CMPT 419/983: Theoretical Foundations of Reinforcement Learning

Lecture 6

Sharan Vaswani

October 13, 2023

Recap

- We have studied algorithms (VI/PI/LP) that use knowledge of the transition probabilities $\mathcal{P}$ and rewards r to compute the optimal policy.
- These quantities are difficult to obtain in practical scenarios, and hence we need methods that can compute the optimal policy without explicitly relying on this information.
- Today, we will consider evaluating a fixed policy π without explicit knowledge of $\mathcal{P}$ and r.

Policy Evaluation

For a fixed policy π and starting state s_0, $v^\pi(s_0) = \mathbb{E}\left[X | S_0 = s_0\right]$ where $X := \sum_{t=0}^{\infty} \gamma^t R_t$.

$$\mathbb{E}\left[X|S_0 = s_0\right] = \mathbb{E}_{A_0}\left[\mathbb{E}\left[X|S_0 = s_0, A_0\right]\right] = \mathbb{E}_{A_0}\left[\mathbb{E}_{S_1|\{S_0,A_0\}}\left[\mathbb{E}\left[X|S_0 = s_0, A_0, S_1\right]\right]\right]$$

(Using that $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$)

$$= \mathbb{E}_{A_0}\,\mathbb{E}_{S_1|\{S_0,A_0\}}\,\mathbb{E}_{A_1|\{S_0,A_0,S_1\}}\ldots\mathbb{E}_{S_t|\{S_0,A_0,\ldots S_{t-1},A_{t-1}\}}\mathbb{E}\left[X|\{S_0, A_0, \ldots, S_{t-1}, A_{t-1}\}\right]$$

(Unrolling recursively)

$$= \mathbb{E}_{A_0}\,\mathbb{E}_{S_1|\{S_0,A_0\}}\,\mathbb{E}_{A_1|\{S_0,A_0,S_1\}}\ldots\mathbb{E}_{S_t|\{S_{t-1},A_{t-1}\}}\mathbb{E}\left[X|\{S_0, A_0, \ldots, S_{t-1}, A_{t-1}\}\right]$$

(Markov assumption)

$$= \mathbb{E}_{A_0}\,\mathbb{E}_{S_1|\{S_0,A_0\}}\,\mathbb{E}_{A_1|S_1}\ldots\mathbb{E}_{S_t|\{S_{t-1},A_{t-1}\}}\mathbb{E}\left[X|\{S_0, A_0, \ldots, S_{t-1}\}\right]$$

(Restricting to Markov policies)

$$= \mathbb{E}_{A_0}\left[R_0 + \mathbb{E}_{S_1|\{S_0,A_0\}}\,\mathbb{E}_{A_1|S_1}\left[\gamma R_1 + \ldots\mathbb{E}_{S_t|\{S_{t-1},A_{t-1}\}}\left[\gamma^t R_t + \ldots\right]\right]\right]$$

(Distributing the sum)

Policy Evaluation

The unrolling on the previous slide suggests a Monte-Carlo sampling scheme:

- Starting from s_0, for $t \geq 0$, sample $a_t \sim \pi(\cdot|s_t)$, the environment transitions to s_{t+1} (equivalent to sampling $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$). This generates a trajectory $\tau = (s_0, a_0, s_1, \ldots)$.
- Collect rewards $r_t = r(s_t, a_t)$, calculate $R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$. Note that $\mathbb{E}[R(\tau)] = v^\pi(s_0)$.
- In order to reduce the variance, generate m trajectories $\{\tau_i\}_{i=1}^m$, calculate $R(\tau_i)$ and output the empirical average: $\hat{v} := \frac{\sum_{i=1}^m R(\tau_i)}{m}$ as an approximation to $v^\pi(s_0)$.

Q: What is the problem with this approach?

Solution 1: Truncate the trajectory to H steps, i.e. calculate $R(\tau) = \sum_{t=0}^{H-1} \gamma^t r_t$.

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t - \sum_{t=H}^{\infty} \gamma^t r_t \implies \mathbb{E}[R(\tau)] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] - \mathbb{E}\left[\sum_{t=H}^{\infty} \gamma^t r_t\right] = v^\pi(s_0) - \sum_{t=H}^{\infty} \gamma^t r_t$$

$$\implies |v^\pi(s_0) - \mathbb{E}[R(\tau)]| \leq \frac{\gamma^H}{1-\gamma} \qquad (r_t \leq 1, \text{ Sum of geometric series.})$$

Policy Evaluation

Claim: Using $m = \frac{\ln(2/\delta)}{2\epsilon^2 (1-\gamma)^2}$ trajectories with $H \geq \frac{\ln(1/\epsilon(1-\gamma))}{\ln(1/\gamma)}$ guarantees that $|\hat{v} - v^\pi(s_0)| \leq \epsilon$ with probability $1 - \delta$.

Proof: Recall that $\hat{v} = \frac{\sum_{i=1}^m R(\tau_i)}{m}$.

$$|v^\pi(s_0) - \mathbb{E}[\hat{v}]| = \left| v^\pi(s_0) - \frac{\sum_{i=1}^m \mathbb{E}[R(\tau_i)]}{m} \right| = \left| \frac{\sum_{i=1}^m [v^\pi(s_0) - \mathbb{E}[R(\tau_i)]]}{m} \right|$$

$$\leq \frac{\sum_{i=1}^m |[v^\pi(s_0) - \mathbb{E}[R(\tau_i)]]|}{m} \leq \frac{\gamma^H}{1-\gamma}$$

$$|\hat{v} - v^\pi(s_0)| = |\hat{v} - \mathbb{E}[\hat{v}] + \mathbb{E}[\hat{v}] - v^\pi(s_0)| \leq |\hat{v} - \mathbb{E}[\hat{v}]| + |\mathbb{E}[\hat{v}] - v^\pi(s_0)|$$

$$\leq |\hat{v} - \mathbb{E}[\hat{v}]| + \frac{\gamma^H}{1-\gamma} \leq |\hat{v} - \mathbb{E}[\hat{v}]| + \frac{\epsilon}{2} \qquad \left(\text{Using } H \geq \frac{\ln(1/\epsilon(1-\gamma))}{\ln(1/\gamma)}\right)$$

$$|\hat{v} - \mathbb{E}[\hat{v}]| = \left| \frac{X_m - \mathbb{E}[X_m]}{m} \right| \qquad \left(X_m := \textstyle\sum_{i=1}^m R(\tau_i)\right)$$

Since the $R(\tau_i)$ r.v's are i.i.d, we can use Hoeffding's inequality.

Policy Evaluation

Recall that $|\hat{v} - v^\pi(s_0)| \leq |\hat{v} - \mathbb{E}[\hat{v}]| + \frac{\epsilon}{2}$. Here, $|\hat{v} - \mathbb{E}[\hat{v}]| = \left|\frac{X_m - \mathbb{E}[X_m]}{m}\right|$ where $X_m := \sum_{i=1}^m R(\tau_i)$.

Hoeffding's Inequality: For m i.i.d. r.v's such that $X_i \in [a_i, b_i]$. For $t > 0$,

$$\Pr[|X_m - \mathbb{E}[X_m]| \geq t] \leq 2 \exp\left(\frac{-2t^2}{\sum_{i=1}^m (b_i - a_i)^2}\right)$$

$R(\tau_i) \in [0, 1/1-\gamma]$. Setting $t = m\,\epsilon$,

$$\Pr\left[\left|\frac{X_m - \mathbb{E}[X_m]}{m}\right| \geq \epsilon\right] \leq 2 \exp\left(-2m\,\epsilon^2\,(1-\gamma)^2\right)$$

$$\implies \Pr\left[\left|\frac{X_m - \mathbb{E}[X_m]}{m}\right| \geq \epsilon\right] \leq \delta \qquad \left(\text{Setting } m = \frac{\ln(2/\delta)}{2\epsilon^2\,(1-\gamma)^2}\right)$$

Putting everything together, with probability $1 - \delta$, $|\hat{v} - v^\pi(s_0)| \leq \epsilon$. $\square$

Solution 2: Randomly truncate the trajectory i.e. sample H from a geometric distribution with parameter $1 - \gamma$, return $R(\tau) = \sum_{t=0}^{H-1} r_t$. Eliminates the bias from using a fixed truncation.

Claim: $\mathbb{E}_H\,\mathbb{E}_\tau[R(\tau)] = v^\pi(s_0)$. Prove in Assignment 2!

Policy Evaluation

- **Problem 1**: To estimate $v^\pi \in \mathbb{R}^S$, we need fresh trajectories for estimating $v^\pi(s)$ for each $s \in \mathcal{S}$. We need to restart the sampling each time, which may not always be possible.
- *Sol*: Sample a single trajectory, estimate $v^\pi(s)$ as the cumulative discounted sum of rewards following the first time state s is visited. This is referred to as "first visit" Monte-Carlo. Can also average the returns following "every visit" to state s. Both strategies can be shown to produce unbiased estimates of v^π. For more details, see [SB18, Chapter 5].
- If $\hat{v}_k$ is the empirical average after sampling $k \in [1, m]$ trajectories, we can update it in an online fashion: $\hat{v}_k = \hat{v}_{k-1} + \frac{R(\tau_k) - \hat{v}_{k-1}}{k-1}$.
- **Problem 2**: Hence, $\hat{v}_k$ is updated only after observing the rewards from the entire trajectory. This could be slow when the trajectories are long. Moreover, Monte-Carlo estimation does not exploit the MDP structure effectively.
- *Sol*: Temporal Difference Learning

Temporal Difference Learning

Idea: Exploit the Bellman equation and combine it with Monte-Carlo estimation.

Recall that, for starting state s, for a fixed policy π,

$$v^\pi(s) = \mathbf{r}_\pi(s) + \gamma \sum_{s'} P_\pi[s, s']\, v^\pi(s') = \sum_{a \in \mathcal{A}} r(s, a)\, \pi[a|s] + \gamma \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{P}[s'|s, a]\, \pi[a|s]\, v^\pi(s')$$

$$= \sum_{a \in \mathcal{A}} \pi[a|s] \left[r(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}[s'|s, a] v^\pi(s') \right] = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[r(s, a) + \gamma\, \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}[v^\pi(s')] \right]$$

$$\implies v^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\, \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \left[r(s, a) + \gamma v^\pi(s') \right]$$

Sampling a from $\pi(\cdot|s)$ and the environment samples $s' \sim \mathcal{P}(\cdot|s, a)$, $\hat{v}^\pi(s) = r(s, a) + \gamma\, v^\pi(s')$.

Since we do not know $v^\pi(s')$ either, we can use the estimate instead, implying that, $\hat{v}^\pi(s) = r(s, a) + \gamma\, \hat{v}^\pi(s')$. This is known as *bootstrapping* since we are using an estimate at s' to estimate the value function at state s.

Using this idea, we can design an iterative algorithm – TD(0).

Temporal Difference Learning

Algorithm Temporal Difference Learning. [TD(0)]

1: **Input**: MDP $M = (\mathcal{S}, \mathcal{A}, \rho)$, $v_0 = 0$, Policy π. Step-sizes $\{\alpha_t\}_{t=0}^{T-1}$.
2: Sample state $s_0 \sim \rho$.
3: **for** $t = 0 \to T-1$ **do**
4: Take action $a_t \sim \pi(\cdot|s_t)$, observe reward $r(s_t, a_t)$ and transition to state s_{t+1}.
5: Update $v_{t+1}(s_t) = (1 - \alpha_t)\, v_t(s_t) + \alpha_t \left[r(s_t, a_t) + \gamma\, v_t(s_{t+1})\right]$.
6: $\forall s \neq s_t,\ v_{t+1}(s) = v_t(s)$
7: **end for**

- Unlike Monte-Carlo estimation, TD(0) does not require waiting until the end of trajectories to start updating the value function estimates.
- Unlike using $\mathcal{T}_\pi$, TD(0) does not require knowledge of $\mathcal{P}$ and r.
- Under some technical assumptions, TD(0) will converge, i.e. $\lim_{t\to\infty} v_t = v^\pi$.
- TD(0) can handle linear function approximation and has non-asymptotic theoretical convergence guarantees. We will prove this next.

Linear Temporal Difference Learning

Linear TD(0)

Assumption: Have access to features $\Phi \in \mathbb{R}^{S \times d}$ such that for every policy π, there exists a $\theta \in \mathbb{R}^d$ such that $v^\pi = \Phi\theta$. For the specific policy π being evaluated, there exists a unique θ^* such that $v^\pi = \Phi\theta^* = v_{\theta^*}$ where $v_\theta := \Phi\theta$.

Define $\phi(s)$ as the feature vector corresponding to state s. Hence, $v_\theta(s) = \langle \phi(s), \theta \rangle$. For convenience, we will assume that $\forall s, \|\phi(s)\| \leq 1$.

Algorithm TD(0) with linear function approximation

Input: MDP $M = (\mathcal{S}, \mathcal{A}, \rho)$, Features $\Phi \in \mathbb{R}^{S \times d}$, Policy π. $\theta_0 \in R^d$, Step-sizes $\{\alpha_t\}_{t=0}^{T-1}$.
Sample state $s_0 \sim \rho$
for $t = 0 \to T - 1$ **do**
  Take action $a_t \sim \pi(\cdot|s_t)$, observe reward $r(s_t, a_t)$ and transition to state s_{t+1}.
  Define $g_t(\theta) = [r_t + \gamma\langle \theta, \phi(s_{t+1})\rangle - \langle \theta, \phi(s_t)\rangle]\, \phi(s_t)$
  Update $\theta_{t+1} = \theta_t + \alpha_t\, g_t(\theta_t)$
end for

If $d = S$ and $\phi(s)$ correspond to one-hot vectors, then we recover TD(0) from the previous slide.

Linear TD(0) Analysis

The TD(0) update is $\theta_{t+1} = \theta_t + \alpha_t \, g_t(\theta)$ where $g_t(\theta) = [r_t + \gamma\langle\theta, \phi(s_{t+1})\rangle - \langle\theta, \phi(s_t)\rangle] \, \phi(s_t)$.

Q: Could we use a Gradient Descent type analysis?

We will analyze Linear TD(0) in 4 steps:

(1) Warmup: Analyze a hypothetical algorithm that performs GD on $f(\theta) := \frac{1}{2} \left\| v_{\theta^*} - v_\theta \right\|_D^2$.

(2) Mean-path: Make an analogy between Linear TD(0) and GD, and analyze Linear TD(0) assuming access to the stationary distribution.

(3) IID: Analyze Linear TD(0) assuming access to (s_t, s_{t+1}) sampled i.i.d from the stationary distribution.

(4) Markovian: Analyze *Projected* Linear TD(0) assuming access to (s_t, s_{t+1}) that are gathered from a “fast-mixing” Markov chain (will not cover this in detail).

Linear TD(0) Analysis

Define $P(s'|s)$ to be the probability of transitioning from s to s' when acting according to π.

Assumption: The Markov chain induced by policy π is ergodic (can visit every state) with a unique stationary distribution $\omega \in \Delta_S$. For $s \in \mathcal{S}$, $\omega(s) = \lim_{t\to\infty} \Pr[s_t = s]$. Hence, $\omega \, \mathbf{P}^\pi = \omega$ meaning that if $s \sim \omega$ and $s' \sim P(\cdot|s)$, then the marginal distribution of s' is ω.

Define a diagonal matrix $D \in \mathbb{R}^{S\times S}$ such that $D_{i,i} = \omega(i)$. For any $u, w \in \mathbb{R}^S$, define $\|u - w\|_D^2 = \sum_s \omega(s) \, [u(s) - w(s)]^2$.
For v_θ and $v_{\theta'}$, define $\Sigma := \sum_s \omega(s) \, \phi(s)\phi(s)^T \in \mathbb{R}^{d\times d}$ and $\lambda := \lambda_{\min}[\Sigma]$.

$$\|v_\theta - v_{\theta'}\|_D^2 = \sum_s \omega(s) \, [v_\theta(s) - v_{\theta'}(s)]^2 = \sum_s \omega(s) \, [\langle \phi(s), \theta - \theta' \rangle]^2$$
$$= (\theta - \theta')^T \sum_s \omega(s) \, \phi(s)\phi(s)^T (\theta - \theta') = \|\theta - \theta'\|_\Sigma^2$$

Q: Prove that $\lambda_{\max}[\Sigma] \leq 1$

Hence, for any θ, $\sqrt{\lambda} \, \|\theta\| \leq \|v_\theta\|_D \leq \|\theta\|$ (by setting $\theta' = 0$ above).

Linear TD(0) Analysis – Warmup

Define $f(\theta) := \frac{1}{2}\left\|v_{\theta^*} - v_\theta\right\|_D^2 = \frac{1}{2}\left\|\theta^* - \theta\right\|_\Sigma^2$. Consider a hypothetical algorithm that performs GD on $f(\theta)$ i.e. at iteration t, $\theta_{t+1} = \theta_t - \alpha \nabla f(\theta_t)$. Note that $\nabla f(\theta) = \Sigma(\theta - \theta^*)$.

$$\left\|\theta_{t+1} - \theta^*\right\|^2 = \left\|\theta_t - \alpha\,\nabla f(\theta_t) - \theta^*\right\|^2 = \left\|\theta_t - \theta^*\right\|^2 + 2\alpha\left\langle \nabla f(\theta_t), \theta^* - \theta_t\right\rangle + \alpha^2\left\|\nabla f(\theta_t)\right\|^2$$

$$\left\langle \nabla f(\theta_t), \theta^* - \theta_t\right\rangle = \left\langle \Sigma(\theta_t - \theta^*), \theta^* - \theta_t\right\rangle = -\left\|\theta_t - \theta^*\right\|_\Sigma^2 = -\left\|v_{\theta_t} - v_{\theta^*}\right\|_D^2$$

For any vector u s.t. $\|u\| \leq 1$,

$$\langle u, \nabla f(\theta)\rangle = \langle u, \Sigma(\theta - \theta^*)\rangle \leq \left\|\Sigma^{1/2}\,u\right\|\,\left\|\Sigma^{1/2}\,(\theta - \theta^*)\right\| \qquad \text{(Cauchy Schwarz)}$$

$$= \|u\|_\Sigma\,\left\|\theta - \theta^*\right\|_\Sigma \leq \lambda_{\max}[\Sigma]\,\|u\|\,\left\|\theta - \theta^*\right\|_\Sigma \leq \left\|v_\theta - v_{\theta^*}\right\|_D \qquad (\lambda_{\max}[\Sigma] \leq 1,\ \|u\| \leq 1)$$

$$\implies \left\|\nabla f(\theta)\right\|^2 \leq \left\|v_\theta - v_{\theta^*}\right\|_D^2 \qquad \left(\text{Setting } u = {}^{\nabla f(\theta)}/_{\|\nabla f(\theta)\|}\right)$$

$$\implies \left\|\theta_{t+1} - \theta^*\right\|^2 \leq \left\|\theta_t - \theta^*\right\|^2 - 2\alpha\,\left\|v_{\theta_t} - v_{\theta^*}\right\|_D^2 + \alpha^2\,\left\|v_{\theta_t} - v_{\theta^*}\right\|_D^2$$

$$\left\|\theta_{t+1} - \theta^*\right\|^2 \leq \left\|\theta_t - \theta^*\right\|^2 - \left\|v_{\theta_t} - v_{\theta^*}\right\|_D^2 \leq (1 - \lambda)\,\left\|\theta_t - \theta^*\right\|^2 \qquad (\text{Set } \alpha = 1,\ \lambda = \lambda_{\min}[\Sigma])$$

$$\implies \left\|\theta_T - \theta^*\right\|^2 \leq (1 - \lambda)^T\,\left\|\theta_0 - \theta^*\right\|^2 \qquad (\text{Recursing from } t = 0 \text{ to } T - 1)$$

Linear TD(0) Analysis – Mean-path

The previous analysis relied on bounding two key quantities: (i) $\langle \nabla f(\theta_t), \theta^* - \theta_t \rangle$ and (ii) $\|\nabla f(\theta)\|^2$. We now consider analyzing Mean-path TD. For this, define $\bar{g}(\theta)$ and the corresponding update as:

$$\bar{g}(\theta) := \mathbb{E}_{s\sim\omega}\mathbb{E}_{s'\sim P(\cdot|s)}\left[r(s,\pi(s)) + \gamma\langle\theta, \phi(s')\rangle - \langle\theta, \phi(s)\rangle\right]\phi(s)$$
$$\theta_{t+1} = \theta_t + \alpha\, \bar{g}(\theta)$$

- Intuitively, $\bar{g}(\theta)$ is the Linear TD update in expectation if s was sampled from the stationary distribution, and the Markov chain transitioned to s'.
- Importantly, recall that the marginal distribution of s' is the stationary distribution ω.
- If $\mathcal{T}_\pi$ is the policy evaluation operator for π, then, $\bar{g}(\theta) = \Phi^T D\,[\mathcal{T}_\pi \Phi\theta - \Phi\theta]$ (Prove in Assignment 3!).

Similar to the warm-up, we will show two important properties for $\bar{g}(\theta)$. For all θ,

(1) $\langle \bar{g}(\theta), \theta^* - \theta\rangle \geq (1-\gamma)\,\|v_\theta - v_{\theta^*}\|_D^2$
(2) $\|\bar{g}(\theta)\| \leq 2\sqrt{2}\,\|v_\theta - v_{\theta^*}\|_D$

Linear TD(0) Analysis – Mean-path

Claim: $\langle \bar{g}(\theta), \theta^* - \theta \rangle \geq (1-\gamma) \, \|v_\theta - v_{\theta^*}\|_D^2$.

Proof: Since $\bar{g}(\theta) = \Phi^T D \, [\mathcal{T}_\pi \Phi\theta - \Phi\theta]$, using the definition of θ^*, $\bar{g}(\theta^*) = \Phi^T D \, [\mathcal{T}_\pi \Phi\theta^* - \Phi\theta^*] = \Phi^T D \, [\mathcal{T}_\pi v^\pi - v^\pi] = 0$. Hence,

$$
\begin{aligned}
\bar{g}(\theta) &= \bar{g}(\theta) - \bar{g}(\theta^*) \\
&= \mathbb{E}_{s,s'} \left[\left[(r(s,\pi(s)) + \gamma\langle \theta, \phi(s')\rangle - \langle \theta, \phi(s)\rangle) - (r(s,\pi(s)) + \gamma\langle \theta^*, \phi(s')\rangle - \langle \theta^*, \phi(s)\rangle) \right] \phi(s) \right] \\
&= \mathbb{E}_{s,s'} \left[(\langle \phi(s), \theta^* - \theta \rangle - \gamma\langle \phi(s'), \theta^* - \theta\rangle) \, \phi(s) \right]
\end{aligned}
$$

Define $\zeta_s := \langle \theta^* - \theta, \phi(s) \rangle$ and $\zeta_{s'} := \langle \theta^* - \theta, \phi(s') \rangle$

$$
\implies \bar{g}(\theta) = \mathbb{E}_{s,s'} \left[(\zeta_s - \gamma\zeta_{s'}) \, \phi(s) \right]
$$

$$
\begin{aligned}
\langle \bar{g}(\theta), \theta^* - \theta \rangle &= \langle \mathbb{E}_{s,s'} \left[(\zeta_s - \gamma\zeta_{s'}) \, \phi(s) \right], \theta^* - \theta \rangle = \mathbb{E}_{s,s'} \left[(\zeta_s - \gamma\zeta_{s'}) \, \langle \phi(s), \theta^* - \theta \rangle \right] \\
&= \mathbb{E}_{s,s'} \left[(\zeta_s - \gamma\zeta_{s'}) \; \zeta_s \right] = \mathbb{E}_{s,s'} \left[\zeta_s^2 - \gamma\zeta_{s'} \, \zeta_s \right]
\end{aligned}
$$

$$
\implies \langle \bar{g}(\theta), \theta^* - \theta \rangle = \mathbb{E}_{s\sim\omega}\mathbb{E}[\zeta_s^2] - \gamma \, \mathbb{E}_{s\sim\omega, s'\sim P(\cdot|s)} \left[\zeta_{s'} \, \zeta_s \right]
$$

Linear TD(0) Analysis – Mean-path

Recall that $\langle \bar{g}(\theta), \theta^* - \theta \rangle = \mathbb{E}_{s\sim\omega}\mathbb{E}[\zeta_s^2] - \gamma\, \mathbb{E}_{s\sim\omega, s'\sim P(\cdot|s)}\left[\zeta_{s'}\, \zeta_s\right]$ where $\zeta_s := \langle \theta^* - \theta, \phi(s) \rangle$.

$$
\begin{aligned}
\langle \bar{g}(\theta), \theta^* - \theta \rangle &= \mathbb{E}_{s\sim\omega}[\zeta_s^2] - \gamma\, \mathbb{E}_{s\sim\omega, s'\sim P(\cdot|s)}\left[\zeta_{s'}\, \zeta_s\right] \\
&\geq \mathbb{E}_{s\sim\omega}\mathbb{E}[\zeta_s^2] - \gamma \sqrt{\mathbb{E}_{s\sim\omega, s'\sim P(\cdot|s)}[\zeta_s^2]}\, \sqrt{\mathbb{E}_{s\sim\omega, s'\sim P(\cdot|s)}[\zeta_{s'}^2]} && \text{(Cauchy Schwarz)} \\
&= \mathbb{E}_{s\sim\omega}[\zeta_s^2] - \gamma \sqrt{\mathbb{E}_{s\sim\omega}[\zeta_s^2]}\, \sqrt{\mathbb{E}_{s'\sim\omega}[\zeta_{s'}^2]} && \text{(ω is the stationary distribution)} \\
&= (1-\gamma)\, \mathbb{E}_{s\sim\omega}[\zeta_s^2] = (1-\gamma) \sum_s \omega(s)\, \zeta^2(s) \\
&= (1-\gamma) \sum_s \omega(s)\, (\theta^* - \theta)^T \phi(s)\phi(s)^T\, (\theta^* - \theta) && \text{(By def. of ζ_s)} \\
&= (1-\gamma)\, \|\theta - \theta^*\|_\Sigma^2 && \text{(By def. of Σ)} \\
\implies \langle \bar{g}(\theta), \theta^* - \theta \rangle &\geq (1-\gamma)\, \|v_\theta - v_{\theta^*}\|_D^2 \quad \square && \text{(Since $\|\theta - \theta^*\|_\Sigma = \|v_\theta - v_{\theta^*}\|_D$)}
\end{aligned}
$$

Linear TD(0) Analysis – Mean-path

Claim: $\|\bar{g}(\theta)\| \leq 2\sqrt{2}\,\|v_\theta - v_{\theta^*}\|_D$.

Proof: Since $\bar{g}(\theta) = \mathbb{E}_{s,s'}\left[(\zeta_s - \gamma\zeta_{s'})\,\phi(s)\right]$,

$$\|\bar{g}(\theta)\| = \|\mathbb{E}_{s,s'}\left[(\zeta_s - \gamma\zeta_{s'})\,\phi(s)\right]\| \leq \mathbb{E}_{s,s'}\,\|[(\zeta_s - \gamma\zeta_{s'})\,\phi(s)]\| \qquad \text{(Jensen's inequality)}$$

$$= \mathbb{E}_{s,s'}\left[|\zeta_s - \gamma\zeta_{s'}|\;\|\phi(s)\|\right] \leq \sqrt{\mathbb{E}\left[(\zeta_s - \gamma\zeta_{s'})^2\right]}\,\sqrt{\mathbb{E}[\|\phi(s)\|^2]} \quad \text{(Cauchy Schwarz)}$$

$$\leq \sqrt{\mathbb{E}\left[(\zeta_s - \gamma\zeta_{s'})^2\right]} \qquad \text{(Since } \|\phi(s)\| \leq 1\text{)}$$

$$\leq \sqrt{2}\,\sqrt{\mathbb{E}\left[\zeta_s^2 + \gamma^2\zeta_{s'}^2\right]} \leq \sqrt{2}\,\sqrt{\mathbb{E}_{s\sim\omega}[\zeta_s^2]} + \sqrt{2}\sqrt{\gamma^2\mathbb{E}_{s\sim\omega,s'\sim P(\cdot|s)}[\zeta_{s'}^2]}$$

(Since $(a+b)^2 \leq 2(a^2+b^2)$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a \geq 0, b \geq 0$)

$$= \sqrt{2}\,\sqrt{\mathbb{E}_{s\sim\omega}[\zeta_s^2]} + \sqrt{2}\,\gamma\sqrt{\mathbb{E}_{s'\sim\omega}[\zeta_{s'}^2]} = \sqrt{2}\,(1+\gamma)\,\sqrt{\mathbb{E}[\zeta_s^2]}$$

(Since ω is the stationary distribution)

$$\leq 2\sqrt{2}\,\sqrt{\mathbb{E}[\zeta_s^2]} \qquad \text{(Since } 1+\gamma < 2\text{)}$$

$$\implies \|\bar{g}(\theta)\| \leq 2\sqrt{2}\,\|v_\theta - v_{\theta^*}\|_D \quad \square \qquad \text{(Using the bound on } \mathbb{E}[\zeta_s^2]\text{)}$$

Linear TD(0) Analysis – Mean-path

Claim: $\|\theta_T - \theta^*\|^2 \leq \left(1 - \frac{(1-\gamma)^2\,\lambda}{8}\right)^T \|\theta_0 - \theta^*\|^2$.

Proof: We have proven (1) $\langle \bar{g}(\theta), \theta^* - \theta\rangle \geq (1-\gamma)\,\|v_\theta - v_{\theta^*}\|_D^2$ and (2) $\|\bar{g}(\theta)\| \leq 2\sqrt{2}\,\|v_\theta - v_{\theta^*}\|_D$.

$$
\begin{aligned}
\|\theta_{t+1} - \theta^*\|^2 &= \|\theta_t + \alpha\bar{g}(\theta) - \theta^*\|^2 = \|\theta_t - \theta^*\|^2 + 2\alpha\langle \bar{g}(\theta_t), \theta_t - \theta^*\rangle + \alpha^2\,\|\bar{g}(\theta_t)\|^2 \\
&\leq \|\theta_t - \theta^*\|^2 - 2\alpha\,(1-\gamma)\,\|v_{\theta_t} - v_{\theta^*}\|_D^2 + 8\,\alpha^2\,\|v_{\theta_t} - v_{\theta^*}\|_D^2 \\
&\leq \|\theta_t - \theta^*\|^2 - \frac{(1-\gamma)^2}{8}\,\|v_{\theta_t} - v_{\theta^*}\|_D^2 && \left(\text{Setting } \alpha = \tfrac{1-\gamma}{8}\right) \\
&= \|\theta_t - \theta^*\|^2 - \frac{(1-\gamma)^2}{8}\,\|\theta_t - \theta^*\|_\Sigma^2 && \left(\text{Since } \|v_\theta - v_{\theta^*}\|_D^2 = \|\theta - \theta^*\|_\Sigma^2\right) \\
&\leq \|\theta_t - \theta^*\|^2 - \lambda_{\min}[\Sigma]\,\frac{(1-\gamma)^2}{8}\,\|\theta_t - \theta^*\|^2 \\
\|\theta_{t+1} - \theta^*\|^2 &\leq \left(1 - \frac{(1-\gamma)^2\,\lambda}{8}\right)\|\theta_t - \theta^*\|^2 && \left(\text{Since } \lambda = \lambda_{\min}[\Sigma]\right) \\
\implies \|\theta_T - \theta^*\|^2 &\leq \left(1 - \frac{(1-\gamma)^2\,\lambda}{8}\right)^T \|\theta_0 - \theta^*\|^2 \quad \square && (\text{Recursing from } t = 0 \text{ to } T-1)
\end{aligned}
$$

Linear TD(0) Analysis – IID

The previous analysis requires $\bar{g}(\theta) = \mathbb{E}_{s\sim\omega}\mathbb{E}_{s'\sim P(\cdot|s)}\left[r(s,\pi(s)) + \gamma\langle\theta,\phi(s')\rangle - \langle\theta,\phi(s)\rangle\right]\phi(s)$.

Since we do not have access to the expectation, we will adapt the previous proof.

We will assume that (s_t, s_{t+1}) are sampled i.i.d. from the stationary distribution, i.e. $s_t \sim \omega$ and $s_{t+1} \sim P(\cdot|s_t) \implies \Pr[s_t = s, s_{t+1} = s'] = \omega(s)\,P(s'|s)$. Hence, taking the expectation over the randomness in s_t, s_{t+1}, we have that for all t and θ,

$$\begin{aligned}\mathbb{E}[g_t(\theta)] &= \mathbb{E}_{s_t,s_{t+1}}[[r(s_t,\pi(s_t)) + \gamma\langle\theta,\phi(s_{t+1})\rangle - \langle\theta,\phi(s_t)\rangle]\,\phi(s_t)] \\ &= \sum_{s,s'}[r(s,\pi(s)) + \gamma\langle\theta,\phi(s')\rangle - \langle\theta,\phi(s)\rangle]\,\phi(s)\,\Pr[s_t = s, s_{t+1} = s'] = \bar{g}(\theta)\end{aligned}$$

Similar to the previous proofs, we will rely on two important properties for $g_t(\theta)$. For a fixed t and θ independent of the randomness in (s_t, s_{t+1}),

(1) $\mathbb{E}\left[\langle g_t(\theta), \theta^* - \theta\rangle\right] = \langle\bar{g}(\theta), \theta^* - \theta\rangle \geq (1-\gamma)\left\|v_\theta - v_{\theta^*}\right\|_D^2$.
(2) $\mathbb{E}[\|g_t(\theta)\|^2] \leq 2\sigma^2 + 8\left\|v_\theta - v_{\theta^*}\right\|_D^2$ where $\sigma^2 := \mathbb{E}_{s_t,s_{t+1}}\left\|g_t(\theta^*)\right\|^2$ is the variance in $g_t(\theta^*)$.
(Prove in Assignment 3!)

Linear TD(0) Analysis – IID

Claim: Assuming (s_t, s_{t+1}) are sampled i.i.d from the stationary distribution, the update $\theta_{t+1} = \theta_t + \alpha_t\, g_t(\theta)$ with $\alpha_t = \frac{1}{\sqrt{T}}$ has the following convergence,

$$\mathbb{E}\left\|v_{\bar{\theta}_T} - v_{\theta^*}\right\|_D^2 \leq \frac{8\,\left\|\theta_0 - \theta^*\right\|^2}{(1-\gamma)^2\sqrt{T}} + \frac{\sigma^2}{4\sqrt{T}},$$

where the expectation is w.r.t. $\{s_t, s_{t+1}\}_{t=0}^{T-1}$ and $\bar{\theta}_T := \frac{\sum_{t=0}^{T-1}\theta_t}{T}$ is the average iterate.

Proof: We have proved that (1) $\mathbb{E}\left[\langle g_t(\theta), \theta^* - \theta\rangle\right] \geq (1-\gamma)\,\|v_\theta - v_{\theta^*}\|_D^2$ and (2) $\mathbb{E}[\|g_t(\theta)\|^2] \leq 2\sigma^2 + 8\,\|v_\theta - v_{\theta^*}\|_D^2$. Proceeding similar to the previous proof,

$$\|\theta_{t+1} - \theta^*\|^2 = \|\theta_t - \theta^*\|^2 + 2\alpha_t\langle g_t(\theta_t), \theta_t - \theta^*\rangle + \alpha_t^2\,\|g_t(\theta)\|^2$$

Taking expectation w.r.t the randomness at iteration t

$$\begin{aligned}\mathbb{E}\,\|\theta_{t+1} - \theta^*\|^2 &= \|\theta_t - \theta^*\|^2 + 2\alpha_t\,\mathbb{E}[\langle g_t(\theta_t), \theta_t - \theta^*\rangle] + \alpha_t^2\,\mathbb{E}\,\|g_t(\theta)\|^2 \\ &\leq \|\theta_t - \theta^*\|^2 - 2\alpha_t\,(1-\gamma)\,\|v_{\theta_t} - v_{\theta^*}\|_D^2 + \alpha_t^2\,\mathbb{E}\,\|g_t(\theta)\|^2\end{aligned}$$

(Using Property (1))

Linear TD(0) Analysis – IID

We have shown that $\mathbb{E}\left\|\theta_{t+1}-\theta^*\right\|^2 \leq \left\|\theta_t-\theta^*\right\|^2 - 2\alpha_t\,(1-\gamma)\,\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2 + \alpha_t^2\,\mathbb{E}\left\|g_t(\theta)\right\|^2$. Using Property (2),

$$\mathbb{E}\left\|\theta_{t+1}-\theta^*\right\|^2 \leq \left\|\theta_t-\theta^*\right\|^2 - 2\alpha_t\,(1-\gamma)\,\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2 + \alpha_t^2\,\left[2\sigma^2 + 8\,\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2\right]$$

$$\leq \left\|\theta_t-\theta^*\right\|^2 - \alpha_t\,(1-\gamma)\,\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2 + 2\alpha_t^2\sigma^2 \quad \left(\text{For } \alpha_t \leq \tfrac{1-\gamma}{8}\right)$$

$$\implies (1-\gamma)\,\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2 \leq \frac{\mathbb{E}[\left\|\theta_t-\theta^*\right\|^2 - \left\|\theta_{t+1}-\theta^*\right\|^2]}{\alpha_t} + 2\alpha_t\,\sigma^2$$

Using constant step-size $\alpha_t = \frac{1-\gamma}{8\sqrt{T}}$, and taking expectation w.r.t the randomness in iterations 0 to $T-1$,

$$(1-\gamma)\,\mathbb{E}\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2 \leq \mathbb{E}\left[\frac{\left\|\theta_t-\theta^*\right\|^2 - \left\|\theta_{t+1}-\theta^*\right\|^2}{\alpha_t}\right] + 2\alpha_t\,\sigma^2$$

$$\leq \frac{8\sqrt{T}}{1-\gamma}\,\mathbb{E}\left[\left\|\theta_t-\theta^*\right\|^2 - \left\|\theta_{t+1}-\theta^*\right\|^2\right] + \frac{\sigma^2\,(1-\gamma)}{4\sqrt{T}}$$

Linear TD(0) Analysis – IID

Recall $(1-\gamma)\,\mathbb{E}\left\|v_{\theta_t} - v_{\theta^*}\right\|_D^2 \leq \frac{8\sqrt{T}}{1-\gamma}\,\mathbb{E}\left[\left\|\theta_t - \theta^*\right\|^2 - \left\|\theta_{t+1} - \theta^*\right\|^2\right] + \frac{\sigma^2\,(1-\gamma)}{4\sqrt{T}}$. Summing from $t = 0$ to $T-1$,

$$(1-\gamma)\sum_{t=0}^{T-1}\mathbb{E}\left\|v_{\theta_t} - v_{\theta^*}\right\|_D^2 \leq \frac{8\sqrt{T}}{1-\gamma}\,\left\|\theta_0 - \theta^*\right\|^2 + \frac{\sigma^2\,(1-\gamma)\,\sqrt{T}}{4}$$

$$\implies \frac{\sum_{t=0}^{T-1}\mathbb{E}\left\|v_{\theta_t} - v_{\theta^*}\right\|_D^2}{T} \leq \frac{8\,\left\|\theta_0 - \theta^*\right\|^2}{(1-\gamma)^2\,\sqrt{T}} + \frac{\sigma^2}{4\,\sqrt{T}} \qquad \text{(Dividing by } (1-\gamma)\,T\text{)}$$

Using Jensen's inequality,

$$\mathbb{E}\left\|v_{\bar{\theta}_T} - v_{\theta^*}\right\|_D^2 \leq \frac{8\,\left\|\theta_0 - \theta^*\right\|^2}{(1-\gamma)^2\,\sqrt{T}} + \frac{\sigma^2}{4\,\sqrt{T}} \quad \square$$

By using more complicated step-size sequences, we can also show convergence for the last-iterate θ_T (similar to the previous proofs).

Linear TD(0) Analysis – Markovian

The previous analysis assumes that (s_t, s_{t+1}) are sampled i.i.d from the stationary distribution. However, (s_t, s_{t+1}) are gathered from a single trajectory of the Markov chain induced by policy π.

Hence, the samples are correlated and assuming that they are i.i.d is not valid. However, under certain standard assumptions, we can adapt the previous proof.

Assumption: The underlying Markov chain is "fast-mixing" i.e. for constants $m > 0$ and $\rho \in (0, 1)$, and all t, if $\mathsf{TV}(P, Q)$ is the total variation distance between distributions P, Q, then,

$$\sup_s \mathsf{TV}(\Pr^\pi[s_t | s_0 = s], \omega) \leq m\,\rho^t$$

i.e. the distribution over states approaches the stationary distribution exponentially fast.

Define $\tau_{\mathsf{mix}}(\epsilon) = \min\{t | \rho^t \leq \epsilon\}$ as the mixing time of the Markov chain.

Linear TD(0) Analysis – Markovian

Projected linear TD(0) update: $\theta_{t+1} = \text{Proj}\left[\theta_{t+1} + \alpha_t g_t(\theta)\right]$. The projection is onto the ball $\mathcal{B} = \{\theta | \, \|\theta\| \leq R\}$ where R is an upper-bound on $\|\theta^*\|$.

Claim: Assuming fast-mixing of the underlying Markov chain, Projected linear TD(0) with $\alpha_t = \frac{1}{\sqrt{T}}$ has the following convergence:

$$\mathbb{E} \left\| v_{\bar{\theta}_T} - v_{\theta^*} \right\|_D^2 \leq O\left(\frac{\left\|\theta_0 - \theta^*\right\|^2}{\sqrt{T}} + \frac{(1+2R)^2 \left(1 + \tau_{\text{mix}}\left(1/\sqrt{T}\right)\right)}{\sqrt{T}} \right).$$

- Intuitively, every cycle of $\tau_{\text{mix}}(\cdot)$ samples provides as much information as a single independent sample from the stationary distribution.
- If (s_t, s_{t+1}) were sampled i.i.d. from ω, $\tau_{\text{mix}}(\cdot) = 0$ and we would obtain the IID result.
- The proof is similar to the i.i.d case except that it needs to carefully handle correlations and bound $\mathbb{E}\left[\langle g_t(\theta_t) - \bar{g}(\theta_t), \theta_t - \theta^* \rangle\right] \neq 0$.
- For more details, refer to [BRS18, Section 8].

References i

- Jalaj Bhandari, Daniel Russo, and Raghav Singal, *A finite time analysis of temporal difference learning with linear function approximation*, Conference on learning theory, PMLR, 2018, pp. 1691–1692.
- Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, MIT press, 2018.