Lecture 10: Approximate Dynamic Programming

Diana Borsa

February 2020, UCL

This Lecture

- Last lectures:
  - MDP, DP, Model-free Prediction, Model-free Control
  - Bellman equations and their corresponding operators.
  - RL under function approximation.
- This lecture:
  - Revisit the framework of Approximate Dynamic Programming.
  - Under the 2 sources of error (estimation + function approximation), what can we say about resulting estimates?
- Next lectures: (more) approximate versions of these paradigms, mainly in the absence of perfect knowledge of the environment + (deep) neural networks parametrisation.

Preliminaries
(Quick Recap)

(Reminder)The Bellman Optimality Operator

Definition (Bellman Optimality Operator $T^*_{\mathcal{V}}$)

Given an MDP, $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, let $\mathcal{V} \equiv \mathcal{V}_{\mathcal{S}}$ be the space of bounded real-valued functions over $\mathcal{S}$. We define, point-wise, the Bellman Expectation operator $T^*_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}$ as:

$$(T^*_{\mathcal{V}} f)(s) = \max_a \left[r(s, a) + \gamma \sum_{s'} p(s'|a, s) f(s') \right], \forall f \in \mathcal{V} \tag{1}$$

As a common convention we drop the index $\mathcal{V}$ and simply use $T^* = T^*_{\mathcal{V}}$

(Reminder)The Bellman Expectation Operator

Definition (Bellman Expectation Operator)

Given an MDP, $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, let $\mathcal{V} \equiv \mathcal{V}_{\mathcal{S}}$ be the space of bounded real-valued functions over $\mathcal{S}$. For any policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, we define, point-wise, the Bellman Expectation operator $T^{\pi}_{\mathcal{V}} : \mathcal{V} \rightarrow \mathcal{V}$ as:

$$(T^{\pi}_{\mathcal{V}} f)(s) = \sum_{a} \pi(s, a) \left[r(s, a) + \gamma \sum_{s'} p(s'|a, s) f(s') \right], \forall f \in \mathcal{V} \tag{2}$$

(Reminder) Dynamic Programming with Bellman Operators

Value Iteration

- Start with v_0.
- Update values: $v_{k+1} = T^* v_k$.

Policy Iteration

- Start with π_0.
- Iterate:
  - Policy Evaluation: v_{π_i}
    - (E.g. For instance, by iterating T^π: $v_k = T^{\pi_i} v_{k-1} \Rightarrow v_k \to v^{\pi_i}$ as $k \to \infty$)
  - Greedy Improvement: $\pi_{i+1} = \arg\max_a q_{\pi_i}(s, a)$

Approximate DP

- More often than not:
  - We won't know the underlying MDP.
    $\Rightarrow$ sampling/estimation error, as we don't have access to the true operators T^{π} (T^{*})
  - We won't be able to represent the value function exactly after each update.
    $\Rightarrow$ approximation error, as we approximate the true value functions within a (parametric) class (e.g. linear functions, neural nets, etc).
- Objective: Under the above conditions, come up with a policy π that is (close to) optimal.

Approximate Value Iteration
(+ friends)

(Reminder) Value Iteration

Value Iteration

- Start with v_0.
- Update values: $v_{k+1} = T^* v_k$.

As $k \to \infty$, $v_k \to_{\|.\|_\infty} v^*$. (Direct application for the Banach's Fixed Point theorem)

Approximate Value Iteration

Approximate Value Iteration

- Start with v_0.
- Update values: $v_{k+1} = \mathcal{A} T^* v_k$. $(v_{k+1} \approx T^* v_k)$
- Return control policy: $\pi_{k+1} = Greedy(v_{k+1})$

Question: As $k \to \infty$, $v_k \to_{\|.\|_\infty} v^*$? Generally **X**. But maybe we don't need to!

Good news: interested in the quality of π_n after n iterations: v_{π_n} (or q_{π_n})

Approximate Value Iteration (q-value version)

Approximate Value Iteration

- Start with q_0.
- Update values: $q_{k+1} = \mathcal{A} T^* q_k$. $(q_{k+1} \approx T^* q_k)$
- Return control policy: $\pi_{k+1} = \mathit{Greedy}(q_{k+1})$

Question: As $k \to \infty$, $q_k \to_{\|.\|_\infty} q^*$? Generally **X**.

Performance of AVI

Theorem (Bertsekas & Tsitsiklis, 1996)

Consider a MDP. And let q_k be the value function returned by AVI after k steps and let π_k be its corresponding greedy policy, then:

$$\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < n} \|T^* q_k - \mathcal{A} T^* q_k\|_\infty + \frac{2\gamma^{n+1}}{(1-\gamma)} \epsilon_0$$

where

$$\epsilon_0 = \|q^* - q_0\|_\infty$$

and T^* is the optimal Bellman operator associated with this MDP

Performance of AVI

Theorem (Bertsekas & Tsitsiklis, 1996)

Consider a MDP. And let q_k be the value function returned by AVI after k steps and let π_k be its corresponding greedy policy, then:

$$\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < n} \|T^* q_k - \mathcal{A} T^* q_k\|_\infty + \frac{2\gamma^{n+1}}{(1-\gamma)} \underbrace{\epsilon_0}_{\textit{(initial error)}}$$

where

$$\epsilon_0 = \|q^* - q_0\|_\infty$$

and T^* is the optimal Bellman operator associated with this MDP

Performance of AVI

Theorem (Bertsekas & Tsitsiklis, 1996)

Consider a MDP. And let q_k be the value function returned by AVI after k steps and let π_k be its corresponding greedy policy, then:

$$\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < n} \underbrace{\|T^* q_k - \mathcal{A} T^* q_k\|_\infty}_{\text{approximation error at iter. } k} + \frac{2\gamma^{n+1}}{(1-\gamma)} \underbrace{\epsilon_0}_{\text{(initial error)}}$$

where

$$\epsilon_0 = \|q^* - q_0\|_\infty$$

and T^* is the optimal Bellman operator associated with this MDP

Performance of AVI (Proof)

Statement: $\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0\leq k<n} \|T^* q_k - \mathcal{A}T^* q_k\|_\infty + \frac{2\gamma^{n+1}}{(1-\gamma)}\|q^* - q_0\|_\infty$

Proof.

Let's denote $\epsilon = \max_{0\leq k<n} \|T^* q_k - \mathcal{A}T^* q_k\|_\infty$. Then for all $k < n$:

$$\begin{aligned}
\|q^* - q_{k+1}\|_\infty &\leq \|q^* - T^* q_k\|_\infty + \|T^* q_k - q_{k+1}\|_\infty && (3)\\
&\leq \|T^* q^* - T^* q_k\|_\infty + \epsilon && (4)\\
&\leq \gamma\|q^* - q_k\|_\infty + \epsilon && (5)
\end{aligned}$$

Thus:

$$\begin{aligned}
\|q^* - q_k\|_\infty &\leq \gamma\|q^* - q_{k-1}\|_\infty + \epsilon && (6)\\
&\leq \gamma(\gamma\|q^* - q_{k-2}\|_\infty + \epsilon) + \epsilon && (7)\\
&\ldots\\
&\leq \gamma^k\|q^* - q_0\|_\infty + \epsilon(1 + \gamma + \cdots + \gamma^{K-1}) && (8)\\
&\leq \gamma^k\|q^* - q_0\|_\infty + \frac{1}{(1-\gamma)}\epsilon && (9)
\end{aligned}$$

Performance of AVI (Proof continued)

Statement: $\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < n} \|T^* q_k - \mathcal{A} T^* q_k\|_\infty + \frac{2\gamma^{n+1}}{(1-\gamma)} \|q^* - q_0\|_\infty$

Proof.

Let's denote $\epsilon = \max_{0 \leq k < n} \|T^* q_k - \mathcal{A} T^* q_k\|_\infty$. Then for all k, we have

$$\|q^* - q_k\|_\infty \leq \gamma^k \|q^* - q_0\|_\infty + \frac{1}{(1-\gamma)}\epsilon \tag{10}$$

Now recall, the performance of a greedy policy, π_k based on q_k:

$$\|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{1-\gamma} \|q^* - q_k\|_\infty \tag{11}$$

Combining the two results, we get the statement of the theorem. □

Performance of AVI: Breakdown

Statement:

$$\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0\leq k<n} \|T^* q_k - \mathcal{A}T^* q_k\|_\infty + \frac{2\gamma^{n+1}}{(1-\gamma)}\|q^* - q_0\|_\infty$$

Some implications:

- As $n \to \infty$, $\Rightarrow 2\gamma^n/(1-\gamma) \to 0$
- What if $q_0 = q^*$?

$$\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0\leq k<n} \|T^* q_k - \mathcal{A}T^* q_k\|_\infty$$

- Consider iteration 1: $q_1 = \mathcal{A}T^* q_0 = \mathcal{A}q^*$.In general $\Rightarrow \|q_1 - q_0\|_\infty > 0$.

Performance of AVI: Breakdown

Statement:

$$\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0\leq k<n} \|T^* q_k - \mathcal{A} T^* q_k\|_\infty + \underbrace{\frac{2\gamma^{n+1}}{(1-\gamma)} \|q^* - q_0\|_\infty}_{0 \text{ as } n \to \infty}$$

- Consider a hypothesis space $\mathcal{F}$.
- What if $\mathcal{A} = \Pi_\infty$ is the projection operator in L_∞:

$$\Pi_\infty g := \arg \inf_{f\in\mathcal{F}} \|g - f\|_\infty$$

- We obtain:

$$q_{k+1} = \Pi_\infty T^* q_k = \arg \inf_{f\in\mathcal{F}} \|T^* q_k - f\|_\infty$$

.

  - Note that $\mathcal{A} T^* = \Pi_\infty T^*$ is a contraction operator in L_∞.
  - Algorithm converges for its fixed point: $f = \Pi_\infty T^* f$
  - If $q^* \in \mathcal{F}$, the above will converge to q^*.

Some concrete instances of AVI

Fitted Q-iteration with Linear Approximation

Propose Algorithm:

$$q_{k+1} = \Pi_\infty T^* q_k = \arg \inf_{f \in \mathcal{F}} \| T^* q_k - f \|_\infty$$

- Consider a linear hypothesis space $\mathcal{F}_\phi = \{q_w(s,a) = w^T \phi(s,a) | \forall w \in B\}$.
- We obtain:

$$\begin{aligned} q_{k+1} &= \arg \inf_{f \in \mathcal{F}_\phi} \| T^* q_k - f \|_\infty && (12) \\ \Leftrightarrow w_{k+1} &= \arg \inf_{w \in B} \| T^* (w_k^T \phi) - w^T \phi \|_\infty && (13) \end{aligned}$$

- Potential problems:
  - P1: L_∞ minimisation typically hard to carry out efficiently.
  - P2: T^* is typically unknown and will be approximated as well.

Fitted Q-iteration with Linear Approximation

Proposals:

- **P1**: $L_\infty \to L_2$, wrt to a probability distribution μ over $\mathcal{S} \times \mathcal{A}$.

$$q_{k+1} = \arg\inf_{f \in \mathcal{F}} \| T^* q_k - f \|_\mu^2.$$

- **P2**: Sample to approximate T^*. (see previous lectures on Model-free control)
  - Sample $(S_t, A_t, R_{t+1}, S_{t+1}) \sim \mu, P$
  - Approximate $T^* q_k(S_t, A_t)$ by

$$Y_t = R_{t+1} + \gamma \max_a q_k(S_{t+1}, a) := \tilde{T}^* q_k$$

- Every iteration k:

$$q_{k+1} \quad = \quad \arg\min_{q_w \in \mathcal{F}} \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} (Y_t - q_w(S_t, A_t))^2$$

Fitted Q-iteration with other Approximations

Algorithm:

- Every iteration $k+1$:

$$
\begin{aligned}
q_{k+1} &= \arg\min_{q_\theta \in \mathcal{F}} \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \left(Y_t - q_\theta(S_t, A_t)\right)^2 && (14) \\
&= \arg\min_{q_\theta \in \mathcal{F}} \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \left(\tilde{T}^* q_k(S_t, A_t) - q_\theta(S_t, A_t)\right)^2 && (15)
\end{aligned}
$$

- $\mathcal{F} = \mathcal{F}_\theta$ can be:
  - Linear functions
  - Neural networks
  - Kernel functions
  - ...

Fitted Q-iteration (General recipe)

Algorithm:

- Every iteration $k+1$:

$$q_{k+1} = \arg\min_{q_\theta \in \mathcal{F}} \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \left(\tilde{T}^* q_k(S_t, A_t) - q_\theta(S_t, A_t) \right)^2$$

for samples $(S_t, A_t, R_{t+1}, S_{t+1}) \sim \mu, P$.

$\mathcal{F} = \mathcal{F}_\theta$ can be:

- Linear functions
- Neural networks
- Kernel functions
- ...

Samples:

- Online
- Fixed Dataset
- Replay Memory
- Generative Model

Targets:

- $\tilde{T}^* q_k = R_{t+1} + \gamma \max_a q_k(S_{t+1}, a)$
- $\tilde{T}^* q_{target} = \tilde{T}^* q_{\theta^-}$
- Off-policy updates (next lecture)
- Multi-step operators (next lecture)

Fitted Q-iteration (General recipe: DQN)

Algorithm:

- Every iteration $k+1$:

$$q_{k+1} = \arg\min_{q_\theta \in \mathcal{F}} \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \left(\tilde{T}^* q_k(S_t, A_t) - q_\theta(S_t, A_t) \right)^2$$

$\mathcal{F} = \mathcal{F}_\theta$ can be:

- Linear functions
- Neural networks
- Kernel functions
- ...

Samples:

- Online
- Fixed Dataset
- Replay Memory
- Generative Model

Targets:

- $\tilde{T}^* q_k = R_{t+1} + \gamma \max_a q_k(S_{t+1}, a)$
- $\tilde{T}^* q_{target} = \tilde{T}^* q_{\theta^-}$
- Off-policy updates (next lecture)
- Multi-step operators (next lecture)

Fitted Q-iteration (General recipe: Batch RL - 1)

Algorithm:

- Every iteration $k+1$:

$$q_{k+1} = \arg\min_{q_\theta \in \mathcal{F}} \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \left(\tilde{T}^* q_k(S_t, A_t) - q_\theta(S_t, A_t) \right)^2$$

$\mathcal{F} = \mathcal{F}_\theta$ can be:

- Linear functions
- Neural networks
- Kernel functions
- ...

Samples:

- Online
- Fixed Dataset
- Replay Memory
- Generative Model

Targets:

- $\tilde{T}^* q_k = R_{t+1} + \gamma \max_a q_k(S_{t+1}, a)$
- $\tilde{T}^* q_{target} = \tilde{T}^* q_{\theta^-}$
- Off-policy updates (next lecture)
- Multi-step operators (next lecture)

Fitted Q-iteration (General recipe: Batch RL - 2)

Algorithm:

- Every iteration $k+1$:

$$q_{k+1} = \arg\min_{q_\theta \in \mathcal{F}} \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \left(\tilde{T}^* q_k(S_t, A_t) - q_\theta(S_t, A_t) \right)^2$$

$\mathcal{F} = \mathcal{F}_\theta$ can be:

- Linear functions
- Neural networks
- Kernel functions
- ...

Samples:

- Online
- Fixed Dataset
- Replay Memory
- Generative Model

Targets:

- $\tilde{T}^* q_k = R_{t+1} + \gamma \max_a q_k(S_{t+1}, a)$
- $\tilde{T}^* q_{target} = \tilde{T}^* q_{\theta^-}$
- Off-policy updates (next lecture)
- Multi-step operators (next lecture)

Fitted Q-iteration (General recipe: Dyna)

Algorithm:

- Every iteration $k+1$:

$$q_{k+1} = \arg\min_{q_\theta \in \mathcal{F}} \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \left(\tilde{T}^* q_k(S_t, A_t) - q_\theta(S_t, A_t) \right)^2$$

$\mathcal{F} = \mathcal{F}_\theta$ can be:

- Linear functions
- Neural networks
- Kernel functions
- ...

Samples:

- Online
- Fixed Dataset
- Replay Memory
- Generative Model

Targets:

- $\tilde{T}^* q_k = R_{t+1} + \gamma \max_a q_k(S_{t+1}, a)$
- $\tilde{T}^* q_{target} = \tilde{T}^* q_{\theta^-}$
- Off-policy updates (next lecture)
- Multi-step operators (next lecture)

Approximate Policy Iteration

(Reminder) Policy Iteration

Policy Iteration

- Start with π_0.
- Iterate:
  - Policy Evaluation: $q_i = q_{\pi_i}$
  - Greedy Improvement: $\pi_{i+1} = \arg\max_a q_{\pi_i}(s, a)$

As $i \to \infty$, $q_i \to_{\|.\|_\infty} q^*$. Thus $\pi_i \to \pi^*$.

(Reminder) Approximate Policy Iteration

Approximate Policy Iteration

- Start with π_0.
- Iterate:
  - Policy Evaluation: $q_i = \mathcal{A} q_{\pi_i}$ $\quad (q_i \approx q_{\pi_i})$
  - Greedy Improvement: $\pi_{i+1} = \arg\max_a q_i(s, a)$

Question 1: As $i \to \infty$, does $q_i \to_{\|.\|_\infty} q^*$?

Question 2: Or does π_i converge to the optimal policy?

In general, what is the quality, q_{π_i}, of the obtained policy π_i?

Performance of API

Theorem (API Performance)

Consider a MDP. And let q_k and π_k be the value function and respectively evaluated (greedy) policy achieved by API at iteration k, then:

$$\lim\sup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \lim\sup_{k\to\infty} \|q_k - q_{\pi_k}\|_\infty$$

Performance of API

Theorem (API Performance)

Consider a MDP. And let q_k and π_k be the value function and respectively evaluated (greedy) policy achieved by API at iteration k, then:

$$\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \underbrace{\|q_{\pi_k} - q_k\|_\infty}_{\text{approximation error at iter. } k}$$

Performance of API (Proof)

Notation:

- Matrix P (transition probabilities): $n_a n_s \times n_s$

$$P((s,a),s') = Prob(s'|s,a)$$

- Matrix P^π (transition probabilities, given policy π): $n_a n_s \times n_a n_s$

$$P((s,a),s',a') = Prob(s',a'|s,a) = Prob(s'|s,a)\pi(a'|s')$$

- Note, that under this notation:

$$T^\pi q = R + \gamma P^\pi q$$

where $R \in \mathbb{R}^{n_s n_a}$ is a vector enumerating all rewards $r(s,a)$.

Performance of API (Proof)

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \| \underbrace{q_{\pi_k} - q_k}_{e_k} \|_\infty$

Proof.

Let's denote $gain_k := q_{\pi_{k+1}} - q_{\pi_k}$, for all iterations k.

$$
\begin{aligned}
gain_k &= q_{\pi_{k+1}} - q_{\pi_k} \\
&= T^{\pi_{k+1}} q_{\pi_{k+1}} - T^{\pi_k} q_{\pi_k} && (16) \\
&= T^{\pi_{k+1}} q_{\pi_{k+1}} - T^{\pi_{k+1}} q_{\pi_k} + && (17) \\
&\quad + T^{\pi_{k+1}} q_{\pi_k} - T^{\pi_{k+1}} q_k + && (18) \\
&\quad + T^{\pi_{k+1}} q_k - T^{\pi_k} q_k + && (19) \\
&\quad + T^{\pi_k} q_k - T^{\pi_k} q_{\pi_k} && (20)
\end{aligned}
$$

$\square$

Performance of API (Proof)

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \|\underbrace{q_{\pi_k} - q_k}_{e_k}\|_\infty$

Proof.

Let's denote $gain_k := q_{\pi_{k+1}} - q_{\pi_k}$, for all iterations k.

$$
\begin{aligned}
gain_k &= q_{\pi_{k+1}} - q_{\pi_k} \\
&= T^{\pi_{k+1}} q_{\pi_{k+1}} - T^{\pi_{k+1}} q_{\pi_k} + \\
&\quad + T^{\pi_{k+1}} q_{\pi_k} - T^{\pi_{k+1}} q_k + \\
&\quad + T^{\pi_{k+1}} q_k - T^{\pi_k} q_k + \\
&\quad + T^{\pi_k} q_k - T^{\pi_k} q_{\pi_k}
\end{aligned}
$$

$\square$

Performance of API (Proof)

Statement: $\limsup_{k\to\infty}\|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\limsup_{k\to\infty}\|\underbrace{q_{\pi_k} - q_k}_{e_k}\|_\infty$

Proof.

Let's denote $gain_k := q_{\pi_{k+1}} - q_{\pi_k}$, for all iterations k.

$$
\begin{aligned}
gain_k &= q_{\pi_{k+1}} - q_{\pi_k} \\
&= T^{\pi_{k+1}} q_{\pi_{k+1}} - T^{\pi_{k+1}} q_{\pi_k} + && = \gamma P^{\pi_{k+1}}(q_{\pi_{k+1}} - q_{\pi_k}) = \gamma P^{\pi_{k+1}} gain_k \\
&\quad + T^{\pi_{k+1}} q_{\pi_k} - T^{\pi_{k+1}} q_k + && = \gamma P^{\pi_{k+1}}(q_{\pi_k} - q_k) = \gamma P^{\pi_{k+1}} e_k \\
&\quad + T^{\pi_{k+1}} q_k - T^{\pi_k} q_k + && \geq 0 \\
&\quad + T^{\pi_k} q_k - T^{\pi_k} q_{\pi_k} && = \gamma P^{\pi_k}(q_k - q_{\pi_k}) = -\gamma P^{\pi_k} e_k
\end{aligned}
$$

Unpacking explicitly $T^\pi q_k \leq T^{\pi_{k+1}} q_k, \forall \pi$

$$
\begin{aligned}
T^{\pi_{k+1}} q_k(s,a) &= r(s,a) + \gamma \sum_{a'} \pi_{k+1}(a'|s') q_k(s',a') \\
&= r(s,a) + \gamma \max_{a'} q_k(s',a') && (\text{as } \pi_{k+1} = \arg\max_{a'} q_k(s',a'))
\end{aligned}
$$

Performance of API (Proof)

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \| \underbrace{q_{\pi_k} - q_k}_{e_k} \|_\infty$

Proof.

Let's denote $gain_k := q_{\pi_{k+1}} - q_{\pi_k}$, for all iterations k.

$$
\begin{aligned}
gain_k &= q_{\pi_{k+1}} - q_{\pi_k} \\
&= T^{\pi_{k+1}} q_{\pi_{k+1}} - T^{\pi_{k+1}} q_{\pi_k} + && = \gamma P^{\pi_{k+1}}(q_{\pi_{k+1}} - q_{\pi_k}) = \gamma P^{\pi_{k+1}} gain_k \\
&\quad + T^{\pi_{k+1}} q_{\pi_k} - T^{\pi_{k+1}} q_k + && = \gamma P^{\pi_{k+1}}(q_{\pi_k} - q_k) = \gamma P^{\pi_{k+1}} e_k \\
&\quad + T^{\pi_{k+1}} q_k - T^{\pi_k} q_k + && \ge 0 \\
&\quad + T^{\pi_k} q_k - T^{\pi_k} q_{\pi_k} && = \gamma P^{\pi_k}(q_k - q_{\pi_k}) = -\gamma P^{\pi_k} e_k \\
&\ge \gamma P^{\pi_{k+1}} gain_k + \gamma(P^{\pi_{k+1}} - P^{\pi_k}) e_k
\end{aligned}
$$

Re-arranging, we get:

$$gain_k \ge \gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k}) e_k$$

$\square$

Performance of API - Performance gain via Greedy step

Statement:

$$gain_k \geq \gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})e_k$$

Some implications:

- What if $e_k = 0$? (perfect evaluation at iter. k)

$$gain_k \geq 0$$

  aka $q_{\pi_{k+1}} \geq q_{\pi_k}$.

- Can $gain_k < 0$?

Performance of API - Performance gain via Greedy step

Q: Can $gain_k := q_{\pi_{k+1}} - q_{\pi_k}$ be negative?

Simple MDP

+1		S_0		**+1**

$$\mathcal{A} = \{\leftarrow, \rightarrow\}$$

Performance of API - Performance gain via Greedy step

- **Q**: Can $gain_k := q_{\pi_{k+1}} - q_{\pi_k}$ be negative?

Deterministic policy π_k:

	s_1	s_0	s_2
$\pi_k(a\|s)$	$\rightarrow$	$\rightarrow$	$\rightarrow$

Evaluation π_k:

$q_{\pi_k}(s,a)$	s_1	s_0	s_2
$a_1 = \rightarrow$	0.81	0.9	1.0
$a_2 = \leftarrow$	1.0	0.73	0.81

Consider an approx. q_k:

$q_k(s,a)$	s_1	s_0	s_2
$a_1 = \rightarrow$	0.8	0.83	0.85
$a_2 = \leftarrow$	1.1	0.75	0.87

Greedy policy π_{k+1}:

	s_1	s_0	s_2
$\pi_{k+1}(a\|s)$	$\leftarrow$	$\rightarrow$	$\leftarrow$

Performance of API - Performance gain via Greedy step

- **Q**: Can $gain_k := q_{\pi_{k+1}} - q_{\pi_k}$ be negative?

Deterministic policy π_k:

	s_1	s_0	s_2
$\pi_k(a\|s)$	$\rightarrow$	$\rightarrow$	$\rightarrow$

Evaluation π_k:

$q_{\pi_k}(s, a)$	s_1	s_0	s_2
$a_1 = \rightarrow$	0.81	0.9	1.0
$a_2 = \leftarrow$	1.0	0.73	0.81

Greedy policy π_{k+1}:

	s_1	s_0	s_2
$\pi_k(a\|s)$	$\leftarrow$	$\rightarrow$	$\leftarrow$

Evaluation π_{k+1}:

$q_{\pi_{k+1}}(s, a)$	s_1	s_0	s_2
$a_1 = \rightarrow$	0.0	0.9	1.0
$a_2 = \leftarrow$	1.0	0.0	0.0

Performance of API (Proof - continuing)

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \| \underbrace{q_{\pi_k} - q_k}_{e_k} \|_\infty$

Proof.

Let's denote $L_k := q^* - q_{\pi_k}$, for all iterations k. ("Loss in performance")

$$\begin{aligned} L_{k+1} &= q^* - q_{\pi_{k+1}} \\ &= T^{\pi^*} q_{\pi^*} - T^{\pi_{k+1}} q_{\pi_{k+1}} && (21) \\ &= T^{\pi^*} q_{\pi^*} - T^{\pi^*} q_{\pi_k} + && (22) \\ & \quad + T^{\pi^*} q_{\pi_k} - T^{\pi^*} q_k + && (23) \\ & \quad + T^{\pi^*} q_k - T^{\pi_{k+1}} q_k + && (24) \\ & \quad + T^{\pi_{k+1}} q_k - T^{\pi_{k+1}} q_{\pi_k} + && (25) \\ & \quad + T^{\pi_{k+1}} q_{\pi_k} - T^{\pi_{k+1}} q_{\pi_{k+1}} && (26) \end{aligned}$$

□

Performance of API (Proof - continuing)

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \| \underbrace{q_{\pi_k} - q_k}_{e_k} \|_\infty$

Proof.

Let's denote $L_k := q^* - q_{\pi_k}$, for all iterations k. ("Loss in performance")

$$
\begin{aligned}
L_{k+1} &= q^* - q_{\pi_{k+1}} \\
&= T^{\pi^*} q_{\pi^*} - T^{\pi^*} q_{\pi_k} + && = \gamma P^{\pi^*}(q_{\pi^*} - q_{\pi_k}) = \gamma P^{\pi^*} L_k \\
&\quad + T^{\pi^*} q_{\pi_k} - T^{\pi^*} q_k + && = \gamma P^{\pi^*}(q_{\pi_k} - q_k) = \gamma P^{\pi^*} e_k \\
&\quad + T^{\pi^*} q_k - T^{\pi_{k+1}} q_k + && \leq 0 \\
&\quad + T^{\pi_{k+1}} q_k - T^{\pi_{k+1}} q_{\pi_k} + && = \gamma P^{\pi_{k+1}}(q_k - q_{\pi_k}) = -\gamma P^{\pi_{k+1}} e_k \\
&\quad + T^{\pi_{k+1}} q_{\pi_k} - T^{\pi_{k+1}} q_{\pi_{k+1}} && = \gamma P^{\pi_{k+1}}(q_{\pi_k} - q_{\pi_{k+1}}) = -\gamma P^{\pi_{k+1}} g_k \\
&\leq \gamma P^{\pi^*} L_k + \gamma(P^{\pi^*} - P^{\pi_{k+1}}) e_k - \gamma P^{\pi_{k+1}} g_k
\end{aligned}
$$

□

Performance of API (Proof - continuing)

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \| \underbrace{q_{\pi_k} - q_k}_{e_k} \|_\infty$

Proof.

Thus we have:

$$
\begin{aligned}
L_{k+1} &\leq \gamma P^{\pi^*} L_k + \gamma (P^{\pi^*} - P^{\pi_{k+1}}) e_k - \gamma P^{\pi_{k+1}} g_k && (27)\\
&\leq \gamma P^{\pi^*} L_k + \gamma (P^{\pi^*} - P^{\pi_{k+1}}) e_k - \gamma P^{\pi_{k+1}} \big(\gamma (I - \gamma P^{\pi_{k+1}})^{-1} (P^{\pi_{k+1}} - P^{\pi_k}) e_k\big) && (28)\\
&\leq \gamma P^{\pi^*} L_k + \gamma \left(P^{\pi^*} + \gamma P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (P^{\pi_{k+1}} - P^{\pi_k}) - P^{\pi_{k+1}}\right) e_k && (29)\\
&\leq \gamma P^{\pi^*} L_k + \gamma \left(P^{\pi^*} + P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (I - \gamma P^{\pi_k})\right) e_k && (30)
\end{aligned}
$$

$\square$

Performance of API (Proof - continuing)

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \| \underbrace{q_{\pi_k} - q_k}_{e_k} \|_\infty$

Proof.

Thus we have:

$$\begin{aligned}
L_{k+1} &\leq \gamma P^{\pi^*} L_k + \gamma (P^{\pi^*} - P^{\pi_{k+1}}) e_k - \gamma P^{\pi_{k+1}} g_k && (31)\\
&\leq \gamma P^{\pi^*} L_k + \gamma (P^{\pi^*} - P^{\pi_{k+1}}) e_k - \gamma P^{\pi_{k+1}} \big(\gamma (I - \gamma P^{\pi_{k+1}})^{-1} (P^{\pi_{k+1}} - P^{\pi_k}) e_k\big) && (32)\\
&\leq \gamma P^{\pi^*} L_k + \gamma \left(P^{\pi^*} + \gamma P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (P^{\pi_{k+1}} - P^{\pi_k}) - P^{\pi_{k+1}} \right) e_k && (33)\\
&\leq \gamma P^{\pi^*} L_k + \gamma \left(P^{\pi^*} + P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (I - \gamma P^{\pi_k}) \right) e_k && (34)
\end{aligned}$$

Asymptotic regime $k \to \infty$:

$$\limsup_{k\to\infty} L_k \leq \gamma (I - \gamma P^{\pi^*})^{-1} \limsup_{k\to\infty} \left(P^{\pi^*} + P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (I - \gamma P^{\pi_k}) \right) e_k$$

$\square$

Performance of API (Proof - continuing)

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \| \underbrace{q_{\pi_k} - q_k}_{e_k} \|_\infty$

Proof.

Asymptotic regime $k \to \infty$:

$$\limsup_{k\to\infty} L_k \leq \gamma(I - \gamma P^{\pi^*})^{-1} \limsup_{k\to\infty} \left(P^{\pi^*} + P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k})\right) e_k$$

Thus, taking the L_∞ norm:

$$\limsup_{k\to\infty} \|L_k\|_\infty \leq \frac{\gamma}{1-\gamma} \limsup_{k\to\infty} \left\| \left(P^{\pi^*} + P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k})\right) \right\| \cdot \|e_k\|_\infty \tag{35}$$

$$\leq \frac{\gamma}{1-\gamma} \left(\frac{1+\gamma}{1-\gamma} + 1\right) \cdot \limsup_{k\to\infty} \|e_k\|_\infty \tag{36}$$

Note: Here we used that $\|P\|_\infty = 1$ for all (row-)stochastic matrices P. □

A concrete instance

(Reminder) TD(λ) with Linear Approximation

- Consider a linear hypothesis space $\mathcal{F}_\phi = \{q_w(s,a) = w^T\phi(s,a) | \forall w \in B\}$.
- Temporal difference error:

$$\delta_t = R_{t+1} + \gamma q_{w_t}(S_{t+1}, \pi(S_{t+1})) - q_{w_t}(S_t, A_t) \tag{37}$$

- Parameters update: $w_{t+1} = w_t + \alpha_t \delta_t \phi(s_t, a_t)$
- Properties:
  - This converges $\lim_{t\to\infty} w_t = w^*$, if $\sum_t \alpha_t = \infty$ and $\sum \alpha_t^2 < \infty$. (Tsitsiklis et Van Roy'97).
  - Furthermore:

$$\|q_{w^*} - q_\pi\|_{2,\mu^\pi} \leq \frac{1-\lambda\gamma}{1-\gamma} \inf_w \|q_w - q_\pi\|_{2,\mu^\pi} \tag{38}$$

TD(λ) with Linear Approximation

Statement:

$$\|q_{w^*} - q_\pi\|_{2,\mu^\pi} \leq \frac{1-\lambda\gamma}{1-\gamma} \inf_w \|q_w - q_\pi\|_{2,\mu^\pi}$$

Some implications:

- **Q**: For which λ is the RHS minimised (tightest bound)?
  - **A**: $\lambda = 1$ (TD(1) = Monte Carlo).
- **Q**: What if $q_\pi \in \mathcal{F}_\phi$?
  - **A :** RHS $= 0$. Thus $q_{w^*} = q_\pi$.
- **Q**: What if $q_\pi \notin \mathcal{F}_\phi$?
  - **A :** RHS $\neq 0$. In general the FP $q_{w^*} \neq \inf_w \|q_w - q_\pi\|_{2,\mu^\pi}$

Summary

AVI in general

Statement:

$$\|q^* - q_{\pi_n}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0\leq k<n} \underbrace{\|T^* q_k - q_{k+1}\|_\infty}_{\epsilon_k} + \cancelto{0 \text{ as } n \to \infty}{\frac{2\gamma^{n+1}}{(1-\gamma)} \|q^* - q_0\|_\infty}$$

Some lessons:

- In general, convergence is not guaranteed. (In practice, fairly well behaved)
- Control the approximation errors ϵ
  - Two sources of error: estimation(sampling) + approximation($\mathcal{F}$)
  - For efficient optimisation: $L_\infty \to L_{2,\mu}$
- Convergence point is not always q^*!
- $q^* \in \mathcal{F}$ is useful, but not enough!

API in general

Statement: $\limsup_{k\to\infty} \|q^* - q_{\pi_k}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \| \underbrace{q_{\pi_k} - q_k}_{e_k} \|_\infty$

Some lessons:

- In general, convergence is not guaranteed. (In practice, fairly well behaved)
- Control the approximation errors e_k
  - Two sources of error: estimation(sampling) + approximation($\mathcal{F}$)
  - For efficient optimisation: $L_\infty \to L_{2,\mu^{\pi_i}}$ (safe on-policy)
- Depending on the conditions/function class, we can obtain convergence:
  - Convergence point is not always q^* or q_π!
  - Convergence points might not be unique.
- $q^* \in \mathcal{F}$ is usually not enough!

Questions?

The only stupid question is the one you were afraid to ask but never did.
-Rich Sutton

For questions that may arise during this lecture please use Moodle and/or the next Q&A session.