Recap of value iteration, policy iteration, and approximations

Daniel Russo

April 12, 2020

Columbia University

Table of Contents

Table of Contents

On policy cost-to-go estimation

Approximating the cost-to-go function J_μ of a fixed policy μ.

- We are observe a Markovian sequence $(s_0, s_1, \cdots, s_n, \cdots)$ with stationary distribution π that is generated by applying policy μ.
- Focus on linear value approximation $J_\theta(s) = \phi(s)^\top \theta$.
  - Can write $J_\theta = \Phi\theta$ where $\Phi \in \mathbb{R}^{|\mathcal{S}| \times d}$ with rows $\Phi_s = \phi(s)^\top$.
- We compared two estimators:
  1. Monte-carlo estimators converge to the projected cost-to-go function $J_{\theta^{\mathrm{MC}}} = \Pi_{2,\pi} J^\mu$
  2. Temporal difference methods converge to the solution to a projected Bellman equation, $J_{\theta^{\mathrm{TD}}} = \Pi_{2,\pi} J_\mu J_{\theta^{\mathrm{TD}}}$.
     - The convergence of these methods and the existence of a fixed point relied on the linear architecture and on-policy sampling.

Table of Contents

Value iteration

Value iteration converges to the optimal cost-to-go function by repeatedly applying the Bellman operator

Value iteration: For $k = 1, 2, \cdots$, $J_{k+1} = TJ_k$, i.e.

$$J_{k+1}(s) = \min_{u \in U(s)} g(s, u) + \alpha \sum_{s' \in \mathcal{S}} P_{ss'}(u) J_k(s') \qquad \forall s \in \mathcal{S}.$$

One possible stopping rule is to stop once $\|J_{k+1} - J_k\|_\infty < \epsilon$. Asynchronous variants offer some advantages.

Fitted Value Iteration (FVI)

FVI uses regression based approximations to Bellman updates.

- Define the weighted norm $\|J\|_{2,\nu} = \sqrt{\mathbb{E}_{s\sim\nu}[J(s)^2]}$.
- Fitted value iteration is the scheme: for $k = 1, 2, \cdots$

$$\theta_{k+1} \in \operatorname*{argmin}_{\theta\in\Theta} \|J_\theta - TJ_{\theta_k}\|_{2,\nu}$$

- Equivalently, this can be viewed as a projected value iteration:

$$J_{k+1} = \Pi_{\mathcal{F},\nu} TJ_k$$

where $\mathcal{F} = \{J_\theta \mid \theta \in \Theta\}$ is the space of value functions approximations and $\Pi_{\mathcal{F},\nu} J = \operatorname{argmin}_{f\in\mathcal{F}} \|f - J\|_{2,\nu}$ projects onto $\mathcal{F}$ in a weighted norm.

Divergence of fitted value iteration

We'll focus on fitted value iteration

$$J_{k+1} = \Pi_{\mathcal{F},\nu} T J_k \quad k = 1, 2, \cdots$$

but the same steps apply for Q functions.

Bad news: ? gives an example where $J_k \to \infty$ with one dimensional linear function approximation.

Reason for non-convergence

T is a contraction in $\|\cdot\|_\infty$

But $\Pi_{\mathcal{F},\nu}$ could be an expansion in $\|\cdot\|_\infty$

- Consider $\mathcal{S} = \{1, 2\}$, $\nu = (.75, .25)$, $\mathcal{F} = \{(1, 2)\theta : \theta \in \mathbb{R}\}$
- Take $J = (2, 1)$. Then, solving

$$\operatorname*{argmin}_{\theta} .75(\theta - J(1))^2 + .25(2\theta - J(2)^2 = \frac{3.5}{2}$$

gives $\|\Pi_{\nu,\mathcal{F}} J\|_\infty = 3.5$.

Convergence of FVI with Low Inherent Bellman Error

Define the inherrent Bellman error of the function class:

$$\epsilon = \sup_{J \in \mathcal{F}} \inf_{\hat{J} \in \mathcal{F}} \|\hat{J} - TJ\|_\infty = \sup_{J \in T\mathcal{F}} \inf_{\hat{J} \in \mathcal{F}} \|\hat{J} - J\|_\infty$$

If $\epsilon = 0$, then $\mathcal{F}$ is *closed under Bellman updates*.

Define the approximate Bellman operator $\hat{T}J = \Pi_{\mathcal{F},\nu} TJ$.

Then,

$$\sup_{J \in \mathcal{F}} \|\hat{T}J - TJ\|_\infty = \epsilon$$

Your homework #6 shows

$$\limsup_{k \to \infty} \|J_k - J^*\|_\infty \le \frac{\epsilon}{1 - \alpha}.$$

Table of Contents

State-Aggregation (a.k.a state abstraction)

- We believe $J^*(s) \approx J^*(\tilde{s})$ if s and $\tilde{s}$ are "similar."
- More formally, take $\phi : \mathcal{S} \to \mathcal{S}$ to a be a mapping that associates $s \in \mathcal{S}$ with a representative state $\phi(s) \in \mathcal{S}$.
  - Simple case is $\mathcal{S} = [0, 1]$ and ϕ maps $s \in [0, .01)$ to $\phi(s) = .005$, $s \in [.01, .02)$ to $\phi(s) = .015$ and so on.
- State aggregated value functions are

$$\begin{aligned}\mathcal{F}_\phi &= \{f | f(s) = f(\phi(s)) \, \forall s\} \\ &= \{f | f(s) = f(\tilde{s}) \quad \text{if } \phi(s) = \phi(\tilde{s})\}\end{aligned}$$

  - In the simple case described above, $J \in \mathcal{F}_\phi$ is defined by the 99 values $J(.005), J(.015), \cdots, J(.995)$.
- $\Phi = \{\phi(s) : s \in \mathcal{S}\} = \{1, \cdots, m\}$. (. . . w.l.o.g) Set $\mathcal{S}_i = \{s \in \mathcal{S} : \phi(s) = i\}$ for $i = 1, \cdots, m$.

Convergence of FVI with state-aggregation

Theorem
Consider the iteration

$$J_{k+1} = \Pi_{\mathcal{F}_\phi,\nu} T J_k \quad k = 0, 1, \cdots$$

where $\nu(s) > 0$ *for all* $s \in \mathcal{S}$. *Then,*

$$\|J_k - \widehat{J}\|_\infty \le \alpha^k \|J_0 - \widehat{J}\|_\infty$$

where $\widehat{J}$ *solves the projected Bellman equation*

$$\widehat{J} = \Pi_{\mathcal{F}_\phi,\nu} T \widehat{J}.$$

Convergence proof

Lemma $\Pi_{\mathcal{F}_\phi,\nu} T$ is a contraction w.r.t $\|\cdot\|_\infty$ with modulus of contraction α.

Crucial fact: $\Pi_{\mathcal{F}_\phi,\nu}$ is a non-expansion in $\|\cdot\|_\infty$.

Proof.

$$\Pi_{\mathcal{F}_\phi,\nu} J \in \underset{\widehat{J} \in \mathcal{F}_\phi}{\operatorname{argmin}} \mathbb{E}_{s\sim\nu}\left[\left(\widehat{J}(s) - J(s)\right)^2\right]$$

It is solved by the conditional mean

$$\widehat{J}(i) = \mathbb{E}_{s\sim\nu}\left[J(s) \mid s \in \mathcal{S}_i\right]$$

$\square$

Performance loss bound

How effective is the greedy policy computed with respect to $\widehat{J}$?

Theorem
Let $\mu \in G(\widehat{J})$ *and* $\epsilon = \|\Pi_{\mathcal{F}_\phi,\nu} J^* - J^*\|_\infty$. *Then,*

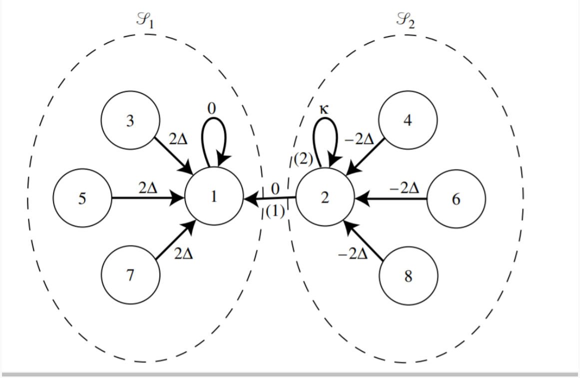
$$\|J_\mu - J^*\|_\infty \leq \frac{\alpha\|\widehat{J} - J\|_\infty}{1-\alpha} \leq \frac{\alpha\epsilon}{(1-\alpha)^2}$$

In this special case, we get a much stronger result than one that depends on the inherent Bellman error of the function class.

Tightness of the Approximation Error Bound

The dependence on $1/(1-\alpha)^2$ is highly problematic.

Unfortunately, it is tight due to an example discussed in detail in Van Roy (2006).



Reducing performance loss via state-relevance weighting

Recall the discounted state occupancy measure

$$d_{\infty}^{\mu} = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t d_0 P_{\mu}^t.$$

We showed a natural online algorithm converges toward a fixed point

$$\widehat{J} = \Pi_{\mathcal{F}_{\phi}, d_{\infty}^{\mu}} T\widehat{J} \qquad \mu \in G(\widehat{J}).$$

This satisfies the performance loss bound

$$\mathbb{E}_{s\sim d_0}\left[J_{\mu}(s) - J^*(s)\right] \leq \frac{\epsilon}{1-\alpha}$$

Comments:

- Saves a factor of $1/(1-\alpha)$ in the worst-case.
- The state-relevance-weighting is the fraction of time spent in a given state under the selected policy.
- This is a fixed point in $\{$the space of cost-to-functions$\} \times \{$the space of policies$\}$.

Table of Contents

Policy iteration

Policy iteration solves the $\min_\mu J_\mu$ *by solving a sequence of single period problems* $\mu_{k+1} \in \operatorname{argmin}_\mu T_\mu J_{\mu_k}$

- For $k = 0, 1, 2 \cdots$
  1. Policy evaluation:

     $$J_{\mu_k} = g_{\mu_k} + \alpha P_{\mu_k} J_{\mu_k}$$

  2. Policy improvement: $\mu_{k+1} \in G(J_{\mu_k}) = \{\mu : T_\mu J_{\mu_k} = T J_{\mu_k}\}$, i.e.

     $$\mu_{k+1}(s) \in \operatorname*{argmin}_{u \in U(s)} g(s,u) + \alpha \sum_{s' \in \mathcal{S}} P_{s,s'}(u) J_{\mu_k}(s') \qquad \forall s \in \mathcal{S}$$

Policy improvement property

Each step of policy iteration produces and improved policy, and the improvement is strict until an optimal policy is reached:

$$J_{\mu_{k+1}} = J_{\mu_k} \iff J_{\mu_k} = TJ_{\mu_k} \iff J_{\mu_k} = J^*.$$

Version 1: Simple bound on policy improvement.

Lemma $J_{\mu_{k+1}} \preceq TJ_{\mu_k} \preceq J_{\mu_k}$

Proof.

$J_{\mu_k} = T_{\mu_k} J_{\mu_k} \succeq TJ_{\mu_k} = T_{\mu_{k+1}} J_{\mu_k} \succeq T^2_{\mu_{k+1}} J_{\mu_k} \succeq \cdots \succeq J_{\mu_{k+1}}.$ $\square$

Version 2: a more precise quantification of policy improvement.

Lemma: $J_{\mu_{k+1}} = J_{\mu_k} - (I - \alpha P_{\mu_{k+1}})^{-1} (J_{\mu_k} - TJ_{\mu_k})$

Policy iteration with Q functions

Define the state-action cost-to-go function

$$Q_\mu(s, u) = g(s, u) + \alpha \sum_{s'} P_{ss'}(u) J_\mu(s')$$

This satisfies the Bellman equation:

$$Q_\mu(s, u) = g(s, u) + \alpha \sum_{s'} P_{ss'}(u)) Q_\mu(s', \mu(s'))$$

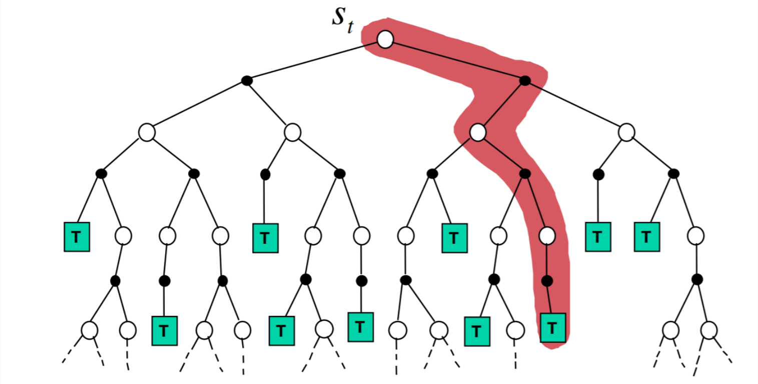
PI with Q functions:

- For $k = 0, 1, 2 \cdots$
  1. Policy evaluation:
  $$Q_{\mu_k}(s, u) = g(s, u) + \alpha \sum_{s'} P_{ss'}(u) Q_{\mu_k}(s', \mu_k(s')) \qquad \forall s, u$$
  2. Policy improvement:
  $$\mu_{k+1}(s) \in \operatorname*{argmin}_{u \in U(s)} Q_{\mu_k}(s, u) \qquad \forall s \in \mathcal{S}$$

Table of Contents

Reminder: Rollout

Choose which action to take at the current state by lookahead.



Reminder: Rollout (one-period lookahead)

- We have a base policy $\bar{\mu}$.
- At time k, in state s_k, we select

$$u_k \in \operatorname*{argmin}_u Q_{\bar{\mu}}(s_k, u)$$

- But we do this without storing the function Q_μ. How?
    - We can evaluate each control u by simulating many trajectories in which we first apply u and apply $\bar{\mu}$ thereafter:

$$\begin{aligned} Q_\mu(s,u) &= g(s,u) + \alpha \sum_{s' \in \mathcal{S}} P_{ss'}(u) J_\mu(s') \\ &= \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k g(s_k, u_k) : u_0 = u, s_0 = s,\ u_k = \bar{\mu}(s_k)\ k > 0\right] \end{aligned}$$

- This is done at decision-time, once s_k is observed.
    - No need to compute a full policy iteration update. Just lookahead in the current subproblem.

Table of Contents

Approximate Policy iteration (with Q functions)

Define the state-action cost-to-go function

$$Q_\mu(s,u) = g_\mu(u) + \alpha \sum_{s'} P_{ss'}(u) Q_\mu(s', \mu(s'))$$

The policy $\operatorname{argmin}_{u \in U(s)} Q_\mu(s,a)$ is a policy iteration update to μ.

Approximate PI:

- For $k = 0, 1, 2 \cdots$
  1. Approximate the cost-to-go under μ_k: $Q_{\theta_k} \approx Q_{\mu_k}$
  2. Solve for an improved policy

$$\mu_{k+1}(s) \in \operatorname*{argmin}_{u \in U(s)} Q_{\theta_k}(s,u) \qquad \forall s \in \mathcal{S}$$

Q_{μ_k} can be approximated by either TD or Monte Carlo methods.

Analysis of approximate PI

Lemma: Let $\epsilon = \max_{i \le k} \|Q_{\theta_i} - Q_{\mu_i}\|_\infty$. Then

$$\|Q_{\mu_k} - Q^*\|_\infty \le \alpha^k \|Q_{\mu_0} - Q^*\|_\infty + \frac{\alpha\epsilon}{(1-\alpha)^2}$$

But, it is difficult to form approximations with low error in the max-norm.