A Theorical Analysis of Fitted Q-Iteration

Jacob Harder
University of Copenhagen

February 27, 2020

1 Abstract

2 Foreword

The main purpose of this master thesis for me, has been to uncover what (at present) it is possible to say (mathematically) about the convergence of Q-learning algorithms. In particular Q-learning algotihms using (deep) ANNs.

I came to realize during my reading of [TODO ref to YangXieWang] that it is quite error-prone with some errors not obviously fixable.

3 Disambiguation

- $[\phi] = 1$ when ϕ is true/holds and 0 otherwise, for a logical formula ϕ.
- $[q] = \{1, \ldots, q\}$ for $q \in \mathbb{N}$.
- $C_{\mathbb{K}}(X) = \{f : X \to \mathbb{K} \mid f \text{ continuous}\}$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. $C(X) = C_{\mathbb{R}}(X)$
- ANN abrv. artificial neural network see definition 2.
- δ_a Dirac-measure of point a. I.e. $\delta_a(A) = [a \in A]$.
- $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying measure space of all random variables and processes when not otherwise specified.

3.1 Notational deviations from [TODO ref YangXieWang]

Because σ is used ambigously in theorem 1 we denote the probability distribution 'σ' from [YangXieWang, thm. 6.2, p. 20] by ν instead.

I dislike the shorthand defined in [YangXieWang, p. 26 bottom]: $\|f\|_n^2 = 1/n \cdot \sum_{i=1}^n f(X_i)^2$. This is partially due to inconsistencies and abuse of this notation employed. For example $\|f\|_n$ is used as $1/n \sum_{i=1}^n f(X_i)$ as opposed another likely interpretation $\sqrt{\|f\|_n^2}$, whereas $\|f\|_n^{-1}$ is used to mean $1/(\|f\|_n)$. This is avoided by using finite dimensional p-norms instead. The conversion to my notation thus becomes $\|f\|_n \rightsquigarrow \|f\|_1 /n$, $\|f\|_n^2 \rightsquigarrow \|f\|^2 /n$, $\|f\|_n^{-1} \rightsquigarrow n\|f\|_1^{-1}$.

4 Introduction

4.1 Reinforcement Learning

In Reinforcement Learning (RL) we are concerned with finding an optimal policy for an agent in some environment. Typically (also in the case of Q-learning) this environment is a Markov decision process

Definition 1. A Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ consists of

- $\mathcal{S}$ a set of states

- $\mathcal{A}$ a set of actions
- $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ its Markov transition kernel
- $R : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathbb{R})$ its immediate reward distribution
- $\gamma \in (0, 1)$ the discount factor

A policy (for an MDP) is a function

$$\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$$

With this we can define the state-value function $V^\pi : \mathcal{S} \to \mathbb{R}$

$$V^\pi(s) = \mathbb{E}\left(\sum_{t \geq 0} \gamma^t R_t \mid R_t \sim R(S_t, A_t), S_t \sim P(S_{t-1}, A_{t-1}), A_t \sim \pi(S_t), S_0 = s\right)$$

And the state-action-value (Q-) function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$Q^\pi(s, a) = \mathbb{E}(R(s, a) + \gamma V^\pi(S_0) \mid S_0 \sim P(s, a))$$

The optimal Q-function is defined as

$$Q^*(s, a) = \sup_\pi Q^\pi(s, a)$$

One can show that there is a policy π^* such that $Q^* = Q^{\pi^*}$. This is the optimal policy - the goal of RL.

Note that V^π, Q^π and Q^* are usually infeasible to calculate to machine precision, unless $\mathcal{S} \times \mathcal{A}$ is finite and not very big.

4.2 Q-Learning

Let $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ be a policy. We define the operator

$$(P^\pi Q)(s, a) = \mathbb{E}(Q(S', A') \mid S' \sim P(s, a), A' \sim \pi(S'))$$

Intuitively this operator yields the expected state-action-value function when looking *one step ahead* following the policy π and taking expectation of Q.

We define the operator T^π called the Bellman operator by

$$(T^\pi Q)(s, a) = \mathbb{E}R(s, a) + \gamma(P^\pi Q)(s, a)$$

This operator adjust the Q function to look more like Q^π making one "iteration" of "propagation of rewards" discounting with γ. Indeed it is easily seen that Q^π is a fixed point for T^π.

A *greedy* policy π with respect to a state-action value function Q is a policy which deterministically chooses an action with maximal value of Q for each state. That is $\pi(s) = \delta_a$ for some $a \in \operatorname{argmax}_a Q(s, a)$. We then write $\pi = \pi_Q$. With this we can define the operator T:

$$TQ = T^{\pi_Q} Q$$

called the Bellman *optimality* operator.

The Bellman optimality *equation* can then be written $Q^* = TQ^*$.

Proposition 1. Q^π is the unique fixed point of T^π.

Proof. Clearly $T^\pi Q^\pi = Q^\pi$. [TODO: rest of this proof] □

4.3 Artificial Neural Networks

Definition 2. An **ANN** (Artificial Neural Network) with structure $\{d_i\}_{i=0}^{L+1} \subseteq \mathbb{N}$, activation functions $\sigma_i = (\sigma_{ij} : \mathbb{R} \to \mathbb{R})_{j=1}^{d_i}$ and weights $\{W_i \in M^{d_i \times d_{i-1}}, v_i \in \mathbb{R}^{d_i}\}_{i=1}^{L+1}$ is the function $F : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$

$$F = w_{L+1} \circ \sigma_L \circ w_L \circ \sigma_{L-1} \circ \cdots \circ w_1$$

where w_i is the affine function $x \mapsto W_i x + v_i$ for all i.

Here $\sigma_i(x_1, \ldots, x_{d_i}) = (\sigma_{i1}(x_1), \ldots, \sigma_{id_i}(x_{d_i}))$.

$L \in \mathbb{N}_0$ is called the number of hidden layers.

d_i is the number of neurons or nodes in layer i.

An ANN is called *deep* if there are two or more hidden layers.

4.4 Fitted Q-Iteration

We here present the algorithm which everything in this paper revolves around:

Algorithm 1: Fitted Q-Iteration Algorithm

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, function class $\mathcal{F}$, sampling distribution ν, number of iterations K, number of samples n, initial estimator $\widetilde{Q}_0$

for $k = 0, 1, 2, \ldots, K-1$ **do**

Sample i.i.d. observations $\{(S_i, A_i), i \in [n]\}$ from ν obtain $R_i \sim R(S_i, A_i)$ and $S_i' \sim P(S_i, A_i)$

Let $Y_i = R_i + \gamma \cdot \max_{a \in \mathcal{A}} \widetilde{Q}_k(S_i', a)$

Update action-value function:

$$\widetilde{Q}_{k+1} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(S_i, A_i))^2$$

Define π_K as the greedy policy w.r.t. $\widetilde{Q}_K$

Output: An estimator $\widetilde{Q}_K$ of Q^* and policy π_K

5 Assumptions

5.1 Assumption 1: Holder Smoothness

Definition 3. For $s, V \in \mathbb{R}$ a (s,V)-**Sparse ReLU Network** is an ANN f with any structure $\{d_i\}_{i \in [L+1]}$, all activation functions being *ReLU* i.e. $\sigma_{ij} = \max(\cdot, 0)$ and any weights (W_ℓ, v_ℓ) satisfying

- $\max_{\ell \in [L+1]} \left\| \widetilde{W}_\ell \right\|_\infty \leq 1$
- $\sum_{\ell=1}^{L+1} \left\| \widetilde{W}_\ell \right\|_0 \leq s$
- $\max_{j \in [d_{L+1}]} \left\| f_j \right\|_\infty \leq V$

Here $\widetilde{W}_\ell = (W_\ell, v_\ell)$.

The set of them we denote $\mathcal{F}(s, V)$.

Definition 4. Let $\mathcal{D} \subseteq \mathbb{R}^r$ be compact and $\beta, H > 0$. A function $f : \mathcal{D} \to \mathbb{R}$ we call Holder smooth if

$$\sum_{\alpha : |\alpha| < \beta} \|\partial^\alpha f\|_\infty + \sum_{\alpha : \|\alpha\|_1 = \lfloor \beta \rfloor} \sup_{x \neq y} \frac{|\partial^\alpha (f(x) - f(y))|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \leq H$$

Where $\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$. We write $f \in C_r(\mathcal{D}, \beta, H)$.

Definition 5. Let $t_j, p_j \in \mathbb{N}$, $t_j \leq p_j$ and $H_j, \beta_j > 0$ for $j \in [q]$. We say that f is a *Composition of Holder smooth Functions* when

$$f = g_q \circ \cdots \circ g_1$$

for some functions $g_j : [a_j, b_j]^{p_j} \to [a_{j+1}, b_{j+1}]^{p_{j+1}}$ that only depend on t_j of their inputs for each of their components g_{jk}, and satisfies $g_{jk} \in C_{t_j}([a_j, b_j]_j^t, \beta_j, H_j)$, i.e. they are Holder smooth. We denote the class of these functions

$$\mathcal{G}(\{p_j, t_j, \beta_j, H_j\}_{j \in [q]})$$

Definition 6. Define

$$\mathcal{F}_0 = \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \mid f(\cdot, a) \in \mathcal{F}(s, V) \ \forall a \in \mathcal{A}\}$$

and

$$\mathcal{G}_0 = \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \mid f(\cdot, a) = \mathcal{G}(\{p_j, t_j, \beta_t, H_j\}_{j \in [q]}) \ \forall a \in \mathcal{A}\}$$

Assumption 1. It is assumed that $Tf \in \mathcal{G}_0$ for any $f \in \mathcal{F}_0$.

I.e. when using the Bellman optimality operator on our sparse ReLU networks, we should stay in the class of compositions of Holder smooth functions.

5.2 Assumption 2: Concentration Coefficients

Definition 7 (Concentration coefficients)**.** Let $\nu_1, \nu_2 \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be probability measures, absolutely continuous w.r.t. m_λ Define

$$\kappa(m, \nu_1, \nu_2) = \sup_{\pi_1,\dots,\pi_m} \left[\mathbb{E}_{v_2} \left(\frac{\mathrm{d}(P^{\pi_m} \dots P^{\pi_1} \nu_1)}{\mathrm{d}\nu_2} \right)^2 \right]^{1/2}$$

Assumption 2. Let ν be the sampling distribution from the algorithm, and μ the distribution over which we measure the error in the main theorem, then we assume

$$(1-\gamma)^2 \sum_{m \geq 1} \gamma^{m-1} m \kappa(m, \mu, \nu) = \phi_{\mu,\nu} < \infty$$

6 Main theorem

Theorem 1 (Yang, Xie, Wang)**.** For any $K \in \mathbb{N}$ let Q^{π_K} be the action-value function corresponding to policy π_K which is returned by Algorithm 1, when run with a sparse ReLU network on the form

$$\mathcal{F}_0 = \{ f(\cdot, a) \in \mathcal{F}(L^*, \{d_j^*\}_{j=0}^{L^*+1}, s^*) \mid a \in \mathcal{A} \}$$

where

$$L^* \lesssim (\log n)^{\xi'}, d_0 = r, d_j^*, d_{L+1} = 1, \lesssim n^{\xi'}, s^* \asymp n^{\alpha^*} \cdot (\log n)^{\xi'}$$

Let μ be any distribution over $\mathcal{S} \times \mathcal{A}$. Under assumption 1 and assumption 2

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq C \cdot \frac{\phi_{\mu,\nu} \cdot \gamma}{(1-\gamma)^2} \cdot |\mathcal{A}| \cdot (\log n)^{\xi^*} \cdot n^{(\alpha^*-1)/2} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} \cdot R_{\max}$$

Here $C, \xi', \xi^*, \phi_{\mu,\nu} \in \mathbb{R}_+$ and $\alpha^* \in (0,1)$ are constants depending on the assumptions and $R_{\max}$ the maximum possible reward.

7 Proofs

Proof of main theorem. Using theorem 2 we get

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq 2\frac{\phi_{\mu,\nu}}{(1-\gamma)^2} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max} \tag{1}$$

where $\varepsilon_{\max} = \max_{k \in [K]} \left\| T\widetilde{Q}_{k-1} - \widetilde{Q}_k \right\|_{2,\nu}$. Using **??** with $Q = \widetilde{Q}_{k-1}$, $\mathcal{F} = \mathcal{F}_0$, $\epsilon = 1$ and $\delta = 1/n$, we get

$$\varepsilon_{\max} \leq 4\omega(\mathcal{F}_0) + C \cdot V_{\max}^2 / n \cdot \log N_0 \tag{2}$$

where $C = 64 + 8/V_{\max}$ and $N_0 = \left| \mathcal{N}(1/n, \mathcal{F}_0, \|\cdot\|_\infty) \right|$. □

Theorem 2 (Error Propagation)**.** Let $\{\widetilde{Q}_i\}_{0 \leq i \leq K}$ be the iterates of the fitted Q-iteration algorithm. Then

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq \frac{2\phi_{\mu,\nu}\gamma}{(1-\gamma)^2} \cdot \varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} \cdot R_{\max}$$

Where

$$\varepsilon_{\max} = \max_{k \in [K]} \left\| T\widetilde{Q}_{k-1} - \widetilde{Q}_k \right\|_{2,\nu}$$

Lemma 1. $TQ \geq T^\pi Q$ for any policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ and any action value function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

Proof.

$$\begin{aligned}
(TQ)(s,a) &= \mathbb{E} \left(R(s,a) + \gamma \max_{a'} Q(S', a') \mid S' \sim P(\cdot \mid s, a) \right) \\
&\geq \mathbb{E} \left(R(s,a) + \gamma Q(S', A') \mid S' \sim P(\cdot \mid s,a), A' \sim \pi(\cdot \mid S') \right) \\
&= T^\pi Q(s,a)
\end{aligned}$$

□

Lemma 2. Let $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ be an action-value function, $\tau_1, \ldots, \tau_m$ be policies and $\mu \in \mathcal{P}(\mathcal{S}\times\mathcal{A})$ be a probability measure. Then

$$\mathbb{E}_\mu[(P^{\tau_m} \ldots P^{\tau_1})(f)] \le \kappa(k - i + j; \mu, \nu)\|f\|_{2,\nu}$$

For any measure $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ which is absolutely continuous w.r.t. $(P^{\tau_m} \ldots P^{\tau_1})(\mu)$. Here κ is the concentration coefficients defined in definition 7.

Proof. Recall that

$$\begin{aligned}\kappa(m; \mu, \nu) &:= \sup_{\pi_1,\ldots,\pi_m} \left[\mathbb{E}_\nu\left|\frac{\mathrm{d}(P^{\pi_m} \ldots P^{\pi_1}\mu)}{\mathrm{d}\nu}\right|^2\right]^{1/2} \\ &= \sup_{\pi_1,\ldots,\pi_m} \left\|\frac{\mathrm{d}(P^{\pi_m} \ldots P^{\pi_1}\mu)}{\mathrm{d}\nu}\right\|_{2,\nu}\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}_\mu[(P^{\tau_m} \ldots P^{\tau_1})(f)] &= \int (P^{\tau_m} \ldots P^{\tau_1})(f)\,\mathrm{d}\mu && (3)\\ &= \int f\,\mathrm{d}(P^{\tau_m} \ldots P^{\tau_1}\mu) && (4)\\ &= \int f\frac{\mathrm{d}(P^{\tau_m} \ldots P^{\tau_1}\mu)}{\mathrm{d}\nu}\,\mathrm{d}\nu && (5)\\ &\le \left\|\frac{\mathrm{d}(P^{\tau_m} \ldots P^{\tau_1}\mu)}{\mathrm{d}\nu}\right\|_{2,\nu} \cdot \|f\|_{2,\nu} && (6)\\ &\le \kappa(m, \mu, \nu)\|f\|_{2,\nu} && (7)\end{aligned}$$

Where eq. (5) is due to the Radon-Nikodym theorem and eq. (6) is Cauchy-Schwarz. □

Proof of theorem 2. First some things to keep in mind during the proof. Recall that $V_{\max} = R_{\max}/(1-\gamma)$ and that π_Q is the greedy policy w.r.t. Q. Denote

$$\pi_i = \pi_{\widetilde{Q}_i},\ Q_{i+1} = T\widetilde{Q}_i,\ \varrho_i = Q_i - \widetilde{Q}_i,\quad \text{for } i \in \{0, \ldots, K+1\}$$

Note that for any policy π, P^π is linear and 1-contrative on $\mathcal{L}^\infty(\mathcal{S} \times \mathcal{A})$. Also

$$T^\pi Q^\pi = Q^\pi,\ TQ = T^{\pi_Q}Q,\ TQ^* = Q^* = Q^{\pi^*}$$

where π^* is greedy w.r.t. Q^*. If $f > f'$ for $f, f' : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ then $P^\pi f \ge P^\pi f'$.

The proof consists of four steps.

Step 1 We start by relating $Q^* - Q^{\pi_K}$, the quantity of interest, to $Q^* - \widetilde{Q}_K$, which is more related to the output of the algorithm. Using lemma 1 we can make the upper bound

$$\begin{aligned}Q^* - Q^{\pi_K} &= T^{\pi^*}Q^* - T^{\pi_K}Q^{\pi_K}\\ &= T^{\pi^*}Q^* + (T^{\pi^*}\widetilde{Q}_K - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T\widetilde{Q}_K) - T^{\pi_K}Q^{\pi_K}\\ &= (T^{\pi^*}\widetilde{Q}_K - T\widetilde{Q}_K) + (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K})\\ &\le (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K})\\ &= (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_K) + (T^{\pi_K}\widetilde{Q}_K - T^{\pi_K}Q^{\pi_K})\\ &= \gamma P^{\pi^*}(Q^* - \widetilde{Q}_K) + \gamma P^{\pi_K}(\widetilde{Q}_K - Q^{\pi_K})\\ &= \gamma(P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}_K) + \gamma P^{\pi_K}(Q^* - Q^{\pi_K}) && (8)\end{aligned}$$

This implies

$$(I - \gamma P^{\pi_K})(Q^* - Q^{\pi_K}) \le \gamma(P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}_K)$$

Since γP^{π_K} is γ-contractive, $U = (I - \gamma P^{\pi_K})^{-1}$ exists as a bounded operator on $\mathcal{L}^\infty(\mathcal{S} \times \mathcal{A})$ and equals

$$U = \sum_{i=0}^{\infty} \gamma^i (P^{\pi_K})^i$$

From this we also see that $f \geq f' \implies Uf \geq Uf'$ for any $f, f' : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Therefore we can apply U on both sides of eq. (8) to obtain

$$Q^* - Q^{\pi_K} \leq \gamma U^{-1}(P^{\pi^*}(Q^* - \widetilde{Q}_K) - P^{\pi_K}(Q^* - \widetilde{Q}_K)) \tag{9}$$

Step 2 Using lemma 1 for any $i \in [K]$ we can get an upper bound

$$\begin{aligned} Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T^{\pi^*}\widetilde{Q}_i - T^{\pi^*}\widetilde{Q}_i) \\ &= (Q^* - T^{\pi^*}\widetilde{Q}_i) + (T\widetilde{Q}_i - \widetilde{Q}_{i+1}) + (T^{\pi^*}\widetilde{Q}_i - T\widetilde{Q}_i) \\ &= (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_i) + \varrho_{i+1} + (T^{\pi^*}\widetilde{Q}_i - T\widetilde{Q}_i) \\ &\leq T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_i + \varrho_{i+1} \\ &= \gamma P^{\pi^*}(Q^* - \widetilde{Q}_i) + \varrho_{i+1} \end{aligned} \tag{10}$$

and a lower bound

$$\begin{aligned} Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T^{\pi_i}Q^* - T^{\pi_i}Q^*) \\ &= (T^{\pi_i}Q^* - T^{\pi_i}\widetilde{Q}_i) + \varrho_{i+1} + (TQ^* - T^{\pi_i}Q^*) \\ &\geq T^{\pi_i}Q^* - T^{\pi_i}\widetilde{Q}_i + \varrho_{i+1} \\ &= \gamma P^{\pi_i}(Q^* - \widetilde{Q}_i) + \varrho_{i+1} \end{aligned} \tag{11}$$

Applying eq. (10) and eq. (11) iteratively we get

$$Q^* - \widetilde{Q}_K \leq \gamma^K (P^{\pi^*})^K (Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi^*})^{K-1-i} \varrho_{i+1} \tag{12}$$

and

$$Q^* - \widetilde{Q}_K \geq \gamma^K (P^{\pi_{K-1}} \dots P^{\pi_0})(Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi_{K-1}} \dots P^{\pi_{i+1}}) \varrho_{i+1} \tag{13}$$

Step 3 Combining eq. (12) and eq. (13) with eq. (9) we get

$$\begin{aligned} Q^* - Q^{\pi_K} \leq U^{-1}\Bigg(&\gamma^{K+1}((P^{\pi^*})^{K+1} - P^{\pi_K} \dots P^{\pi_0})(Q^* - \widetilde{Q}_0) \\ &+ \sum_{i=0}^{K-1} \gamma^{K-i}((P^*)^{K-i} - P^{\pi_K} \dots P^{\pi_{i+1}}) \varrho_{i+1} \Bigg) \end{aligned} \tag{14}$$

For shorthand define constants

$$\alpha_i = \frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}} \text{ for } 0 \leq i \leq K-1 \text{ and } \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} \tag{15}$$

(note that $\sum_{i=0}^{K} \alpha_i = 1$) and operators

$$O_i = (1-\gamma)/2U^{-1}[(P^{\pi^*})^{K-i} + (P^{\pi_K} \dots P^{\pi_{i+1}})] \tag{16}$$

$$O_K = (1-\gamma)/2U^{-1}[(P^{\pi^*})^{K+1} + (P^{\pi_K} \dots P^{\pi_0})] \tag{17}$$

Then by eq. (14)

$$|Q^* - Q^{\pi_K}| \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{i=0}^{K-1} \alpha_i O_i |\varrho_{i+1}| + \alpha_K O_K \left| Q^* - \widetilde{Q}_0 \right| \right] \tag{18}$$

So by linearity of expectation

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} = \mathbb{E}_\mu |Q^* - Q^{\pi_K}| \tag{19}$$

$$\leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{i=0}^{K-1} \alpha_i \mathbb{E}_\mu(O_i|\varrho_{i+1}|) + \alpha_K \mathbb{E}_\mu(O_K \left| Q^* - \widetilde{Q}_0 \right|) \right] \tag{20}$$

With the bound on rewards we (crudely) estimate

$$\mathbb{E}_\mu O_K \left| Q^* - \widetilde{Q}_0 \right| \leq 2V_{\max} = 2R_{\max}/(1-\gamma) \tag{21}$$

The remaining difficulty lies in $\mathbb{E}_\mu(O_i|\varrho_{i+1}|)$.

Step 4 Using the sum expansion of U^{-1} we get

$$\begin{aligned}
&\mathbb{E}_\mu(O_i|\varrho_{i+1}|) && (22)\\
&= \frac{1-\gamma}{2}\mathbb{E}_\mu\left(U^{-1}[(P^{\pi_K})^{K-i} + P^{\pi_K}\dots P^{\pi_{i+1}}]|\varrho_{i+1}|\right) && (23)\\
&= \frac{1-\gamma}{2}\mathbb{E}_\mu\left(\sum_{j=0}^{\infty}[(P^{\pi_K})^j(P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1}P^{\pi_{K-1}}\dots P^{\pi_{i+1}}]|\varrho_{i+1}|\right) && (24)\\
&= \frac{1-\gamma}{2}\sum_{j=0}^{\infty}\mathbb{E}_\mu\left([(P^{\pi_K})^j(P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1}P^{\pi_{K-1}}\dots P^{\pi_{i+1}}]|\varrho_{i+1}|\right) && (25)
\end{aligned}$$

Notice that there are $K-i+j$ P-operators on both terms in the sum. Therefore were can employ lemma 2 twice. Moreover define $\varepsilon_{\max} = \max_{i\in[K]}\|\varrho_i\|_{2,\nu}$. Then

$$\begin{aligned}
\mathbb{E}_\mu(O_i|\varrho_{i+1}|) &\leq (1-\gamma)\sum_{j=0}^{\infty}\gamma^j\kappa(K-i+j;\mu,\nu)\|\varrho_{i+1}\|_{2,\nu}\\
&\leq \varepsilon_{\max}(1-\gamma)\sum_{j=0}^{\infty}\gamma^j\kappa(K-i+j;\mu,\nu) && (26)
\end{aligned}$$

Using eq. (20), eq. (21) and eq. (26)

$$\begin{aligned}
\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq \frac{2\gamma(1-\gamma^{K+1})}{1-\gamma}&\left[\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\gamma^j\kappa(K-i+j;\mu,\nu)\right]\varepsilon_{\max}\\
&+\frac{4\gamma(1-\gamma^{K+1})}{(1-\gamma)^3}\alpha_K R_{\max}
\end{aligned} \tag{27}$$

Focusing on the first term on RHS of eq. (27), if we then we can take the norm out of the sum as a constant. We are left with

$$\begin{aligned}
&\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\gamma^j\kappa(K-i+j;\mu,\nu)\\
&= \sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\frac{(1-\gamma)\gamma^{K-i+j-1}}{1-\gamma^{K+1}}\kappa(K-i+j;\mu,\nu)\\
&= \frac{1-\gamma}{1-\gamma^{K+1}}\sum_{j=0}^{\infty}\sum_{i=0}^{K-1}\gamma^{K-i+j-1}\kappa(K-i+j;\mu,\nu)\\
&\leq \frac{1-\gamma}{1-\gamma^{K+1}}\sum_{m=0}^{\infty}\gamma^{m-1}\cdot m\cdot\kappa(m;\mu,\nu)\\
&\leq \frac{1}{1-\gamma^{K+1}(1-\gamma)}\phi_{\mu,\nu} && (28)
\end{aligned}$$

Where the last inequality is due to assumption 2. Combining eq. (27) and eq. (28) we arrive at

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \leq \frac{2\gamma\cdot\phi_{\mu,\nu}}{(1-\gamma)^2}\cdot\varepsilon_{\max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2}\cdot R_{\max} \tag{29}$$

□

Theorem 3 (One-step Approximation Error)**.** Let

- $\mathcal{F} \subseteq \mathcal{B}(\mathcal{S} \times \mathcal{A}, V_{\max})$ be a class ofbounded measurable functions
- $\nu \in \mathcal{P}(\mathcal{S}, \mathcal{A})$ be a probability measure
- $(S_i, A_i)_{i\in[n]}$ be n i.i.d. samples following ν
- $(R_i, S_i')_{i\in[n]}$ be the rewards and next states corresponding to the samples
- $Q \in \mathcal{F}$ be fixed
- $Y_i = R_i + \gamma \max_{a\in\mathcal{A}} Q(S_i', a)$
- $\widehat{Q} = \operatorname{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(S_i, A_i) - Y_i)^2$
- $\epsilon \in (0, 1],\ \delta > 0$ be fixed
- $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$ a minimal δ-covering of $\mathcal{F}$ w.r.t. $\|\cdot\|_\infty$
- $N_\delta = |\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)|$ the number of elements in this covering

Then

$$(1+\epsilon)^2 + \omega(\mathcal{F}) + C \cdot V_{\max}^2/(n+\epsilon) \cdot N_\delta + C' \cdot V_{\max} \cdot \delta$$

where $C = 64, C' = 8$ and

$$\omega(\mathcal{F}) = \sup_{g\in\mathcal{F}} \inf_{f\in\mathcal{F}} \|f - Tg\|_{2,\nu}^2$$

Proposition 2. $Y_i(\mathbb{P})$ and $TQ(X_i)(\mathbb{P})$ are i.i.d.

Proof. Recall that $X_i = (S_i, A_i)$,

$$Y_i = \mathbb{E}R(X_i) + \gamma \max_{a\in\mathcal{A}} Q(S_{i+1}, a)$$

where $S_{i+1} \sim P(X_i)$, and

$$TQ(X_i) = \mathbb{E}R(X_i) + \gamma \mathbb{E}_{S'} Q(S', \operatorname*{argmax}_{a\in\mathcal{A}} Q(S', a))$$

where $S' \sim P(X_i)$. It is a fundamental assumption/definition that S' and S_{i+1} are i.i.d. $\square$

Proof of theorem 3. First some introductory fixing of notation and variables. Fix a minimal δ-covering of $\mathcal{F}$ with centers $f_1, \ldots, f_{N_\delta}$. Define

$$\widetilde{Q} := \operatorname*{argmin}_{f\in\mathcal{F}} \|f - TQ\|_\nu^2$$

$$k^* := \operatorname*{argmin}_{k\in[N_\delta]} \left\|f_k - \widehat{Q}\right\|_\infty$$

and $X_i := (S_i, A_i)$. Notice that $\widetilde{Q}$ differs from $\widehat{Q}$ in that $\widetilde{Q}$ approximates TQ w.r.t. $\|\cdot\|_\nu^2$ while $\widehat{Q}$ approximates $Y = (Y_1, \ldots, Y_n)$ in mean squared error over $X = (X_1, \ldots, X_n)$. We shall be loose about applying functions to vectors (of random variables) in the sense that they are applied entry-wise. We use $\|\cdot\|_p$ to denote the (finite dimensional) p-norm (p ommitted when $p = 2$). When talking about p-norms on the random variables we always specify the distribution (e.g. $\|\cdot\|_\nu$). When the sample (e.g. X) is clear from context we omit it writing $\|f\| = \|f(X)\|$.

Step 1 By definion (of $\widehat{Q}$) for all $f \in \mathcal{F}$ we have $\left\|\widehat{Q}(X) - Y\right\|^2 \le \|f(X) - Y\|^2$, leading to

$$\begin{aligned}
&\|Y\|^2 + \left\|\widehat{Q}\right\|^2 - 2Y \cdot Q \le \|Y\|^2 + \|f\|^2 - 2Y \cdot f \\
\iff &\left\|\widehat{Q}\right\|^2 + \|TQ\|^2 - 2\widehat{Q} \cdot TQ \le \|f\|^2 + \|TQ\|^2 - 2f \cdot TQ + 2Y \cdot \widehat{Q} - 2Y \cdot f - 2\widehat{Q} \cdot TQ + 2f \cdot TQ \\
\iff &\|Q - TQ\|^2 \le \|f - TQ\|^2 + 2(Y - TQ) \cdot (\widehat{Q} - f) \\
\iff &\|Q - TQ\|^2 \le \|f - TQ\|^2 + 2\xi \cdot (\widehat{Q} - f)
\end{aligned}$$

Where $\xi_i := Y_i - TQ(X_i)$ and $\xi := (\xi_1, \ldots, \xi_n)$. By proposition 2 we have that Y_i and $TQ(X_i)$ are i.i.d. So satisfy $\mathbb{E}(\xi_i \mid X_i) = 0$. Therefore $\mathbb{E}(\xi_i g(X_i)) = 0$ for any function $g : \mathbb{R} \to \mathbb{R}$. So

$$\mathbb{E}\xi \cdot (\widehat{Q} - f) = \mathbb{E}\xi \cdot (\widehat{Q} - TQ) \tag{30}$$

To bound this we insert f_{k^*} by the triangle inequality

$$\left|\mathbb{E}\xi \cdot (\widehat{Q} - TQ)\right| \leq \left|\mathbb{E}\xi \cdot (\widehat{Q} - f_{k^*})\right| + \left|\mathbb{E}\xi \cdot (f_{k^*} - TQ)\right| \tag{31}$$

We now bound these two terms. The first by Cauchy-Schwarz

$$\left|\mathbb{E}\xi \cdot (\widehat{Q} - f_{k^*})\right| \leq \mathbb{E}\left(\|\xi\| \left\|\widehat{Q} - f_{k^*}\right\|\right) \leq \mathbb{E}(\|\xi\|)\sqrt{n}\delta \leq 2nV_{\max}\delta \tag{32}$$

where we have used that $\left\|\widehat{Q} - f_{k^*}\right\|_\infty \leq \delta$ so

$$\left\|\widehat{Q} - f_{k^*}\right\|^2 = \sum_{i=1}^{n} (\widehat{Q}(X_i) - f_{k^*}(X_i))^2 \leq \sum_{i=1}^{n} \delta^2 = n\delta^2 \tag{33}$$

and that $|Y_i|, TQ(X_i) \leq V_{\max}$ so

$$\|\xi\|^2 = \sum_{i=1}^{n} (Y_i - TQ(X_i))^2 \leq \sum_{i=1}^{n} (2V_{\max})^2 = 4V_{\max}^2 n \tag{34}$$

To bound the second term in eq. (31) define

$$Z_j := \sqrt{n}\, \xi \cdot (f_j - TQ) \left\|f_j - TQ\right\|_1^{-1} \tag{35}$$

Then

$$\begin{aligned}
\mathbb{E}(\xi \cdot (f_{k^*} - TQ)) &= \frac{1}{\sqrt{n}} \mathbb{E}(\|f_{k^*} - TQ\|_1 |Z_{k^*}|) \\
&\leq \frac{1}{\sqrt{n}} \mathbb{E}\left(\left(\left\|\widehat{Q} - TQ\right\|_1 + \left\|\widehat{Q} - f_{k^*}\right\|_1\right) |Z_{k^*}|\right) \\
&\leq \frac{1}{\sqrt{n}} \mathbb{E}\left(\left(\left\|\widehat{Q} - TQ\right\|_1 + \delta\right) |Z_{k^*}|\right)
\end{aligned}$$

□

8 Appendices

8.1 Various lemmas

Proposition 3. For $x > 0$.

$$\int_x^\infty e^{-t^2/2}\, \mathrm{d}t \leq \frac{1}{x} e^{-x^2/2}$$

Proof. Observe that for $t \geq x > 0$ we have $1 \leq t/x$ so

$$\begin{aligned}
\int_x^\infty e^{-t^2/2}\, \mathrm{d}t &\leq \int_x^\infty \frac{t}{x} e^{-t^2/2}\, \mathrm{d}t \\
&\leq \frac{1}{x} e^{-x^2/2}
\end{aligned}$$

□