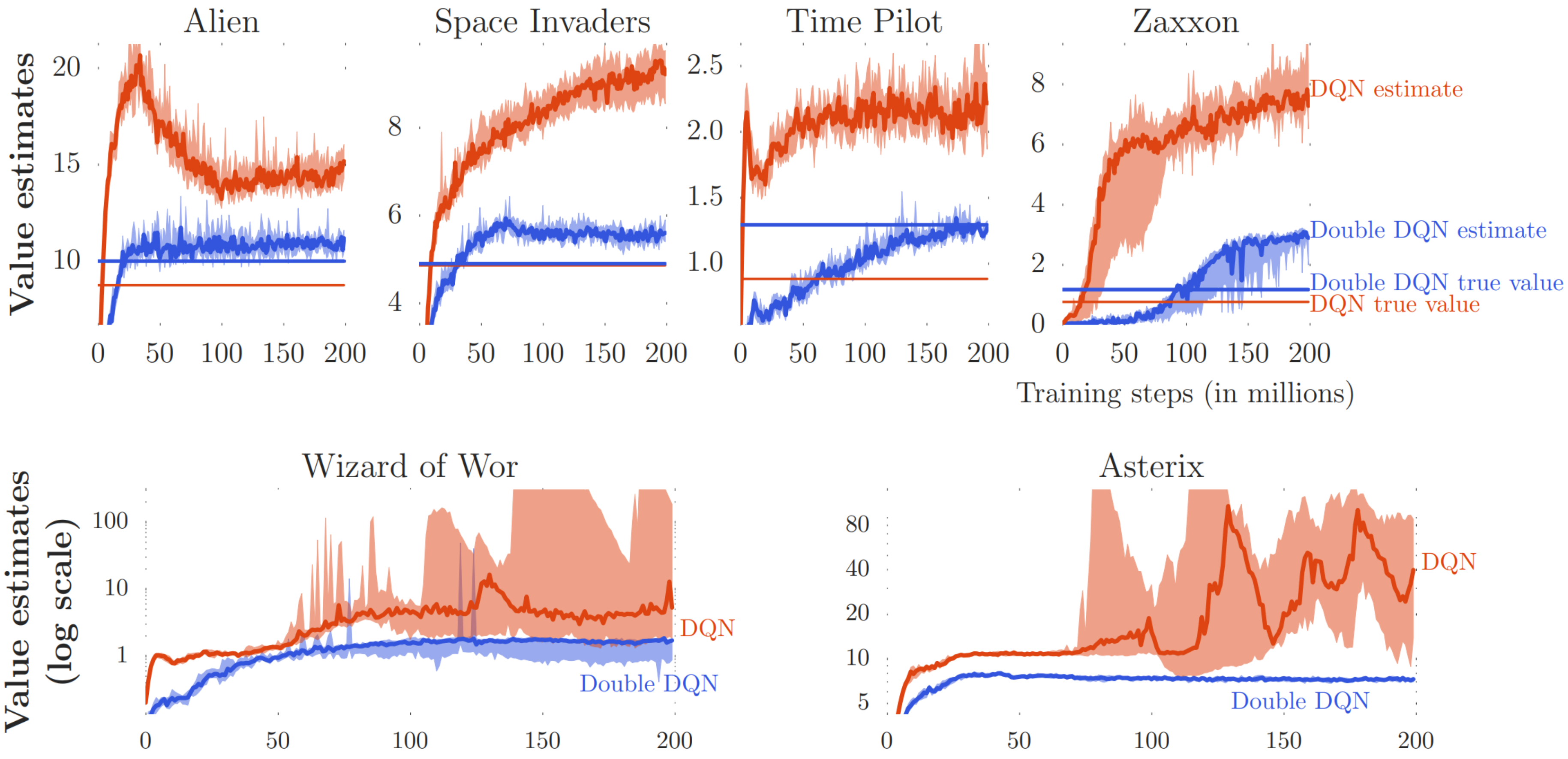
(s, a, r, s')

$$\text{loss} = \left(Q_{\theta}(s,a) - r - \gamma\, Q_{\theta_{\text{tar}}}(s', a')\right)^2$$

$$\underset{a'}{\operatorname{argmax}}\, Q_{\theta}(s', a')$$

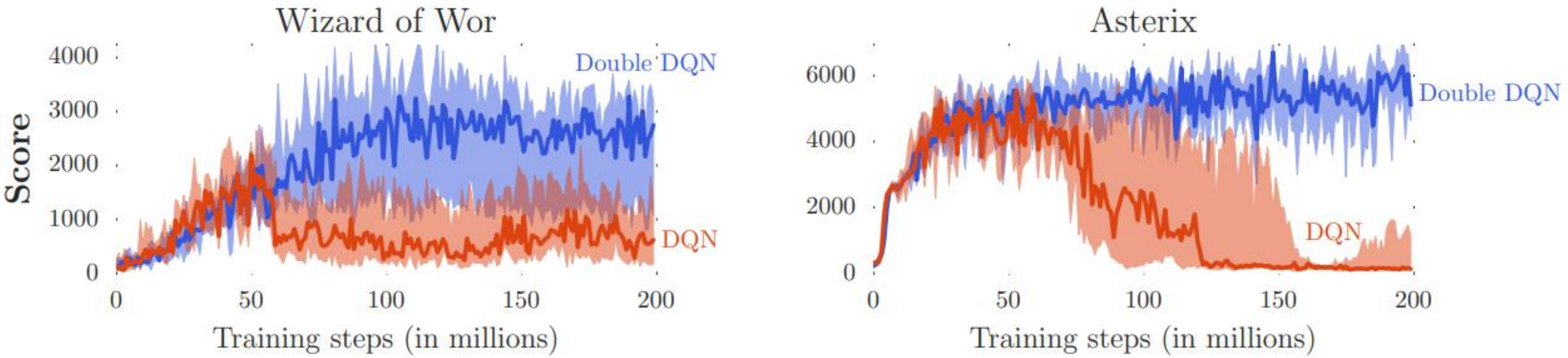
Double Deep Q-Network (DDQN)

DDQN mitigates over-estimation



Hado van Hasselt, Arthur Guez, David Silver. [Deep Reinforcement Learning with Double Q-learning](). 2015.

DDQN mitigates over-estimation



Hado van Hasselt, Arthur Guez, David Silver. [Deep Reinforcement Learning with Double Q-learning](). 2015.

Summary for Deep Q-Learning (1/3)

- Deep Q-learning is performing **approximate value iteration**
- Ideally, it would like generate $\theta_1, \theta_2, \ldots$ that approximates

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

- To successfully achieve this, we need
  - Sufficiently expressive function approximation (Bellman completeness)
  - Sufficient exploration over state-actions

Summary for Deep Q-Learning (2/3)

- There are two candidate updates

$$\theta_k = \underset{\theta}{\text{argmin}} \sum_{(s,a,r,s')} \left(Q_\theta(s,a) - r - \gamma \max_{a'} Q_{\theta_{k-1}}(s',a')\right)^2$$ Least-Square Value iteration

$$\theta_k = \theta_k - \alpha \nabla_\theta \left(Q_\theta(s,a) - r - \gamma \max_{a'} Q_{\theta_{k-1}}(s',a')\right)^2$$ Watkins's Q-Learning

  Only LSVI is stable under function approximation

- In order to implement LSVI, we use double-loop (double-time-scale) updates, where the **target network** is updated in a slow rate.
- When target network is fixed, the main network uses SGD to perform regression. We use **replay buffer + sampling** to reuse data and decorrelate samples.

Summary for Deep Q-Learning (3/3)

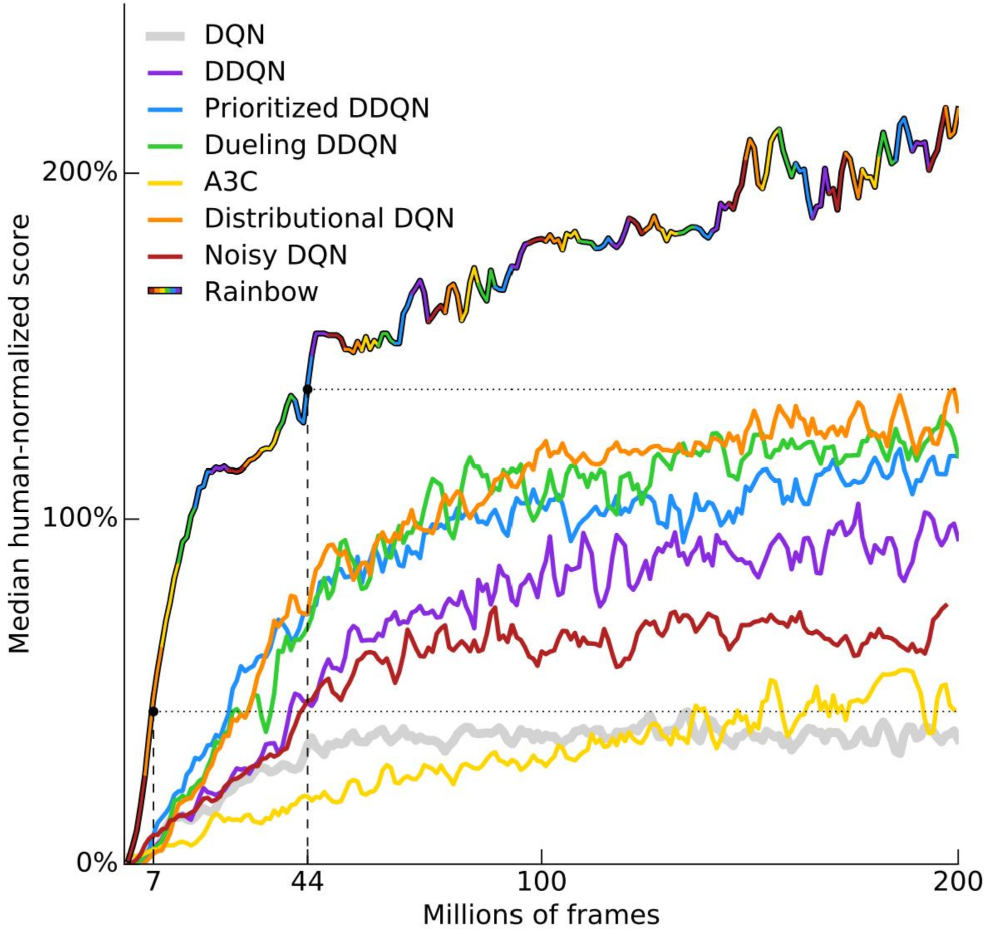
- When the idealized update

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

  is not perfect, there is over-estimation bias. We can use **double DQN** to mitigate the bias.

Combining More Techniques in DQN

Rainbow: Combining Improvements in Deep Reinforcement Learning. 2018.



A Remark on (Deep) Q-Learning in Episodic Settings

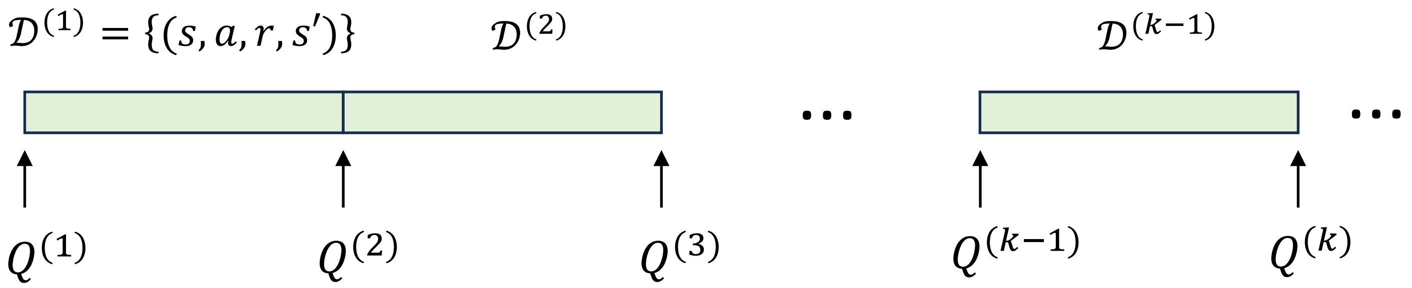
dummy state

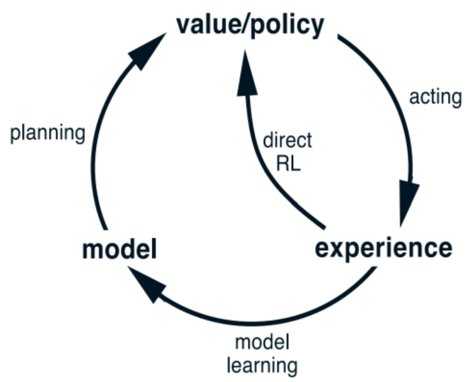
(s, a, r, s')

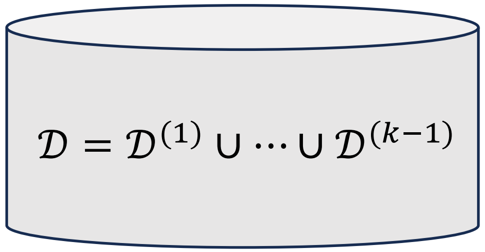
terminal states

$$\text{loss} = \Big(Q_\theta(s,a) - r(s,a) - \underbrace{\gamma \max_{a'} Q_{\theta_{tar}}(s', a')}_{= 0 \text{ for dummy state}} \Big)^2$$

A Remark on Model-Free vs. Model-Based Approaches

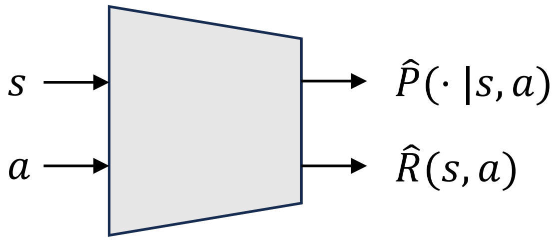






$$\theta_k \leftarrow \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - r - \gamma \max_{a'} Q_{\theta_{k-1}}(s',a')\right)^2\right]$$

Model-free



Trained with $\mathcal{D}$

$$\theta_k \leftarrow \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{(s,a)\sim\mathcal{D},\, r\sim\hat{R}(s,a),\, s'\sim\hat{P}(\cdot|s,a)}\left[\left(Q_\theta(s,a) - r - \gamma \max_{a'} Q_{\theta_{k-1}}(s',a')\right)^2\right]$$

or any other ways to find the optimal policy under $\hat{R}$ and $\hat{P}$

Model-based (Sutton & Barto Section 8)