B9140 Dynamic Programming & Reinforcement Learning Lecture 6 - Oct 16, 2017

Fitted Value Iteration and SGD

Lecturer: Daniel Russo *Scribe: Mauro Escobar, Kleanthis Karakolios, Jingtong Zhao*

1 Projects

Work in groups of reasonable size.

Topics:

1. Theoretical
2. Applied/Modeling
    - Online education
    - Adaptive 'data-driven' healthcare (e.g. see work by usan Murphy)
    - Robotics
    - Arcade Games
    - Repeated customer interactions in e-commerce
3. Implementation for simple problems
4. Try 1, 2, 3 but turn in a thoughtful literature review

Advice:

a) Start simple.

b) For proofs, focus first on idealized settings (i.e. no noise).

c) Focus on understanding.

2 Fitted Value Iteration

Last time, we argued that infinite data, fitted value iteration essentially applies the updates: $V_{\theta_{k+1}} = \Pi T V_{\theta_k}$.

To recap, recall that $V_\theta = \Phi\theta$, and $V_\theta(s) = \phi(s)^\top \theta$.

The projection operator is $\Pi V = \underset{V' \in \text{span}(\Phi)}{\operatorname{argmin}} ||V' - V||_\pi$, where $\pi(s) = \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{s_t = s\}$.

So now we can recognize that fitted value iteration updates a value function $V_{\theta'}$ to a new value function:

$$\begin{aligned} &\underset{V_\theta}{\operatorname{argmin}} \frac{1}{|H|} \sum_{(s,r,s')\in H} (V_\theta(s) - (r + \gamma V_{\theta'}(s')))^2 \\ \approx\ &\underset{V_\theta}{\operatorname{argmin}} ||V_\theta - TV_{\theta'}||_\pi \\ \approx\ &\Pi T V_{\theta'}. \end{aligned}$$

Proposition 1. *$\Pi T : \mathbb{R}^S \to \mathbb{R}^S$ is a contraction w.r.t. $||\cdot||_\pi$ with modulus γ.*

Corollary 2. *V_{θ_k} converges at a geometric rate (in $||\cdot||_\pi$) to a solution of $V_\theta = \Pi T V_\theta$.*

Proposition 1 is established through a sequence of lemmas. The first is standard, and shows that a projection operator with respect to some norm is a non-expansion with respect to that norm. In two dimensions, this is an obvious consequence of the triangle inequality and can be seen by drawing a picture.

Lemma 3. *Π is a non-expansion w.r.t. $||\cdot||_\pi$ (i.e. $||\Pi V||_\pi \leq ||V||_\pi$).*

Proof.

$$\begin{aligned} ||V||_\pi^2 &= ||\Pi V + (I - \Pi)V||_\pi^2 \\ &= ||\Pi V||_\pi^2 + ||(I-\Pi)V||_\pi^2 \\ &\geq ||\Pi V||_\pi^2. \end{aligned}$$

□

The next lemma shows P is a non-expansion with respect to π. For this result it is crucial that $\pi(s)$ is a stationary distribution under P.

Lemma 4. *$||PV||_\pi \leq ||V||_\pi$.*

Proof. Recall that $P_{SS'} = P(S_{t+1} = S'|S_t = S)$. Hence we have

$$\begin{aligned} ||PV||_\pi^2 &= \sum_S \pi(S)(\sum_{S'} P_{SS'} V(S'))^2 \\ &\leq \sum_S \pi(S) \sum_{S'} P_{SS'} V(S')^2 \\ &= \sum_{S'} \sum_S \pi(S) P_{SS'} V(S')^2 \\ &= \sum_{S'} \pi(S') V(S')^2 \\ &= ||V||_\pi^2 \end{aligned}$$

where the second line follows from Jensen's inequality, and the fourth line follows from the definition of π (i.e. $\pi(S') = \sum_S \pi(S) P_{SS'}$). □

We now complete the proof of Proposition 1.
Proof of Proposition 1: Note that $TV = r + \gamma PV$ where $r(s)$ is the expected immediate reward under μ from state s.

$$\begin{aligned} ||\Pi TV - \Pi TV'||_\pi &= ||\Pi(TV - TV')||_\pi \\ &\leq ||TV - TV'||_\pi \\ &= ||\gamma PV - \gamma PV'||_\pi \\ &= \gamma ||P(V - V')||_\pi \\ &\leq \gamma ||V - V'||_\pi \end{aligned}$$

where the second line follows from Lemma 3, and the last line follows from Lemma 4. □

Note that the use of linear value function approximation was very important to this proof. Indeed, even the first line of the proof of Proposition 1 uses the linearity of the projection operator Π onto the space of approximate value functions $\{V_\theta\}$.

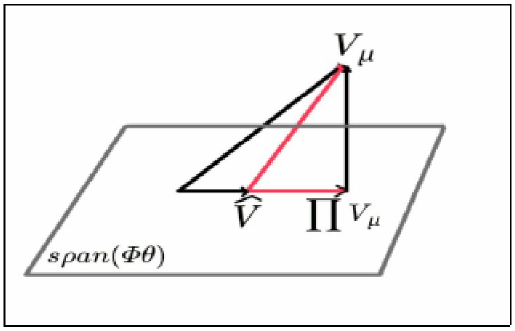
Hence, we know that fitted value iteration converges to a point V_∞ where the temporal error $V_\infty - TV_\infty$ is orthogonal to the features. **Is this any good?** We were hoping that $V_\infty = \Pi V_\mu$, that is $V_\infty = \min\limits_{V=\Phi\theta} \mathbb{E}_{S\sim\pi}\left[(V(S) - V_\mu(S))^2\right]$.

A conservative bound

Consider $\hat{V} = \Pi T \hat{V}$. Our benchmark is the best possible mean square error, $\|\Pi V_\mu - V_\mu\|_\pi$.

Proposition 5. $\|V_\mu - \hat{V}\|_\pi \leq \dfrac{\|V_\mu - \Pi V_\mu\|_\pi}{\sqrt{1-\gamma^2}}$.

This result, and many others in this lecture, were established by Tsitsiklis and Van Roy[1]. The dependence on the discount factor is a severe limitation of this bound, and one may wonder whether it is necessary. This bound is often conservative, but the dependence on the discount factor cannot be avoided in general. One may also consider whether this is the right performance measure at all. Perhaps we simply want that following policies derived from an approximate value function leads to effective decisions. Another paper of Van Roy[2] establishes performance loss bounds of this type which avoid the ugly dependence on the discount factor.



Proof. Since $\hat{V} = \Pi \hat{V}$, by orthogonality,

$$\begin{aligned}
\|V_\mu - \hat{V}\|_\pi^2 &= \|V_\mu - \Pi V_\mu\|_\pi^2 + \|\Pi V_\mu - \hat{V}\|_\pi^2 \\
&= \|V_\mu - \Pi V_\mu\|_\pi^2 + \|\Pi T_\mu V_\mu - \Pi T_\mu \hat{V}\|_\pi^2 \\
&\leq \|V_\mu - \Pi V_\mu\|_\pi^2 + \gamma^2 \|V_\mu - \hat{V}\|_\pi^2
\end{aligned}$$

$$\Rightarrow (1-\gamma^2)\|V_\mu - \hat{V}\|_\pi^2 \leq \|V_\mu - \Pi V_\mu\|_\pi^2$$

□

Divergence with off policy sampling

Consider a two-state MDP, $S = \{1,2\}$, where there exists only one policy with only one available action at each state. The reward is always 0.

We have a one-dimensional function approximation:

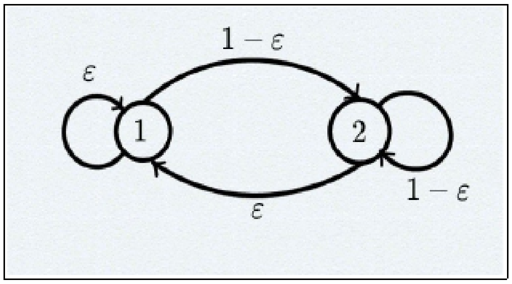
$$\Phi = (1,2) \qquad \Rightarrow \qquad \begin{aligned} (\Phi\theta)(1) &= \theta \\ (\Phi\theta)(2) &= 2\theta \end{aligned}$$

Notice that $V_\mu = \Phi \cdot 0$. Previous error bound will imply that there is no error.

Consider the follwoing MDP:

[1] Tsitsiklis, J. and Van Roy, B. (1997). An Analysis of Temporal-Difference Learning with Function Approximation. *IEEE Transactions on Automatic Control*, 42(5): 674- 690.

[2] Van Roy, B. (2006). Performance Loss Bounds for Approximate Value Iteration with State Aggregation. *Mathematics of Operations Research*, 31(2): 234-244.



Let $\Pi V = \underset{\Phi\theta}{\operatorname{argmin}} \|\Phi\theta - V\|_2$, which is the projection operator that arises when states are sampled uniformly, rather than from the stationary distribution of the MDP. Then

$$(\Phi\theta_{k+1}) = \Pi T(\Phi\theta_k) \qquad \text{and} \qquad TV(1) = \gamma(\varepsilon V(1) + (1-\varepsilon)V(2)) = TV(2).$$

Hence, $(T\Phi\theta_k)(1) = \gamma(2-\varepsilon)\theta_k = (T\Phi\theta_k)(2)$. Now

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmin}} \left\| \begin{pmatrix} \theta \\ 2\theta \end{pmatrix} - \gamma(2-\varepsilon)\theta_k \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\| = \frac{3}{5}\gamma(2-\varepsilon)\theta_k = c\theta_k = c^{k+1}\theta_0, \qquad \text{for } c = \frac{3}{5}\gamma(2-\varepsilon).$$

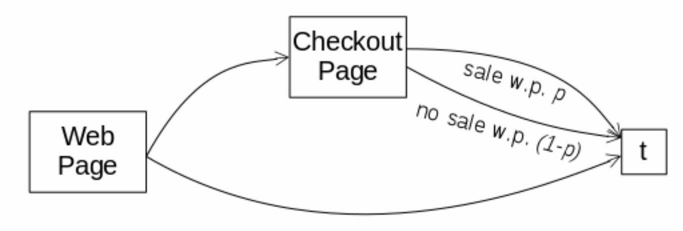
However, if $\gamma \approx 1$ and $\varepsilon \approx 0$, then $c > 1$ and $\theta_k = c^k\theta_0$ will grow exponentially to infinity.

(Area of research: ‘Off Policy Reinforcement Learning’)

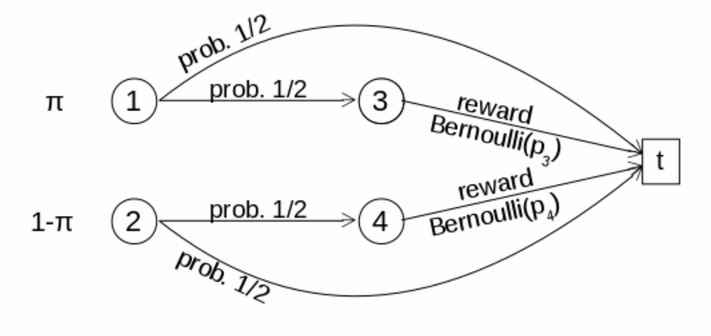
Example

Let us build intuition about the convergence point of fitted value iteration and why it may differ from that of the the monte-carlo estimator.

Consider again the example of a webpage with an advertisement. A customer either clicks in the ad and goes to the checkout page or ignores the ad, in which case we consider that he goes to a terminal state. From the checkout page, with probability p the customer buys the item advertised.



Suppose there are 2 ads, the first one appears with probability π and the second one with probability $1-\pi$. In each case, the probability that the customer clicks on the ad is $1/2$, and the reward obtained is 1 if the sale is made, with probability p_3 for ad 1 and p_4 for ad 2. The following diagram describes the sale process for states 1, 2, 3, and 4.



We will consider a feature matrix that collapses states 3 and 4, that is

$$\Phi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then, $\{\Phi\theta : \theta \in \mathbb{R}^3\} = \{V \in \mathbb{R}^4 : V(3) = V(4)\}$. In words, the model assumes the sale probability once the checkout page is reached does not depend on the initial ad that was shown, whereas it does in relaity.

Let's compute now the MC and TD estimators for V:

$$\begin{aligned} \hat{V}_{MC}(1) &= p_3/2 = V_\mu(1) & \hat{V}_{TD}(3) &= \hat{V}_{TD}(4) = \pi p_3 + (1-\pi)p_4 \\ \hat{V}_{MC}(2) &= p_4/2 = V_\mu(2) & \hat{V}_{TD}(1) &= \tfrac{1}{2}\cdot 0 + \tfrac{1}{2}\cdot \hat{V}_{TD}(3) \neq V_\mu(1) \\ \hat{V}_{MC}(3) &= \hat{V}_{MC}(4) = \pi p_3 + (1-\pi)p_4 & \hat{V}_{TD}(2) &= \tfrac{1}{2}\cdot 0 + \tfrac{1}{2}\cdot \hat{V}_{TD}(3) \neq V_\mu(2) \end{aligned}$$

We see that in this example MC captures a better estimator than TD. Both produce incorrect estimates at states 3 and 4, but MC nevertheless reaches the correct estimate at states 1 and 2. The TD method tries to be as *temporally consistent* as possible, and as a result the incorrect values at states 3 and 4 propagate backward to generate incorrect estimates at states 1 and 2.

3 Incremental Training

How to implement $V_{\theta_{k+1}} = \Pi T V_{\theta_k}$? Recall

$$\begin{aligned} \theta_{k+1} &= \operatorname*{argmin}_\theta \frac{1}{|H|} \sum_{(S,r,S')} \left(V_\theta(S) - (r + \gamma V_{\theta_k}(S'))\right)^2 \\ &= \operatorname*{argmin}_\theta \frac{1}{|H|} \sum_{(S,r,S')} \left(\Phi(S)\theta - (r + \gamma V_{\theta_k}(S'))\right)^2 \\ &= \operatorname*{argmin}_\theta \|A\theta - y\|_2, \end{aligned}$$

for some suitable A and y. To solve this:

Idea 1. Solve normal equations: complexity $O(d^3)$, where d is the dimension of θ.

Idea 2. Gradient descent: $O(|H|)$ per iteration.

Idea 3. **Stochastic Gradient Descent (SGD)**

Algorithm 1: Stochastic Gradient Descent for computing θ_{k+1}

```
Input  : $\theta_k$ and a step size sequence $(\alpha_t : t \in \mathbb{N})$
Output: $\theta_{k+1}$
let $\theta := \theta_k$
for $t = 1, 2, 3, \ldots$ do
    sample $(s, r, s')$ from $H$
    set $y := r + \gamma V_{\theta_k}(s')$
    compute $g := \frac{\partial}{\partial \theta}(V_\theta(s) - y)^2$
    assign $\theta \leftarrow \theta - \alpha_t g$
end
return $\theta_{k+1} := \theta$
```