CMPT 419/983: Theoretical Foundations of Reinforcement Learning

Lecture 7

Sharan Vaswani

October 20, 2023

Recap

- **Monte-Carlo estimation for policy evaluation**
  - Generate trajectory $\tau = (s_0, a_0, s_1, \ldots)$ and calculate $R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$.
  - Generate m trajectories $\{\tau_i\}_{i=1}^m$ and calculate $\hat{v} := \frac{\sum_{i=1}^m R(\tau_i)}{m}$ as an approximation to $v^\pi(s_0)$.
  - Using Monte-Carlo estimation with $m = \frac{\ln(2/\delta)}{2\epsilon^2 (1-\gamma)^2}$ 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 $1 - \delta$.
- **Linear TD(0)**:
  - *Assumption*: For the fixed policy π being evaluated, there exists a unique θ^* such that $v^\pi = \Phi\theta^* = v_{\theta^*}$.
  - *Update*: $\theta_{t+1} = \theta_t + \alpha_t\, g_t(\theta_t)$ where $g_t(\theta) = [r_t + \gamma\langle\theta, \phi(s_{t+1})\rangle - \langle\theta, \phi(s_t)\rangle]\ \phi(s_t)$.
  - *Mean-path TD(0)*: $\theta_{t+1} = \theta_t + \alpha\, \bar{g}(\theta)$ where $\bar{g}(\theta) := \mathbb{E}_{s\sim\omega}\mathbb{E}_{s'\sim P(\cdot|s)}\,[r(s,\pi(s)) + \gamma\langle\theta, \phi(s')\rangle - \langle\theta, \phi(s)\rangle]\ \phi(s)$ and ω is the stationary distribution.
  - By using an analysis similar to GD, we showed that Mean-path TD(0) converges to θ^*at a linear rate.

Linear TD(0) Analysis – IID

Mean-path TD 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) $\mathbb{E}\left[\langle g_t(\theta), \theta^* - \theta\rangle\right] = \langle\bar{g}(\theta), \theta^* - \theta\rangle \geq (1-\gamma)\;\|v_\theta - v_{\theta^*}\|_D^2$.

(2) $\mathbb{E}[\|g_t(\theta)\|^2] \leq 2\sigma^2 + 8\;\|v_\theta - v_{\theta^*}\|_D^2$ where $\sigma^2 := \mathbb{E}_{s_t,s_{t+1}}\|g_t(\theta^*)\|^2$ 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 $\alpha_t = \frac{1-\gamma}{8\sqrt{T}}$ has the following convergence,

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

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),

$$\begin{aligned}
\mathbb{E}\left\|\theta_{t+1}-\theta^*\right\|^2 &\leq \left\|\theta_t-\theta^*\right\|^2 - 2\alpha_t(1-\gamma)\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2 + \alpha_t^2\left[2\sigma^2 + 8\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2\right] \\
&\leq \left\|\theta_t-\theta^*\right\|^2 - \alpha_t(1-\gamma)\left\|v_{\theta_t}-v_{\theta^*}\right\|_D^2 + 2\alpha_t^2\sigma^2 \quad (\text{For } \alpha_t \leq \tfrac{1-\gamma}{8})
\end{aligned}$$

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

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

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

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 $\tau_{\text{mix}}(\epsilon) = \min\{t | \rho^t \leq \epsilon\}$ 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:

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

- 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].

Interpolating between TD(0) and Monte-Carlo

- Recall the derivation of TD(0): (i) use the Bellman equation: $v^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\,\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\,[r(s,a) + \gamma v^\pi(s')]$, (ii) sampling a from $\pi(\cdot|s)$, $s' \sim \mathcal{P}(\cdot|s,a)$ gives $\hat{v}^\pi(s) = r(s,a) + \gamma\, v^\pi(s')$, (iii) using estimate $\hat{v}^\pi(s')$ in place of $v^\pi(s')$ (bootstrapping) results in the TD(0) update.
- Instead, (i) use the Bellman equation for $v^\pi(s')$, meaning that: $\hat{v}^\pi(s) = r(s,a) + \gamma\, v^\pi(s_1) = r(s,a) + \gamma\,\mathbb{E}_{\mathbf{a_1}\sim\pi(\cdot|\mathbf{s_1})}\,\mathbb{E}_{\mathbf{s_2}\sim\mathcal{P}(\cdot|\mathbf{s_1},\mathbf{a_1})}\,[r(s_1,a_1) + \gamma v^\pi(s_2)]$, (ii) sampling a_1 from $\pi(\cdot|s_1)$, $s_2 \sim \mathcal{P}(\cdot|s_1,a_1)$ gives $\hat{v}^\pi(s) = r(s,a) + \gamma\, r(s_1,a_1) + \gamma^2\, v^\pi(s_2)$, (iii) using estimate $\hat{v}^\pi(s_2)$ in place of $v^\pi(s_2)$ (bootstrapping) results in the TD(1) update.
- Similarly, we can derive $TD(n)$ updates for $n \geq 0$, $\hat{v}^\pi(s) = \sum_{t=0}^{n} \gamma^t r_t + \gamma^{n+1}\hat{v}^\pi(s_{n+1})$.
- As $n \to \infty$, we get the update $\hat{v}^\pi(s) = \sum_{t=0}^{\infty} \gamma^t r_t$ corresponding to Monte-Carlo estimation.
- TD(0) has a higher bias, lower variance, while Monte-Carlo estimation has lower bias, higher variance. As n increases, the bias (proportional to γ^n) decays exponentially fast.
- For more details, refer to [SB18, Chapter 7].

Approximate Policy Iteration

Approximate Policy Iteration

For approximate policy iteration (without access to $\mathcal{P}$, r), we will make use of q functions.

State-action value function for policy π: $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that for $s \in \mathcal{S}$, $a \in \mathcal{A}$,

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

i.e. $q^\pi(s, a)$ corresponds to the cumulative discounted reward obtained when starting at state s, taking action a and following policy π from then on. (See Assignment 2 for details)

Algorithm Approximate Policy Iteration

1: **Input**: MDP $M = (\mathcal{S}, \mathcal{A}, \rho)$, π_0.
2: **for** $k = 0 \to K$ **do**
3: **Policy Evaluation**: Compute the estimate $\hat{q}^{\pi_k}$ (for example, using TD, Monte-Carlo).
4: **Policy Improvement**: $\forall s$, $\pi_{k+1}(s) = \arg\max_a \hat{q}^{\pi_k}(s, a)$.
5: **end for**

First, we will study how the error in estimating the q function affects $v^{\pi_{K+1}}$, the value function corresponding to the policy output by the algorithm.

Policy Improvement with Errors

Claim: For Markov policies $\pi, \tilde{\pi}$, define $\hat{q} \in \mathbb{R}^{S\times A}$ as an estimate of q^π s.t. $\hat{q}^\pi = q^\pi + \epsilon$ for some $\epsilon \in \mathbb{R}^{S\times A}$. If $\tilde{\pi}$ is the greedy policy w.r.t $\hat{q}^\pi$, then,

$$\left\| v^* - v^{\tilde{\pi}} \right\|_\infty \leq \gamma \left\| v^* - v^\pi \right\|_\infty + \frac{1}{1-\gamma} \left\| \epsilon \right\|_\infty$$

Proof: Since π^* is optimal, using the fundamental theorem, $\mathcal{T} v^* = v^* = \mathcal{T}_{\pi^*} v^*$. Since $v^{\tilde{\pi}}$ is the fixed point of $\mathcal{T}_{\tilde{\pi}}$, $v^{\tilde{\pi}} = \mathcal{T}_{\tilde{\pi}} v^{\tilde{\pi}}$. Hence,

$$\begin{aligned}
v^* - v^{\tilde{\pi}} &= \mathcal{T}_{\pi^*} v^* - \mathcal{T}_{\tilde{\pi}} v^{\tilde{\pi}} \\
&= \mathcal{T}_{\pi^*} v^* - \mathcal{T}_{\pi^*} v^\pi + \mathcal{T}_{\pi^*} v^\pi - \mathcal{T}_{\tilde{\pi}} v^\pi + \mathcal{T}_{\tilde{\pi}} v^\pi - \mathcal{T}_{\tilde{\pi}} v^{\tilde{\pi}} && (\text{Add/subtract } \mathcal{T}_{\pi^*} v^\pi \text{ and } \mathcal{T}_{\tilde{\pi}} v^\pi) \\
&= [[\mathbf{r}_{\pi^*} + \gamma \mathbf{P}_{\pi^*} v^*] - [\mathbf{r}_{\pi^*} + \gamma \mathbf{P}_{\pi^*} v^\pi]] + \mathcal{T}_{\pi^*} v^\pi - \mathcal{T}_{\tilde{\pi}} v^\pi + \left[[\mathbf{r}_{\tilde{\pi}} + \gamma \mathbf{P}_{\tilde{\pi}} v^\pi] - [\mathbf{r}_{\tilde{\pi}} + \gamma \mathbf{P}_{\tilde{\pi}} v^{\tilde{\pi}}]\right] \\
& && (\text{Since } \mathcal{T}_\pi v = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v) \\
&= \gamma \mathbf{P}_{\pi^*} [v^* - v^\pi] + \mathcal{T}_{\pi^*} v^\pi - \mathcal{T}_{\tilde{\pi}} v^\pi + \gamma \mathbf{P}_{\tilde{\pi}} [v^\pi - v^{\tilde{\pi}}] \\
&\leq \gamma \mathbf{P}_{\pi^*} [v^* - v^\pi] + \mathcal{T} v^\pi - \mathcal{T}_{\tilde{\pi}} v^\pi + \gamma \mathbf{P}_{\tilde{\pi}} [v^\pi - v^{\tilde{\pi}}] && (\text{Since } \mathcal{T}_{\pi^*} v^\pi \leq \mathcal{T} v^\pi) \\
&= \gamma \mathbf{P}_{\pi^*} [v^* - v^\pi] + \delta + \gamma \mathbf{P}_{\tilde{\pi}} [v^\pi - v^{\tilde{\pi}}] && (\text{Define } \delta := \mathcal{T} v^\pi - \mathcal{T}_{\tilde{\pi}} v^\pi)
\end{aligned}$$

Policy Improvement with Errors

Recall that $v^* - v^{\tilde{\pi}} \leq \gamma \mathbf{P}_{\pi^*}[v^* - v^{\pi}] + \delta + \gamma \mathbf{P}_{\tilde{\pi}}[v^{\pi} - v^{\tilde{\pi}}]$, where $\delta = \mathcal{T} v^{\pi} - \mathcal{T}_{\tilde{\pi}} v^{\pi}$.

$$
\begin{aligned}
v^{\pi} - v^{\tilde{\pi}} &= (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} [v^{\pi} - \mathcal{T}_{\tilde{\pi}} v^{\pi}] && \text{(Value Difference Lemma)} \\
&\leq (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} [\mathcal{T} v^{\pi} - \mathcal{T}_{\tilde{\pi}} v^{\pi}] = (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} \delta \\
& && \text{(Since } v^{\pi} = \mathcal{T}_{\pi} v^{\pi} \leq \mathcal{T} v^{\pi} \text{ and for } u \leq w,\ (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} u \leq (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} w) \\
\implies v^* - v^{\tilde{\pi}} &\leq \gamma \mathbf{P}_{\pi^*}[v^* - v^{\pi}] + \delta + \gamma \mathbf{P}_{\tilde{\pi}} \left((I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} \delta\right) \\
&= \gamma \mathbf{P}_{\pi^*}[v^* - v^{\pi}] + \left[I + \gamma \mathbf{P}_{\tilde{\pi}} (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1}\right] \delta = \gamma \mathbf{P}_{\pi^*}[v^* - v^{\pi}] + (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} \delta \\
& && \text{(Since } I + \gamma \mathbf{P}_{\tilde{\pi}} (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} = (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1}) \\
\left\|v^{\pi} - v^{\tilde{\pi}}\right\|_{\infty} &\leq \left\|\gamma \mathbf{P}_{\pi^*}[v^* - v^{\pi}] + (I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} \delta\right\|_{\infty} && \text{(Taking norms on both sides)} \\
&\leq \left\|\gamma \mathbf{P}_{\pi^*}[v^* - v^{\pi}]\right\|_{\infty} + \left\|(I - \gamma \mathbf{P}_{\tilde{\pi}})^{-1} \delta\right\|_{\infty} && \text{(Triangle inequality)} \\
\implies \left\|v^{\pi} - v^{\tilde{\pi}}\right\|_{\infty} &\leq \gamma \left\|v^* - v^{\pi}\right\|_{\infty} + \frac{1}{1-\gamma} \left\|\delta\right\|_{\infty} && \text{(Using the Neumann series)}
\end{aligned}
$$

Policy Improvement with Errors

Recall that $\left\|v^* - v^{\tilde{\pi}}\right\|_\infty \leq \gamma \left\|v^* - v^\pi\right\|_\infty + \frac{1}{1-\gamma} \|\delta\|_\infty$ where $\delta = \mathcal{T} v^\pi - \mathcal{T}_{\tilde{\pi}} v^\pi$.
In order to bound $\|\delta\|_\infty$, recall the following definitions from Assignment 2: $\mathcal{M}_\pi : \mathbb{R}^{S\times A} \to \mathbb{R}^S$, $\mathbb{P} : \mathbb{R}^S \to \mathbb{R}^{S\times A}$ and $\mathcal{M} : \mathbb{R}^{S\times A} \to \mathbb{R}^S$, such that for $u \in \mathbb{R}^{S\times A}$ and $w \in \mathbb{R}^S$,

$$(\mathcal{M}_\pi u)(s) = \sum_a \pi(a|s)\, u(s,a) \; ; \; (\mathbb{P} w)(s,a) = \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)\, w(s') \; ; \; (\mathcal{M} u)(s) = \max_{a\in\mathcal{A}} u(s,a)$$

$$\begin{aligned}
\mathcal{T} v^\pi &\geq \mathcal{T}_{\tilde{\pi}}\, v^\pi && \text{(Since } \mathcal{T} \text{ is the Bellman optimality operator)} \\
&= \mathcal{M}_{\tilde{\pi}}\,(r + \gamma \mathbb{P}\, v^\pi) && \text{(Since } \mathcal{T}_\pi w = \mathcal{M}_\pi (r + \gamma\, \mathbb{P}\, w) \text{ for all } w \in \mathbb{R}^S) \\
&= \mathcal{M}_{\tilde{\pi}} q^\pi && \text{(By definition of } q^\pi) \\
&= \mathcal{M}_{\tilde{\pi}} [\hat{q}^\pi - \epsilon] && \text{(Since } q^\pi = \hat{q}^\pi - \epsilon) \\
&= \mathcal{M}_{\tilde{\pi}} \hat{q}^\pi - \mathcal{M}_{\tilde{\pi}} \epsilon && (\mathcal{M}_\pi \text{ is a linear operator}) \\
&= \mathcal{M} \hat{q}^\pi - \mathcal{M}_{\tilde{\pi}} \epsilon && \text{(Since } \tilde{\pi} \text{ is greedy w.r.t } \hat{q}^\pi) \\
&= \mathcal{M} (q^\pi + \epsilon) - \mathcal{M}_{\tilde{\pi}} \epsilon && \text{(Since } \hat{q}^\pi = q^\pi + \epsilon) \\
\implies \mathcal{T} v^\pi \geq \mathcal{T}_{\tilde{\pi}} v^\pi &\geq \mathcal{M}(q^\pi - \|\epsilon\|_\infty 1) - \mathcal{M}_{\tilde{\pi}} \epsilon && \text{(Since } \epsilon \geq -\|\epsilon\|_\infty 1 \text{ and } \mathcal{M} \text{ is monotone)}
\end{aligned}$$

Policy Improvement with Errors

Recall that $\mathcal{T} v^\pi \geq \mathcal{T}_{\tilde{\pi}} v^\pi \geq \mathcal{M}(q^\pi - \|\epsilon\|_\infty 1) - \mathcal{M}_{\tilde{\pi}} \epsilon$

$$\mathcal{T} v^\pi \geq \mathcal{T}_{\tilde{\pi}} v^\pi \geq \mathcal{M} q^\pi - \|\epsilon\|_\infty 1 - \mathcal{M}_{\tilde{\pi}} \epsilon$$

(Since $\mathcal{M}$ is non-expansive, $\|\mathcal{M}(q^\pi - \|\epsilon\|_\infty 1) - \mathcal{M} q^\pi\|_\infty \leq \|\epsilon\|_\infty$)

$$\geq \mathcal{M} q^\pi - \|\epsilon\|_\infty 1 - \|\epsilon\|_\infty 1$$

(Since $\mathcal{M}_\pi$ is non-expansive, $\|\mathcal{M}_\pi(\|\epsilon\|_\infty 1)\|_\infty \leq \|\epsilon\|_\infty$)

$$= \mathcal{M} q^\pi - 2\|\epsilon\|_\infty 1 = \mathcal{M}(r + \gamma \mathbb{P} v^\pi) - 2\|\epsilon\|_\infty 1 = \mathcal{T} v^\pi - 2\|\epsilon\|_\infty 1$$

(By def. of q and since $\mathcal{T} u = \mathcal{M}(r + \gamma \mathbb{P} u)$)

$$\implies \mathcal{T} v^\pi \geq \mathcal{T}_{\tilde{\pi}} v^\pi \geq \mathcal{T} v^\pi - 2\|\epsilon\|_\infty 1$$

$$\implies \delta = \mathcal{T} v^\pi - \mathcal{T}_{\tilde{\pi}} v^\pi \leq 2\|\epsilon\|_\infty 1 \implies \|\delta\|_\infty \leq 2\|\epsilon\|_\infty \quad \text{(Taking norms on both sides)}$$

Putting everything together,

$$\left\| v^* - v^{\tilde{\pi}} \right\|_\infty \leq \gamma \|v^* - v^\pi\|_\infty + \frac{2\|\epsilon\|_\infty}{1 - \gamma} \quad \square$$

Approximate Policy Iteration

For approximate policy iteration, $\pi_{k+1}(s) = \arg\max_a \hat{q}^{\pi_k}(s,a)$, i.e. π_{k+1} is greedy w.r.t $\hat{q}^{\pi_k}$.

For each iteration $k \in [K]$, if we can estimate $\hat{q}^{\pi_k}$ such that $\hat{q}^{\pi_k} = q^{\pi_k} + \epsilon_k$, then, by using the previous claim,

$$\left\| v^* - v^{\pi_{k+1}} \right\|_\infty \leq \gamma \left\| v^* - v^{\pi_k} \right\|_\infty + \frac{2 \left\| \epsilon_k \right\|_\infty}{1-\gamma}$$

Claim: If the policy evaluation error at iteration k is controlled s.t. $\hat{q}^{\pi_k} = q^{\pi_k} + \epsilon_k$, then, approximate policy iteration has the following convergence,

$$\left\| v^{\pi_{K+1}} - v^* \right\|_\infty \leq \gamma^K \left\| v^{\pi_0} - v^* \right\|_\infty + \frac{2 \max_{k \in \{0,\ldots,K-1\}} \left\| \epsilon_k \right\|_\infty}{(1-\gamma)^2}$$

Prove in Assignment 3!

- This generalizes the claim for exact policy iteration (corresponding to $\epsilon_k = 0$) in Lecture 5.
- The convergence is only to a *neighbourhood* of v^* and the error ϵ is amplified by $\frac{2}{(1-\gamma)^2}$.
- This error amplification is tight for approximate policy iteration. See Csaba's notes for the formal lower-bound.

Policy Evaluation for Approximate Policy Iteration

For Approximate Policy Iteration to be effective, we need to control the policy evaluation error in each iteration. We have seen that,

- Without any structural assumption, Monte-Carlo estimation required rolling out trajectories from each state, making it sample inefficient.
- TD(0) can exploit the linear assumption in an efficient manner.
- However, for TD(0) to have theoretical guarantees, we needed to make assumptions about the ergodicity (can reach all states) and mixing of the underlying Markov chain. This side-steps the important issue of exploration in MDPs.
- In order to handle exploration and still be sample-efficient, we will use Monte-Carlo estimation with a linear assumption on $q^{\pi}(s, a)$ along with experimental design. This will enable us to control the policy evaluation error in theoretically principled manner.

Policy Evaluation for Approximate Policy Iteration

Assumption: Have access to features $\Phi \in \mathbb{R}^{SA \times d}$, such that the q functions for policy π are $\varepsilon_{\mathbf{b}}$-close to the span of Φ. Consider a fixed π. There exists a θ^* s.t.

$$\max_{(s,a)} |q^\pi(s,a) - \langle \theta^*, \phi(s,a) \rangle| \leq \varepsilon_{\mathbf{b}}$$

- Given a "good" estimate of $\hat{\theta}$, we can estimate $q^\pi(s,a)$ by $\hat{q}^\pi(s,a) = \langle \hat{\theta}, \phi(s,a) \rangle$.

Algorithm Idea:

- Choose a set $\mathcal{C} \subset \mathcal{S} \times \mathcal{A}$, and for each $(s,a) \in \mathcal{C}$, rollout trajectories (truncated to horizon H) starting from state s, taking action a and then following policy π.
- For each trajectory τ, calculate the cumulative discounted reward $\sum_{t=0}^{H} \gamma^t r_t$.
- For each $(s,a) \in \mathcal{C}$, run m trajectories and use the average as an estimate for $q^\pi(s,a)$.
- Define $z := (s,a)$ and the corresponding empirical mean as $\hat{R}(z)$. For weights $\zeta \in \Delta_{|\mathcal{C}|}$ (to be determined later), compute the estimate $\hat{\theta}$ by weighted linear regression:

$$\hat{\theta} := \arg\min_{\theta} \frac{1}{2} \sum_{z \in \mathcal{C}} \zeta(z) \left[\langle \theta, \phi(z) \rangle - \hat{R}(z) \right]^2$$

Policy Evaluation for Approximate Policy Iteration

Similar to the proof in Lecture 6, we have the following result that shows that the error in estimating $q^\pi(z)$ for $z \in \mathcal{C}$ can be controlled.

Claim: Using $m = \frac{\ln(2\,|\mathcal{C}|/\delta)}{2\varepsilon_{\mathbf{s}}^{\mathbf{2}}\,(1-\gamma)^{\mathbf{2}}}$ trajectories with $H \geq \frac{\ln(1/\varepsilon_{\mathbf{s}}\,(1-\gamma))}{\ln(1/\gamma)}$ guarantees that $|\hat{R}(z) - q^\pi(z)| \leq \varepsilon_{\mathbf{s}}$ with probability $1-\delta$ for all $z \in \mathcal{C}$.

For the policy evaluation to be effective,

(i) We require control over the *generalization error*, the estimation error for $z \notin \mathcal{C}$.

(ii) For computational efficiency, we want that $|\mathcal{C}|$ not depend on $|\mathcal{S}|$.

Next class, we will see how to choose $\mathcal{C}$ such that both (i) and (ii) are satisfied.

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.