

# A Theoretical Analysis of Fitted Q-Iteration

Jacob Harder  
University of Copenhagen

February 20, 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

- $[q] = \{1, \dots, q\}$  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: artificial neural network see definition 2.

## 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  $\pi^*$  such that  $Q^* = Q^{\pi^*}$ . This is the optimal policy - the goal of RL.

Note that  $V^\pi$ ,  $Q^\pi$  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  $\pi$  and taking expectation of  $Q$ .

We define the operator  $T^\pi$  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  $Q^\pi$  making one "iteration" of "propagation of rewards" discounting with  $\gamma$ . Indeed it is easily seen that  $Q^\pi$  is a fixed point for  $T^\pi$ .

A *greedy* policy  $\pi$  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.**  $Q^\pi$  is the unique fixed point of  $T^\pi$ .

*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  $\nu$ , 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  $\nu$  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:

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

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

**Output:** An estimator  $\tilde{Q}_K$  of  $Q^*$  and policy  $\pi_K$

## 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  $(W_\ell, v_\ell)$  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

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

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  $g_{jk} \in C_{t_j}([a_j, b_j]_{t_j}^{t_j}, \beta_j, H_j)$ , 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

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

**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.  $m_\lambda$ . 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  $\nu$  be the sampling distribution from the algorithm, and  $\mu$  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  $Q^{\pi_K}$  be the action-value function corresponding to policy  $\pi_K$  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  $\mu$  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

**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

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

**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  $\nu$
- $(R_i, S'_i)_{i \in [n]}$  be the rewards and next states corresponding to the samples
- $Q \in \mathcal{F}$  be fixed
- $Y_i = R_i + \gamma \max_{a \in \mathcal{A}} Q(S'_i, a)$
- $\hat{Q} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(S_i, A_i) - Y_i)^2$
- $\epsilon \in (0, 1], \delta > 0$  be fixed
- $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$  a minimal  $\delta$ -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

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

*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  $N_0 = |\mathcal{N}(1/n, \mathcal{F}_0, \|\cdot\|_\infty)|$ .  $\square$

**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,  $\tau_1, \dots, \tau_m$  be policies and  $\mu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$  be a probability measure. Then

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

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  $\kappa$  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  $\pi_Q$  is the greedy policy w.r.t.  $Q$ . Denote

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

Note that for any policy  $\pi$ ,  $P^\pi$  is linear and 1-contractive 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  $\pi^*$  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  $\gamma P^{\pi_K}$  is  $\gamma$ -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

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

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)

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

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

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

□

## 8 Appendices

### 8.1 Various lemmas

**Proposition 2.**

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

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

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

□