

A Theoretical Analysis of Fitted Q-Iteration

Jacob Harder
University of Copenhagen

March 2, 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 algorithms 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 .
- $[q] = \{1, \dots, q\}$ for $q \in \mathbb{N}$.
- $C_{\mathbb{K}}(X) = \{f : X \rightarrow \mathbb{K} \mid f \text{ continuous}\}$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. $C(X) = C_{\mathbb{R}}(X)$
- 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 ambiguously 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
 - $P : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{S})$ its Markov transition kernel
 - $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathbb{R})$ its immediate reward distribution
 - $\gamma \in (0, 1)$ the discount factor
- A policy (for an MDP) is a function

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

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

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

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

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

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 $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$ 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 $\sigma_i = (\sigma_{ij} : \mathbb{R} \rightarrow \mathbb{R})_{j=1}^{d_i}$ 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 $F : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_{L+1}}$

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

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

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

$L \in \mathbb{N}_0$ 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 \tilde{Q}_0

for $k = 0, 1, 2, \dots, K - 1$ **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 $Y_i = R_i + \gamma \cdot \max_{a \in \mathcal{A}} \tilde{Q}_k(S'_i, a)$

 Update action-value function:

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

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

Output: An estimator \tilde{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

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

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

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 $f : \mathcal{D} \rightarrow \mathbb{R}$ we call Holder smooth if

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

Where $\alpha = (\alpha_1, \dots, \alpha_r) \in \mathbb{N}^r$. 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

$$f = g_q \circ \dots \circ g_1$$

for some functions $g_j : [a_j, b_j]^{p_j} \rightarrow [a_{j+1}, b_{j+1}]^{p_{j+1}}$ that only depend on t_j of their inputs for each of their components g_{jk} , and satisfies $g_{jk} \in C_{t_j}([a_j, b_j]^{t_j}, \beta_j, H_j)$, 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

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

and

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

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

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

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

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

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

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

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

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

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

Theorem 2 (Error Propagation). Let $\{\tilde{Q}_i\}_{0 \leq i \leq K}$ 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

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

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

Proof.

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

\square

Lemma 2. Let $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ be an action-value function, τ_1, \dots, τ_m be policies and $\mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ be a probability measure. Then

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

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

Proof. Recall that

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

Thus

$$\mathbb{E}_\mu[(P^{\tau_m} \dots P^{\tau_1})(f)] = \int (P^{\tau_m} \dots P^{\tau_1})(f) d\mu \quad (3)$$

$$= \int f d(P^{\tau_m} \dots P^{\tau_1} \mu) \quad (4)$$

$$= \int f \frac{d(P^{\tau_m} \dots P^{\tau_1} \mu)}{d\nu} d\nu \quad (5)$$

$$\leq \left\| \frac{d(P^{\tau_m} \dots P^{\tau_1} \mu)}{d\nu} \right\|_{2, \nu} \cdot \|f\|_{2, \nu} \quad (6)$$

$$\leq \kappa(m, \mu, \nu) \|f\|_{2, \nu} \quad (7)$$

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

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

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

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

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

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

The proof consists of four steps.

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

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

This implies

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

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 $f, f' : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. Therefore we can apply U on both sides of eq. (8) to obtain

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

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

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

and a lower bound

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

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

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

and

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

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

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

For shorthand define constants

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

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

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

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

Then by eq. (14)

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

So by linearity of expectation

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

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

With the bound on rewards we (crudely) estimate

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

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

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

$$\mathbb{E}_\mu(O_i | \varrho_{i+1}) \quad (22)$$

$$= \frac{1 - \gamma}{2} \mathbb{E}_\mu \left(U^{-1} [(P^{\pi_K})^{K-i} + P^{\pi_K} \dots P^{\pi_{i+1}}] | \varrho_{i+1} \right) \quad (23)$$

$$= \frac{1 - \gamma}{2} \mathbb{E}_\mu \left(\sum_{j=0}^{\infty} [(P^{\pi_K})^j (P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1} P^{\pi_{K-1}} \dots P^{\pi_{i+1}}] | \varrho_{i+1} \right) \quad (24)$$

$$= \frac{1 - \gamma}{2} \sum_{j=0}^{\infty} \mathbb{E}_\mu \left([(P^{\pi_K})^j (P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1} P^{\pi_{K-1}} \dots P^{\pi_{i+1}}] | \varrho_{i+1} \right) \quad (25)$$

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

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

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

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

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

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

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

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

□

Theorem 3 (One-step Approximation Error). Let

- $\mathcal{F} \subseteq \mathcal{B}(\mathcal{S} \times \mathcal{A}, V_{\max})$ be a class of bounded 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
- $Y_i = R_i + \gamma \max_{a \in \mathcal{A}} Q(S'_i, a)$
- $\hat{Q} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(S_i, A_i) - Y_i)^2$
- $\kappa \in (0, 1]$, $\delta > 0$ 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. Let v be a random vector in \mathbb{R}^n then

$$\mathbb{E}\|v\|_1 \leq \sqrt{n} \sqrt{\mathbb{E}\|v\|_2^2}$$

Proof. Denote v 's coordinates $v = (v_1, \dots, v_n)$. Cauchy-Schwarz applied to some vector w and $(1, \dots, 1)$ yields

$$\|w\|_1 \leq \sqrt{n} \|w\|_2$$

Now let $w = (\mathbb{E}v_1, \dots, \mathbb{E}v_n)$. Then by linearity of expectation and Jensens inequality

$$\mathbb{E}\|v\| = \|w\| \leq \sqrt{n} \sqrt{\sum_{i=1}^n (\mathbb{E}v_i)^2} \leq \sqrt{n} \sqrt{\mathbb{E} \sum_{i=1}^n v_i^2} = \sqrt{n} \sqrt{\mathbb{E}\|v\|_2^2}$$

□

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

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

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

and $X_i := (S_i, A_i)$. Notice that \tilde{Q} differs from \hat{Q} in that \tilde{Q} approximates TQ w.r.t. $\|\cdot\|_\nu^2$ while \hat{Q} approximates $Y = (Y_1, \dots, Y_n)$ in mean squared error over $X = (X_1, \dots, X_n)$. 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 omitted 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 definition (of \hat{Q}) for all $f \in \mathcal{F}$ we have $\|\hat{Q}(X) - Y\|^2 \leq \|f(X) - Y\|^2$, leading to

$$\|Y\|^2 + \|\hat{Q}\|^2 - 2Y \cdot \hat{Q} \leq \|Y\|^2 + \|f\|^2 - 2Y \cdot f \quad (30)$$

$$\iff \|\hat{Q}\|^2 + \|TQ\|^2 - 2\hat{Q} \cdot TQ \leq \|f\|^2 + \|TQ\|^2 - 2f \cdot TQ + 2Y \cdot \hat{Q} - 2Y \cdot f - 2\hat{Q} \cdot TQ + 2f \cdot TQ \quad (31)$$

$$\iff \|\hat{Q} - TQ\|^2 \leq \|f - TQ\|^2 + 2(Y - TQ) \cdot (\hat{Q} - f) \quad (32)$$

$$\iff \|\hat{Q} - TQ\|^2 \leq \|f - TQ\|^2 + 2\xi \cdot (\hat{Q} - f) \quad (33)$$

Where $\xi_i := Y_i - TQ(X_i)$ and $\xi := (\xi_1, \dots, \xi_n)$. Now we proof a minor lemma

Proposition 3. $\mathbb{E}(\xi_i g(X_i)) = 0$ for any function $g : \mathbb{R} \rightarrow \mathbb{R}$.

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

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

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

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

where $S' \sim P(X_i)$, $R'_i \sim R(X_i)$. Since S' and S_{i+1} are i.i.d.

$$\begin{aligned} \mathbb{E}_{X_i} \xi_i &= \mathbb{E}_{X_i} (Y_i - TQ(X_i)) \\ &= \mathbb{E}_{X_i} R_i - \mathbb{E}_{X_i} R'_i + \gamma \left(\mathbb{E}_{X_i} \left(\max_{a \in \mathcal{A}} Q(S_{i+1}, a) \right) - \mathbb{E}_{X_i} \operatorname{argmax}_{a \in \mathcal{A}} (Q(S', a)) \right) \\ &= 0 \end{aligned}$$

Therefore $\mathbb{E}(\xi_i g(X_i)) = 0$. □

By this lemma we can deduce

$$\mathbb{E} \left(\xi \cdot (\widehat{Q} - f) \right) = \mathbb{E} \left(\xi \cdot (\widehat{Q} - TQ) \right) \quad (34)$$

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

$$\left| \mathbb{E} \left(\xi \cdot (\widehat{Q} - TQ) \right) \right| \leq \left| \mathbb{E} \left(\xi \cdot (\widehat{Q} - f_{k^*}) \right) \right| + \left| \mathbb{E} \left(\xi \cdot (f_{k^*} - TQ) \right) \right| \quad (35)$$

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

$$\left| \mathbb{E} \xi \cdot (\widehat{Q} - f_{k^*}) \right| \leq \mathbb{E} \left(\|\xi\| \left\| \widehat{Q} - f_{k^*} \right\| \right) \leq \mathbb{E}(\|\xi\|) \sqrt{n} \delta \leq 2n V_{\max} \delta \quad (36)$$

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

$$\left\| \widehat{Q} - f_{k^*} \right\|^2 = \sum_{i=1}^n (\widehat{Q}(X_i) - f_{k^*}(X_i))^2 \leq \sum_{i=1}^n \delta^2 = n \delta^2 \quad (37)$$

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

$$\|\xi\|^2 = \sum_{i=1}^n (Y_i - TQ(X_i))^2 \leq \sum_{i=1}^n (2V_{\max})^2 = 4V_{\max}^2 n \quad (38)$$

To bound the second term in eq. (35) define

$$Z_j := \xi \cdot (f_j - TQ) \|f_j - TQ\|^{-1} \quad (39)$$

Then

$$\mathbb{E}(\xi \cdot (f_{k^*} - TQ)) = \mathbb{E}(\|f_{k^*} - TQ\| |Z_{k^*}|) \quad (40)$$

$$\leq \mathbb{E}\left(\left(\|\hat{Q} - TQ\| + \|\hat{Q} - f_{k^*}\|\right) |Z_{k^*}|\right) \quad (41)$$

$$\leq \mathbb{E}\left(\left(\|\hat{Q} - TQ\| + n\delta\right) |Z_{k^*}|\right) \quad (42)$$

$$\leq \left(\mathbb{E}\left(\|\hat{Q} - TQ\| + n\delta\right)^2\right)^{1/2} \left(\mathbb{E}Z_{k^*}^2\right)^{1/2} \quad (43)$$

$$\leq \mathbb{E}\left(\|\hat{Q} - TQ\| + n\delta\right) \left(\mathbb{E}Z_{k^*}^2\right)^{1/2} \quad (44)$$

$$\leq \left(\sqrt{\mathbb{E}\|\hat{Q} - TQ\|_2^2} + n\delta\right) \left(\mathbb{E}Z_{k^*}^2\right)^{1/2} \quad (45)$$

$$\leq \left(\sqrt{\mathbb{E}\|\hat{Q} - TQ\|_2^2} + n\delta\right) 2V_{\max} \quad (46)$$

Where eq. (40) to eq. (41) is by the triangle inequality, eq. (44) to eq. (45) is proposition 2 and eq. (45) to eq. (46) is due to the following

Proposition 4.

$$|Z_j| \leq 2V_{\max}$$

Proof. For any $i \in [n]$

$$|\xi_i| = |Y_i - \hat{Q}(X_i)| \leq |Y_i| + |\hat{Q}(X_i)| \leq 2V_{\max}$$

Thus

$$\begin{aligned} \xi \cdot (f_j - TQ) &\leq \|\xi\| \cdot \|f_j - TQ\| \cdot \|f_j - TQ\|^{-1} \\ &\leq 2V_{\max} \end{aligned}$$

□

Combining eq. (33), eq. (35), eq. (36) and eq. (46)

$$\mathbb{E}\|\hat{Q} - TQ\|^2 \leq \mathbb{E}\|f - TQ\|^2 + 4nV_{\max}\delta + \left(\sqrt{\mathbb{E}\|\hat{Q} - TQ\|^2} + \sqrt{n}\delta\right) 2V_{\max} \quad (47)$$

$$= 2\sqrt{\mathbb{E}\|\hat{Q} - TQ\|^2} V_{\max} + 6n\delta V_{\max} + \mathbb{E}\|f - TQ\|^2 \quad (48)$$

Lemma 3. Let $a, b > 0, \kappa \in (0, 1]$ then

$$a^2 \leq 2ab + c \implies a^2 \leq (1 + \kappa)^2 b^2 / \kappa + (1 + \kappa)c$$

Proof. $0 \leq (x - y) = x^2 + y^2 - 2xy \implies 2xy \leq x^2 + y^2$ for any $x, y \in \mathbb{R}$ so

$$\begin{aligned} 2ab &= 2\sqrt{\frac{\kappa}{1 + \kappa}} a \sqrt{\frac{1 + \kappa}{\kappa}} b \\ &\leq \frac{\kappa}{1 + \kappa} a^2 + \frac{1 + \kappa}{\kappa} b^2 \end{aligned}$$

□

By lemma 3 applied to eq. (48)

$$\frac{1}{n} \mathbb{E} \left\| \widehat{Q} - TQ \right\|^2 \leq \frac{(1+\kappa)^2}{\kappa} \frac{1}{n} V_{\max}^2 + (1+\kappa) \left(6\delta V_{\max} + \frac{1}{n} \mathbb{E} \|f - TQ\|^2 \right) \quad (49)$$

$$\leq \frac{(1+\kappa)^2}{\kappa} \frac{1}{n} V_{\max}^2 + (1+\kappa) \left(6\delta V_{\max} + \inf_{f \in \mathcal{F}} \frac{1}{n} \mathbb{E} \|f - TQ\|^2 \right) \quad (50)$$

Here we can take infimum since eq. (49) holds for all $f \in \mathcal{F}$.

Step 2 Here we link up $\left\| \widehat{Q} - TQ \right\|_{\sigma}^2$ with $\mathbb{E} \frac{1}{n} \left\| \widehat{Q} - TQ \right\|^2$.

$$\left| \left(\widehat{Q}(x) - TQ(x) \right)^2 - \left(f_{k^*}(x) - TQ(x) \right)^2 \right| = \left| \widehat{Q}(x) - f_{k^*}(x) \right| \cdot \left| \widehat{Q}(x) + f_{k^*}(x) - 2TQ(x) \right| \quad (51)$$

$$\leq 4V_{\max} \delta \quad (52)$$

Using this twice we can say

$$\left(\widehat{Q}(\tilde{X}_i) - TQ(\tilde{X}_i) \right)^2 \quad (53)$$

$$\leq \left(\widehat{Q}(\tilde{X}_i) - TQ(\tilde{X}_i) \right)^2 - \left(f_{k^*}(\tilde{X}_i) - TQ(\tilde{X}_i) \right)^2 + \left(f_{k^*}(\tilde{X}_i) - TQ(\tilde{X}_i) \right)^2 \quad (54)$$

$$\leq \left(f_{k^*}(\tilde{X}_i) - TQ(\tilde{X}_i) \right)^2 + \left(\widehat{Q}(X_i) - TQ(X_i) \right)^2 - \left(\widehat{Q}(X_i) - TQ(X_i) \right)^2 + \left(f_{k^*}(X_i) - TQ(X_i) \right)^2 - \left(f_{k^*}(X_i) - TQ(X_i) \right)^2 + 4V_{\max} \delta \quad (55)$$

$$\leq \left(\widehat{Q}(X_i) - TQ(X_i) \right)^2 + \left(f_{k^*}(\tilde{X}_i) - TQ(\tilde{X}_i) \right)^2 - \left(f_{k^*}(X_i) - TQ(X_i) \right)^2 + 8V_{\max} \delta \quad (56)$$

Thus we get

$$\left\| \widehat{Q} - TQ \right\|_{\sigma}^2 \quad (57)$$

$$= \mathbb{E} \frac{1}{n} \sum_{i=1}^n \left(\widehat{Q}(\tilde{X}_i) - TQ(\tilde{X}_i) \right)^2 \quad (58)$$

$$\leq \mathbb{E} \frac{1}{n} \sum_{i=1}^n \left(\left(\widehat{Q}(X_i) - TQ(X_i) \right)^2 + \left(f_{k^*}(\tilde{X}_i) - TQ(\tilde{X}_i) \right)^2 - \left(f_{k^*}(X_i) - TQ(X_i) \right)^2 \right) + 8V_{\max} \delta \quad (59)$$

$$= \frac{1}{n} \left\| \widehat{Q} - TQ \right\|^2 + \frac{1}{n} \sum_{i=1}^n h_{k^*}(X_i, \tilde{X}_i) + 8V_{\max} \quad (60)$$

Where we define

$$h_j(x, y) := \left(f_j(y) - TQ(y) \right)^2 - \left(f_j(x) - TQ(x) \right)^2 \quad (61)$$

□

8 Appendices

8.1 Various lemmas

Proposition 5. For $x > 0$.

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

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

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

□