Approximate Value Iteration and Variants

Chen-Yu Wei

Value Iteration

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

Find $Q^{(k)}$ such that

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

(regression)

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

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

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

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

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

$$\left|\phi(s,a)^\top \theta' - \left(R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a'} \phi(s',a')^\top \theta\right]\right)\right| \le \epsilon_{\mathrm{fa}}$$

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

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

This is similar to the linear assumption $\left|\phi(s,a)^\top \theta^\star - R(s,a)\right| \le \epsilon_{\mathrm{fa}}$ in contextual bandits, but is qualitatively stronger because the assumption require “for any θ”.

Analysis of LSVI under Certain Assumptions

2. Coverage Assumption: The dataset $\{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ collected in each iteration allows us to find θ_k so that for any s, a,

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

Recall from linear contextual bandit analysis, with

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

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 aims to make $\sqrt{\beta}\|\phi(s,a)\|_{\Lambda^{-1}}$ small for all s, a.

Analysis of LSVI under Certain Assumptions

Under approximate Bellman completeness and coverage assumptions, LSVI ensures

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

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,

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