A Theorical Analysis of Fitted Q-Iteration

Jacob Harder
University of Copenhagen

March 25, 2020

1 Abstract

2 Foreword

The main purpose of this master thesis for me, has been to investigate what has been proven about the convergence of Q-learning algorithms. In particular Q-learning algorithms using (deep) ANNs.

3 Disambiguation

- $\mathrm{id} := x \mapsto x$ the identity function.
- $[\phi] := 1$ when ϕ is true/holds and 0 otherwise, for a logical formula ϕ.
- $[q] := \{1, \dots, q\}$ for $q \in \mathbb{N}$.
- $C_{\mathbb{K}}(X) := \{f : X \to \mathbb{K} \mid f \text{ continuous}\}$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. $C(X) = C_{\mathbb{R}}(X)$
- $C_b(X) := \{f \in C(X) \mid \exists K : \|f\|_\infty < K\}$. I.e. the space of bounded continuous functions.
- ANN: abrv. artificial neural network see definition 4.
- 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.
- $\mathbb{B}_n$ the n-dimensional Borel σ-algebra.
- λ^n the n-dimension Lebesgue measure.

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 avoid the shorthand defined in [YangXieWang, p. 26 bottom]: $\|f\|_n^2 = 1/n \cdot \sum_{i=1}^n f(X_i)^2$. and use p-norms instead. The conversion to my notation thus becomes $\|f\|_n \rightsquigarrow \|f\| / n$.

4 Introduction

Throughout we are working with a background probability space denoted $(\Omega, \mathcal{H}, \mathbb{P})$.

4.1 Reinforcement Learning

Through the probability space, denote this

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

- $(\mathcal{S}, \Sigma_{\mathcal{S}})$ a measurable space of **states**.
- $(\mathcal{A}, \Sigma_{\mathcal{A}})$ a measurable space of **actions**.
- $P(\cdot \mid \cdot) : \Sigma_{\mathcal{S}} \times (\mathcal{S} \times \mathcal{A}) \to [0, 1]$ a probability kernel (of **transition** probabilities).
- $R(\cdot \mid \cdot) : \mathbb{B} \times (\mathcal{S} \times \mathcal{A}) \to [0, 1]$ a probability kernel (of **reward** probabilities).
- $\gamma \in (0, 1)$ a **discount** factor.

In order for this to make sense we here include

Definition 2 (Probability kernel)**.** Let $(X, \Sigma_X), (Y, \Sigma_Y)$ be measurable spaces. A function $\kappa(\cdot \mid \cdot) : \Sigma_Y \times X \to [0, 1]$ is a **probability kernel** provided

- $B \mapsto \kappa(B \mid x) \in \mathcal{P}(\Sigma_Y)$ that is $\kappa(\cdot \mid x)$ is a probability measure for any $x \in X$.
- $x \mapsto \kappa(B \mid x) \in \mathcal{M}(\Sigma_X, \Sigma_Y)$ that is $\kappa(B \mid \cdot)$ is $(\Sigma_X/\Sigma_Y$-) measurable for any $B \in \Sigma_Y$.

Note that both P and R to be stochastic and that R can depend on the action as well as the state. This is perhaps the most general way to define an MDP, generalizing some definitions. Common variations include that R depends on $\mathcal{S}$ only, R is deterministic, or P is deterministic.

Definition 3 (Policy)**.** A (**randomized**, **stationary**) **policy** π is probability kernel

$$\pi(\cdot \mid \cdot) : \Sigma_{\mathcal{A}} \times \mathcal{S} \to [0, 1]$$

An MDP together with a policy and an initial distribution $\mu \in \mathcal{P}(\mathcal{S})$ give rise to a countable stochastic process, $(X_i)_{i \in \mathbb{N}} = (S_i, A_i, R_i)_{i \in \mathbb{N}}$ that is a probability measure P^{π}_{μ}. on $(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{\mathbb{N}}$. See [ref. to Feinberg On Meas. and Repr. of Str. Meas. p. 31-32, and pos. Ionescu Tulcea] for full details of how this is constructed. Intuitively S_1 is drawn from μ, then for all $i \in \mathbb{N}$ A_i is drawn from $\pi(\cdot \mid S_i)$, a reward is then drawn from $R(\cdot \mid S_i, A_i)$, then S_{i+1} is drawn from $P(\cdot \mid S_i, A_i)$ and so on. We let E^{π}_{μ} denote the expectation taken w.r.t P^{π}_{μ}. When μ is a Dirac measure δ_x, i.e. $\mu(\{x\}) = 1$ for some x, we shall generally write x instead of μ, E.g. $\mathbb{E}^{\pi}_{s}$ the expectation taken with respect to $P^{\pi}_{\delta_s}$

4.2 Q-Learning

Fix an MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and a policy π. We assume from now on that any $R \sim R(\cdot \mid s, a)$ is bounded with $|R| \leq R_{\max}$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Generally in value iteration and Q-learning, any function $\mathcal{S} \to \mathbb{R}$ is called a (**state**) **value** function. Similarly any function $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is called a (**state**) **action value** or **Q**- function. The idea behind such functions are (usually) to estimate the cumulative rewards associated with a state or state-action pair and the trajectory of states it can lead to. One has the reduced the problem of finding a good strategy to choosing the action that leads to the highest value. Of course this value of a state will depend strongly on the policy being followed.

The **ideal** value function w.r.t. π, $V^{\pi} : \mathcal{S} \to \mathbb{R}$

$$V^{\pi}(s) := \mathbb{E}^{\pi}_{s} \sum_{t=1}^{\infty} \gamma^t R_t$$

where R_t are the projections of the random process generated from the MDP with starting distribution δ_s onto the rewards of each step. This is well-define because each R_t is bounded by $R_{\max}$ so $V^{\pi}(s) \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = \frac{1}{1-\gamma} R_{\max} := V_{\max} < \infty$. The value function gives the expected accumulated reward when starting in state s and following policy π.

The **ideal** Q-function w.r.t. π, $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$Q^\pi(s,a) := r(s,a) + \gamma \mathbb{E}(V^\pi(S) \mid S \sim P(\cdot \mid s,a))$$

where $r(s,a) := \mathbb{E}(R \mid R \sim R(\cdot \mid s,a))$.

One could think that it is a bit superfluous to define both a value and an action value function. According to [todo: ref] the main reason to work with Q-functions is that it is more difficult to work with value function for several reasons. Firstly to know what is the best action a^* given a state s and a value function V, one has to calculate for each action a the distribution of the next state s' and take expectation over $V(s')$. This either requires full knowledge of the transition kernel (this falls outside the so called model-free approaches) or some way of estimating it, and in both cases at computational cost. Whereas Q-functions simply requires finding $\operatorname{argmax}_{a \in \mathcal{A}} Q(s,a)$.

We define the operator P^π

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

Intuitively this operator yields the expected action value function when looking *one step ahead* following the policy π and taking expectation of Q. Note that $\|P^\pi Q\|_\infty \leq \|Q\|_\infty$. $\|\cdot\|_\infty$.

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

$$(T^\pi Q)(s,a) := r(s,a) + \gamma (P^\pi Q)(s,a)$$

The Bellman operator adjusts an action value function Q to look more like Q^π. This is intuitively by making one iteration of reward-propagation discounting with γ. And indeed

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

Proof. Q^π is a fixed point because for any $s \in \mathcal{S}, a \in \mathcal{A}$

$$\begin{aligned}
T^\pi Q^\pi(s,a) &= r(s,a) + \gamma (P^\pi Q^\pi)(s,a) \\
&= r(s,a) + \gamma \mathbb{E}(Q^\pi(S',A') \mid S' \sim P(\cdot \mid s,a), A' \sim \pi(\cdot, S')) \\
&= r(s,a) + \gamma \mathbb{E}(r(S',A') + \gamma \mathbb{E}(V^\pi(S'') \mid S'' \sim (\cdot \mid S',A')) \mid S' \sim P(\cdot \mid s,a), A' \sim \pi(\cdot, S')) \\
&= r(s,a) + \gamma \mathbb{E}(V^\pi(S') \mid S' \sim P(\cdot \mid s,a)) \\
&= Q^\pi(s,a)
\end{aligned}$$

Now since

$$Q^\pi - T^\pi Q = T^\pi Q^\pi - T^\pi Q = \gamma P^\pi (Q^\pi - Q)$$

by induction

$$Q^\pi - (T^\pi)^n Q = (\gamma P^\pi)^n (Q^\pi - Q)$$

And since γP^π contracts to 0, in fact every bounded Q-function converges to Q^π when iteratively applying T^π. In particular Q^π is the only fixed point of T^π. □

If an action value function Q satisfies that $Q(s,\mathcal{A})$ has a greatest value for every $s \in \mathcal{S}$ then we can define **greedy** policy π with respect to Q to be a policy choosing 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 $\pi = \pi_Q$.

Let π_0 be a policy and $Q_0 = Q^{\pi_0}$ be its ideal Q-function. One can now consider the greedy policy $\pi_1 = \pi_{Q_0}$. Note that

$$\begin{aligned}
T^{\pi_1} Q^{\pi_0} &= r(s,a) + \gamma \mathbb{E}\left(\sup_{a' \in \mathcal{A}} Q^{\pi_0}(S',a') \mid S' \sim P(\cdot \mid s,a) \right) && (1) \\
&\geq r(s,a) + \gamma \mathbb{E}\left(Q^{\pi_0}(S',A') \mid S' \sim P(\cdot \mid s,a), A' \sim \pi_0(\cdot \mid S') \right) && (2) \\
&= Q^{\pi_0} && (3)
\end{aligned}$$

so applying T^{π_1} iteratively on Q_0 creates a monotonically increasing sequence of Q-functions which by proposition 1 converge to Q^{π_1}, proving that $Q^{\pi_1} \geq Q^{\pi_0}$. One can then repeat the process with π_1 and obtain a increasing sequences of (ideal) Q-functions with associated policies $(Q_0, Q_1, \dots)$, $(\pi_0, \pi_1, \dots)$. Variations of this idea is called **policy iteration** and has been studied

a lot. Variations include stopping the "value iteration" (applying the T^{π_i} operator) at various stages before again updating the policy to be greedy w.r.t. to the next Q-function. An important special case is where we simply alternate between updating the policy and the Q-function in every step. We can capture this in a single operator called the Bellman *optimality* operator T, defined as

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

The optimal Q-function is defined as

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

where the supremum is taken over all policies.

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.

4.3 Artificial Neural Networks

Definition 4. 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.

Definition 5 (Sparse ReLU Networks)**.** 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

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

Here $\widetilde{W}_\ell = (W_\ell, v_\ell)$. The set of them we denote $\mathcal{F}(s, V)$.

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:

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

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 Measure Theory

We are mostly concerned with a random process

$$(Z_i)_{i=1}^K = (S_i, A_i, R_i)_{i=1}^K \in (\mathcal{S} \times \mathcal{A} \times (0, R_{\max}))^K \tag{4}$$

where $\mathcal{S} \subseteq \mathbb{R}^d$ is compact and $\mathcal{A}$ is finite, so we can model this as a discrete (and finite) time random process in a compact subset of $\mathbb{R}^{d+1}$ having the Markov property, namely that

$$\mathbb{P}(Z_j \in A \mid Z_{j-1}, \ldots, Z_1) = \mathbb{P}(Z_j \in A \mid Z_{j-1}) \tag{5}$$

6 Assumptions

6.1 Assumption 1: Hölder Smoothness

Definition 6 (Hölder smoothness)**.** For $f : \mathcal{S} \to \mathbb{R}$ we define

$$\|f\|_{C_r} := \sum_{|\alpha|<\beta} \|\partial^\alpha f\|_\infty + \sum_{\|\alpha\|_1=\lfloor\beta\rfloor} \sup_{x\neq y} \frac{|\partial^\alpha(f(x) - f(y))|}{\|x - y\|_\infty^{\beta-\lfloor\beta\rfloor}} \tag{6}$$

Where $\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}_0^r$. And ∂^k is the partial derivative w.r.t. the kth variable. If $\|f\|_{C_r} < \infty$ then f is **Hölder smooth**. Given a compact subset $\mathcal{D} \subseteq \mathbb{R}^r$ the space of Hölder smooth functions on $\mathcal{D}$ with norm bounded by $H > 0$ is denoted

$$C_r(\mathcal{D}, \beta, H) := \left\{ f : \mathcal{D} \to \mathbb{R} \;\middle|\; \|f\|_{C_r} \leq H \right\}$$

Definition 7. 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 Hölder 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 8. Define

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

and

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

Assumption 1. $T\mathcal{F}_0 \subseteq \mathcal{G}_0$. I.e. t is assumed that $Tf \in \mathcal{G}_0$ for any $f \in \mathcal{F}_0$, so when using the Bellman optimality operator on our sparse ReLU networks, we should stay in the class of compositions of Holder smooth functions.

If also $\mathcal{G}_\prime$ is well approximated by functions in $\mathcal{F}_0$ then this assumption implies that $\mathcal{F}_0$ is approximately closed under the Bellman operator T and thus that Q^* is close to $\mathcal{F}_0$.

We now look at a case where $\mathcal{D} = [0, 1]^r$, $q = 1$ and both the expected reward function and transition kernel is Hölder smooth.

Proposition 2. Assume for all $a \in \mathcal{A}$ that $P(s, a)$ and $R(s, a)$ are absolutely continuous w.r.t. λ^k and for all $s' \in \mathcal{S}$ that $s \mapsto P\left(s' \;\middle|\; s, a\right)$ and $s \mapsto \mathbb{E}R(s, a)$ are both in $C_r([0, 1]^r, \beta, H)$. Then $T\mathcal{F}_0 \subseteq C_r([0, 1]^r, \beta, (1 + \gamma V_{\max})H)$.

Proof. Let $f \in \mathcal{F}_0$ and $\alpha \in \mathbb{N}_0^r$. Observe that

$$\begin{aligned}
\partial^\alpha(Tf)(s, a) &= \partial_s^\alpha \left(\mathbb{E}R(s, a)\right) + \gamma \int_{\mathcal{S}} \partial_s^\alpha \left[\max_{a'\in\mathcal{A}} f(s', a') P\left(s' \;\middle|\; s, a\right)\right] \mathrm{d}s' \\
&\leq \partial_s^\alpha \left(\mathbb{E}R(s, a)\right) + \gamma V_{\max} \sup_{s'\in\mathcal{S}} \partial_s^\alpha P\left(s' \;\middle|\; s, a\right)
\end{aligned}$$

similarly

$$\partial^\alpha(Tf)(s,a) - \partial^\alpha(Tf)(s',a) \le \partial_s^\alpha\left(\mathbb{E}R(s,a)\right) - \partial_s^\alpha\left(\mathbb{E}R(s',a)\right) + \gamma V_{\max} \sup_{s''\in\mathcal{S}} \left(\partial_s^\alpha P(s'' \mid s,a) - \partial_s^\alpha P(s'' \mid s',a)\right)$$

Thus

$$\begin{aligned}\|Tf\|_{C_r} &\le \sum_{|\alpha|<\beta} \left(\left\|\partial^\alpha \mathbb{E}R(\cdot,a)\right\|_\infty + \gamma V_{\max} \sup_{s\in\mathcal{S}} \left\|\partial^\alpha P(s \mid \cdot, a)\right\|_\infty \right) \\ &+ \sum_{\|\alpha\|_1=\lfloor\beta\rfloor} \sup_{x\neq y} \left(\frac{\left|\partial^\alpha(\mathbb{E}R(x,a) - \mathbb{E}R(y,a))\right|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}} + \gamma V_{\max} \sup_{s\in\mathcal{S}} \frac{\left|\partial^\alpha(\mathbb{E}P(s \mid x,a) - \mathbb{E}R(s \mid y,a))\right|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}} \right) \\ &\le H + \gamma V_{\max} H = (1+\gamma V_{\max})H\end{aligned}$$

□

6.2 Assumption 2: Concentration Coefficients

Definition 9 (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

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

7 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

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

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.

8 Proofs

Proof of main theorem. Using theorem 2 we get

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \le \frac{2\phi_{\mu,\nu}\gamma}{(1-\gamma)^2} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} R_{\max} \tag{7}$$

where $\varepsilon_{\max} = \max_{k\in[K]} \left\| T\widetilde{Q}_{k-1} - \widetilde{Q}_k \right\|_{2,\nu}$. Using theorem 3 with $Q = \widetilde{Q}_{k-1}$, $\mathcal{F} = \mathcal{F}_0$, $\epsilon = 1$ and $\delta = 1/n$, we get

$$\varepsilon_{\max} \le 6n^{-1}C_2^2 V_{\max}^2 \log(N_\delta) + 2\omega(\mathcal{F}_0) + 8\sqrt{2}V_{\max} n^{-1/2}\sqrt{\log N_0} + 16 V_{\max} n^{-1} \tag{8}$$

where $N_0 = \left|\mathcal{N}(1/n, \mathcal{F}_0, \|\cdot\|_\infty)\right|$. The remains only to wound $\omega(\mathcal{F}_0)$ and N_0, starting with $\omega(\mathcal{F}_0)$.

Step 1. We want to employ the following lemma by [Schmidt-Hieber 2019, thm. 5, p. 22]

Lemma 1 (Approximation of Hölder Smooth Functions)**.** Let $m, M \in \mathbb{Z}_+$ with $N \geq \max\{(\beta+1)^r, (H+1)e^r\}$, $L = 8 + (m+5)(1+\lceil \log_2(r+\beta)\rceil)$, $d_0 = r, d_j = 6(r+\lceil\beta\rceil)N, d_{L+1} = 1$. Then for any $g \in \mathcal{C}_r\left([0,1]^r, \beta, H\right)$ there exists a ReLU network $f \in \mathcal{F}\left(L, \{d_j\}_{j=0}^{L+1}, s, \infty\right)$ with $s \leq 141(r+\beta+1)^{3+r}N(m+6)$ such that

$$\|f-g\|_\infty \leq (2H+1)6^r N(1+r^2+\beta^2)2^{-m} + H3^\beta N^{-\beta/r}$$

to each Hölder smooth part of g and then piece it together somehow, using that ReLU networks are easily stitched together into bigger ReLU networks. Therefore the first step is to refit our Hölder Smooth compositions in $\mathcal{G}_0$ to be defined on a hyper-cube instead. This is a relatively simple procedure:

Let $f \in \mathcal{G}_0$ then $f(\cdot, a) \in \mathcal{G}(\{p_j, t_j, \beta_j, H_j\})$ for all $a \in \mathcal{A}$. Therefore $f(\cdot, a) = g_q \circ \cdots \circ g_1$ where the (sub-)components $(g_{jk})_{k=1}^{p_{j+1}} = g_j$ satisfy

$$g_{jk} \in C_{t_j}([a_j, b_j]^{t_j}, \beta_j, H_j), \qquad j \in [q], k \in [p_{j+1}] \tag{9}$$

Here $a_1 = 0, b_1 = 1$ and, $a_j < b_j \in \mathbb{R}$ are some real numbers for $2 \leq j \leq q$. Notice that the Hölder smooth condition implies that $g_{jk}([a_j, b_j]^{t_j}) \subseteq [-H_j, H_j]$. Define

$$\begin{aligned} h_1 =& g_1/(2H_1) + 1/2 \\ h_j(u) =& g_j(2H_{j-1}u - H_{j-1})/(2H_j) + 1/2, \qquad j \in \{2, \ldots, q-1\} \\ h_q(u) =& g_q(2H_{q-1}u - H_{q-1}) \end{aligned} \tag{10}$$

Then $g_q \circ \cdots \circ g_1 = h_q \circ \cdots \circ h_1$ and

$$\begin{aligned} h_{1k} &\in C_{t_1}([0,1]^{t_1}, \beta_1, 1) \\ h_{jk} &\in C_{t_j}([0,1]^{t_j}, \beta_j, (2H_{j-1})^{\beta_j}), \qquad j \in \{2, \ldots, q-1\} \\ h_q &\in C_{t_q}([0,1]^{t_q}, \beta_q, H_q(2H_{q-1})^{\beta_q}) \end{aligned} \tag{11}$$

Define $\eta = \log\left((2W+1)6^{t_j}N/(W3^{\beta_j}N^{-\beta_j/t_j})\right)$, and $m = \eta\lceil \log_2 n\rceil$

$$W := \max\left(\left\{(2H_{j-1})^{\beta_j} \mid 1 \leq j \leq q-1\right\} \cup \left\{H_q(2H_{q-1})^{\beta_q}, 1\right\}\right) \tag{12}$$

By lemma 1 there exists a ReLU network

$$\widehat{h}_{jk} \in \mathcal{F}\left(L_j + 2, \left\{t_j, \widetilde{d}_j p_{j+1}, \ldots, \widetilde{d}_j p_{j+1}, p_{j+1}\right\}, (\widetilde{s}_j + 4)\cdot p_{j+1}\right) \tag{13}$$

where $\widetilde{d}_j = 6(t_j + \lceil\beta_j\rceil)N$ and $\widetilde{s}_j \leq 141(t_j+\beta_j+1)^{3+t_j}N(m+6)$ such that

$$\left\|\widehat{h}_{jk} - h_{jk}\right\|_\infty \leq (2W+1)6^{t_j}N2^{-m} + W3^{\beta_j}N^{-\beta_j/t_j} \leq 2W3^{-\beta_j}N^{-\beta_j/t_j} \tag{14}$$

since $n \leq 4 > e$. Since h_{j+1} is defined on $[0,1]^{t_{j+1}}$ but $\widetilde{h}_j$ takes values in $\mathbb{R}$ we need to restrict $\widetilde{h}_j$ somehow to stitch the two together (by function composition). This is easily done by

Lemma 2. Restriction to $[0,1]$ is expressible as a two-layer ReLU network with 4 non-zero weights.

Proof. Namely $\tau(u) = 1 - (1-u)_+ = \min\left\{\max\{u, 0\}, 1\right\}$. □

Now define $\widetilde{h}_{jk} = \tau \circ \widehat{h}_{jk}$ (and $\widetilde{h}_j = (\widetilde{h}_{jk})_{k\in[p_{j+1}]}$). Then

$$\widetilde{h}_{jk} \in \mathcal{F}\left(L_j + 2, \left\{t_j, \widetilde{d}_j, \ldots, \widetilde{d}_j, 1\right\}, (\widetilde{s}_j + 4)p_{j+1}\right) \tag{15}$$

and since $h_{jk}([0,1]^{t_j}) \in [0,1]$ by eq. (14)

$$\begin{aligned} \left\|\widetilde{h}_{jk} - h_{jk}\right\|_\infty &= \left\|\tau\circ\widehat{h}_{jk} - \tau\circ h_{jk}\right\|_\infty && (16) \\ &\leq \left\|\widehat{h}_{jk} - h_{jk}\right\|_\infty && (17) \\ &\leq 2W3^{-\beta_j}N^{-\beta_j/t_j} && (18) \end{aligned}$$

Step 2. Now define $\widetilde{f} : \mathcal{S} \to \mathbb{R}$ as $\widetilde{f} = \widetilde{h}_1 \circ \cdots \circ \widehat{h}_1$. If we set $\widetilde{L} := \sum_{j=1}^{q}(L_j + 2)$, $\widetilde{d} := \max j \in [q]\widetilde{d}_j p_{j+1}$ and $\widetilde{s} := \sum_{j=1}^{q}(\widetilde{s}_j + 4)p_{j+1}$. Then $\widetilde{f} \in \mathcal{F}\left(\widetilde{L}, \left\{r, \widetilde{d}, \ldots, \widetilde{d}, 1\right\}, \widetilde{s}\right)$.

□

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

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

Where

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

Lemma 3. $TQ \ge T^\pi Q$ 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.

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

□

Lemma 4. 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 9.

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

$$\begin{aligned}
\mathbb{E}_\mu[(P^{\tau_m} \ldots P^{\tau_1})(f)] &= \int (P^{\tau_m} \ldots P^{\tau_1})(f)\, \mathrm{d}\mu && (19)\\
&= \int f\, \mathrm{d}(P^{\tau_m} \ldots P^{\tau_1}\mu) && (20)\\
&= \int f \frac{\mathrm{d}(P^{\tau_m} \ldots P^{\tau_1}\mu)}{\mathrm{d}\nu}\, \mathrm{d}\nu && (21)\\
&\le \left\| \frac{\mathrm{d}(P^{\tau_m} \ldots P^{\tau_1}\mu)}{\mathrm{d}\nu} \right\|_{2,\nu} \cdot \|f\|_{2,\nu} && (22)\\
&\le \kappa(m, \mu, \nu)\|f\|_{2,\nu} && (23)
\end{aligned}$$

Where eq. (21) is due to the Radon-Nikodym theorem and eq. (22) 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

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

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 $P^\pi f \geq P^\pi f'$.

The proof consists of four steps.

Step 1 We start by relating $Q^* - Q^{\pi_K}$, the quantity of interest, to $Q^* - \widetilde{Q}_K$, which is more related to the output of the algorithm. Using lemma 3 we can make the upper bound

$$\begin{aligned}
Q^* - Q^{\pi_K} &= T^{\pi^*} Q^* - T^{\pi_K} Q^{\pi_K} \\
&= T^{\pi^*} Q^* + (T^{\pi^*} \widetilde{Q}_K - T^{\pi^*} \widetilde{Q}_K) + (T\widetilde{Q}_K - T\widetilde{Q}_K) - T^{\pi_K} Q^{\pi_K} \\
&= (T^{\pi^*} \widetilde{Q}_K - T\widetilde{Q}_K) + (T^{\pi^*} Q^* - T^{\pi^*} \widetilde{Q}_K) + (T\widetilde{Q}_K - T^{\pi_K} Q^{\pi_K}) \\
&\leq (T^{\pi^*} Q^* - T^{\pi^*} \widetilde{Q}_K) + (T\widetilde{Q}_K - T^{\pi_K} Q^{\pi_K}) \\
&= (T^{\pi^*} Q^* - T^{\pi^*} \widetilde{Q}_K) + (T^{\pi_K} \widetilde{Q}_K - T^{\pi_K} Q^{\pi_K}) \\
&= \gamma P^{\pi^*} (Q^* - \widetilde{Q}_K) + \gamma P^{\pi_K} (\widetilde{Q}_K - Q^{\pi_K}) \\
&= \gamma (P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}_K) + \gamma P^{\pi_K} (Q^* - Q^{\pi_K})
\end{aligned} \tag{24}$$

This implies

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

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. (24) to obtain

$$Q^* - Q^{\pi_K} \leq \gamma U^{-1} (P^{\pi^*} (Q^* - \widetilde{Q}_K) - P^{\pi_K} (Q^* - \widetilde{Q}_K)) \tag{25}$$

Step 2 Using lemma 3 for any $i \in [K]$ we can get an upper bound

$$\begin{aligned}
Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T^{\pi^*} \widetilde{Q}_i - T^{\pi^*} \widetilde{Q}_i) \\
&= (Q^* - T^{\pi^*} \widetilde{Q}_i) + (T\widetilde{Q}_i - \widetilde{Q}_{i+1}) + (T^{\pi^*} \widetilde{Q}_i - T\widetilde{Q}_i) \\
&= (T^{\pi^*} Q^* - T^{\pi^*} \widetilde{Q}_i) + \varrho_{i+1} + (T^{\pi^*} \widetilde{Q}_i - T\widetilde{Q}_i) \\
&\leq T^{\pi^*} Q^* - T^{\pi^*} \widetilde{Q}_i + \varrho_{i+1} \\
&= \gamma P^{\pi^*} (Q^* - \widetilde{Q}_i) + \varrho_{i+1}
\end{aligned} \tag{26}$$

and a lower bound

$$\begin{aligned}
Q^* - \widetilde{Q}_{i+1} &= Q^* + (T\widetilde{Q}_i - T\widetilde{Q}_i) - \widetilde{Q}_{i+1} + (T^{\pi_i} Q^* - T^{\pi_i} Q^*) \\
&= (T^{\pi_i} Q^* - T^{\pi_i} \widetilde{Q}_i) + \varrho_{i+1} + (TQ^* - T^{\pi_i} Q^*) \\
&\geq T^{\pi_i} Q^* - T^{\pi_i} \widetilde{Q}_i + \varrho_{i+1} \\
&= \gamma P^{\pi_i} (Q^* - \widetilde{Q}_i) + \varrho_{i+1}
\end{aligned} \tag{27}$$

Applying eq. (26) and eq. (27) iteratively we get

$$Q^* - \widetilde{Q}_K \leq \gamma^K (P^{\pi^*})^K (Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi^*})^{K-1-i} \varrho_{i+1} \tag{28}$$

and

$$Q^* - \widetilde{Q}_K \geq \gamma^K (P^{\pi_{K-1}} \dots P^{\pi_0})(Q^* - \widetilde{Q}_0) + \sum_{i=0}^{K-1} \gamma^{K-1-i} (P^{\pi_{K-1}} \dots P^{\pi_{i+1}}) \varrho_{i+1} \tag{29}$$

Step 3 Combining eq. (28) and eq. (29) with eq. (25) we get

$$\begin{aligned}
Q^* - Q^{\pi_K} \leq U^{-1} \Bigg(& \gamma^{K+1} ((P^{\pi^*})^{K+1} - P^{\pi_K} \dots P^{\pi_0})(Q^* - \widetilde{Q}_0) \\
& + \sum_{i=0}^{K-1} \gamma^{K-i} ((P^*)^{K-i} - P^{\pi_K} \dots P^{\pi_{i+1}}) \varrho_{i+1} \Bigg)
\end{aligned} \tag{30}$$

For shorthand define constants

$$\alpha_i = \frac{(1-\gamma)\gamma^{K-i-1}}{1-\gamma^{K+1}} \quad \text{for } 0 \le i \le K-1 \text{ and } \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} \tag{31}$$

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

$$O_i = (1-\gamma)/2U^{-1}[(P^{\pi^*})^{K-i} + (P^{\pi_K} \dots P^{\pi_{i+1}})] \tag{32}$$

$$O_K = (1-\gamma)/2U^{-1}[(P^{\pi^*})^{K+1} + (P^{\pi_K} \dots P^{\pi_0})] \tag{33}$$

Then by eq. (30)

$$|Q^* - Q^{\pi_K}| \le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{i=0}^{K-1} \alpha_i O_i |\varrho_{i+1}| + \alpha_K O_K \left| Q^* - \widetilde{Q}_0 \right| \right] \tag{34}$$

So by linearity of expectation

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} = \mathbb{E}_\mu |Q^* - Q^{\pi_K}| \tag{35}$$

$$\le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{i=0}^{K-1} \alpha_i \mathbb{E}_\mu (O_i |\varrho_{i+1}|) + \alpha_K \mathbb{E}_\mu (O_K \left| Q^* - \widetilde{Q}_0 \right|) \right] \tag{36}$$

With the bound on rewards we (crudely) estimate

$$\mathbb{E}_\mu O_K \left| Q^* - \widetilde{Q}_0 \right| \le 2V_{\max} = 2R_{\max}/(1-\gamma) \tag{37}$$

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

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

$$\mathbb{E}_\mu(O_i|\varrho_{i+1}|) \tag{38}$$

$$= \frac{1-\gamma}{2} \mathbb{E}_\mu \left(U^{-1}[(P^{\pi_K})^{K-i} + P^{\pi_K} \dots P^{\pi_{i+1}}] |\varrho_{i+1}| \right) \tag{39}$$

$$= \frac{1-\gamma}{2} \mathbb{E}_\mu \left(\sum_{j=0}^{\infty} [(P^{\pi_K})^j (P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1} P^{\pi_{K-1}} \dots P^{\pi_{i+1}}] |\varrho_{i+1}| \right) \tag{40}$$

$$= \frac{1-\gamma}{2} \sum_{j=0}^{\infty} \mathbb{E}_\mu \left([(P^{\pi_K})^j (P^{\pi_K})^{K-i} + (P^{\pi_K})^{j+1} P^{\pi_{K-1}} \dots P^{\pi_{i+1}}] |\varrho_{i+1}| \right) \tag{41}$$

Notice that there are $K-i+j$ P-operators on both terms in the sum. Therefore were can employ lemma 4 twice. Moreover define $\varepsilon_{\max} = \max_{i \in [K]} \|\varrho_i\|_{2,\nu}$. Then

$$\mathbb{E}_\mu(O_i|\varrho_{i+1}|) \le (1-\gamma) \sum_{j=0}^{\infty} \gamma^j \kappa(K-i+j; \mu, \nu) \|\varrho_{i+1}\|_{2,\nu}$$

$$\le \varepsilon_{\max}(1-\gamma) \sum_{j=0}^{\infty} \gamma^j \kappa(K-i+j; \mu, \nu) \tag{42}$$

Using eq. (36), eq. (37) and eq. (42)

$$\|Q^* - Q^{\pi_K}\|_{1,\mu} \le \frac{2\gamma(1-\gamma^{K+1})}{1-\gamma} \left[\sum_{i=0}^{K-1} \sum_{j=0}^{\infty} \alpha_i \gamma^j \kappa(K-i+j; \mu, \nu) \right] \varepsilon_{\max} + \frac{4\gamma(1-\gamma^{K+1})}{(1-\gamma)^3} \alpha_K R_{\max} \tag{43}$$

Focusing on the first term on RHS of eq. (43), if we then we can take the norm out of the sum as a constant. We are left with

$$\begin{aligned}
&\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\alpha_i\gamma^j\kappa(K-i+j;\mu,\nu)\\
&=\sum_{i=0}^{K-1}\sum_{j=0}^{\infty}\frac{(1-\gamma)\gamma^{K-i+j-1}}{1-\gamma^{K+1}}\kappa(K-i+j;\mu,\nu)\\
&=\frac{1-\gamma}{1-\gamma^{K+1}}\sum_{j=0}^{\infty}\sum_{i=0}^{K-1}\gamma^{K-i+j-1}\kappa(K-i+j;\mu,\nu)\\
&\leq\frac{1-\gamma}{1-\gamma^{K+1}}\sum_{m=0}^{\infty}\gamma^{m-1}\cdot m\cdot\kappa(m;\mu,\nu)\\
&\leq\frac{1}{1-\gamma^{K+1}(1-\gamma)}\phi_{\mu,\nu} \qquad (44)
\end{aligned}$$

Where the last inequality is due to assumption 2. Combining eq. (43) and eq. (44) we arrive at

$$\|Q^*-Q^{\pi_K}\|_{1,\mu}\leq\frac{2\gamma\cdot\phi_{\mu,\nu}}{(1-\gamma)^2}\cdot\varepsilon_{\max}+\frac{4\gamma^{K+1}}{(1-\gamma)^2}\cdot R_{\max} \qquad (45)$$

□

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
- $\mathcal{G}=T(\mathcal{F})$ the class of functions obtainable by applying T to some function in $\mathcal{F}$.
- $\nu\in\mathcal{P}(\mathcal{S},\mathcal{A})$ 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
- $Y_i=R_i+\gamma\max_{a\in\mathcal{A}}Q(S_i',a)$
- $\widehat{Q}=\operatorname{argmin}_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(f(S_i,A_i)-Y_i)^2$
- $\kappa\in(0,1],\ \delta>0$ be fixed
- $\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_\infty)$ a minimal δ-covering of $\mathcal{F}$ w.r.t. $\|\cdot\|_\infty$
- $N_\delta=\left|\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_\infty)\right|$ the number of elements in this covering

Then

$$\begin{aligned}
\left\|\widehat{Q}-TQ\right\|_\nu^2\leq&\frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2V_{\max}^2\log(N_\delta)+(1+\kappa)\left(\delta C_2^2V_{\max}^2\log(N_\delta)+\omega(\mathcal{F})\right)\\
&+8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta}+8V_{\max}(n^{-1}+\delta)
\end{aligned}$$

Where

$$\omega(\mathcal{F})=\sup_{g\in\mathcal{G}}\inf_{f\in\mathcal{F}}\frac{1}{n}\mathbb{E}\|f-TQ\|^2$$

where

Lemma 5 (Rotation invariance)**.** Let $(X_i)_{i=1}^n$ be independent, centered and sub-gaussian. Then $\sum_{i=1}^n X_i$ is centered and sub-gaussian with

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2}^2\leq C\sum_{i=1}^n\|X_i\|_{\psi_2}^2$$

Proof. See [Vershynin 2010, p. 12]. □

Definition 10 (Sub-exponential norm)**.** For a random variable define

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} \|X\|_p$$

called the sub-exponential norm, said to 'exist' if finite. In that case X is said to be 'sub-exponential'.

Lemma 6 (Sub-gaussian squared is sub-exponential)**.** A random variable X is sub-gaussian if and only if X^2 is sub-exponential and

$$\|X\|_{\psi_2}^2 \leq \left\|X^2\right\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2$$

Proof. See [Vershynin 2010, p. 14] □

Proposition 3. 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 $v = (v_1, \ldots, v_n)$. Cauchy-Schwarz applied to some vector w and $(1, \ldots, 1)$ yields

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

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

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

□

Theorem 4 (Bernstein's inequality)**.** Suppose $U_1, \ldots, U_n$ are independent with $\mathbb{E}U_i = 0, |U_i| \leq M$ for all $i \in [n]$. Then for all $t > 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^n U_i\right| \geq t\right) \leq \exp\left(\frac{-t^2}{2/3Mt + 2\sigma^2}\right)$$

where $\sigma^2 = \sum_{i=1}^n V(U_i)$.

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 $\|f\| = \left\|f(X)\right\|$.

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

$$\|Y\|^2 + \left\|\widehat{Q}\right\|^2 - 2Y \cdot \widehat{Q} \leq \|Y\|^2 + \|f\|^2 - 2Y \cdot f \tag{46}$$

$$\iff \left\|\widehat{Q}\right\|^2 + \|TQ\|^2 - 2\widehat{Q} \cdot TQ \leq \|f\|^2 + \|TQ\|^2 - 2f \cdot TQ + 2Y \cdot \widehat{Q} - 2Y \cdot f - 2\widehat{Q} \cdot TQ + 2f \cdot TQ \tag{47}$$

$$\iff \left\|\widehat{Q} - TQ\right\|^2 \leq \|f - TQ\|^2 + 2(Y - TQ) \cdot (\widehat{Q} - f) \tag{48}$$

$$\iff \left\|\widehat{Q} - TQ\right\|^2 \leq \|f - TQ\|^2 + 2\xi \cdot (\widehat{Q} - f) \tag{49}$$

Where $\xi_i := Y_i - TQ(X_i)$ and $\xi := (\xi_1, \ldots, \xi_n)$. Let $\Sigma = (X_1, \ldots, X_n)^{-1}(\mathbb{B}_n) \in \mathcal{H}$ be the σ-algebra generated by the samples. Now we proof a minor lemma

Proposition 4. $\mathbb{E}(\xi_i \mid \Sigma) = 0$ and thus $\mathbb{E}(\xi_i g(X_i)) = 0$ for any function $g : \mathbb{R} \to \mathbb{R}$.

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

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

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

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

Therefore $\mathbb{E}(\xi_i \mid \Sigma) = 0$. □

By this lemma we can deduce

$$\mathbb{E}\left(\xi \cdot (\widehat{Q} - f)\right) = \mathbb{E}\left(\xi \cdot (\widehat{Q} - TQ)\right) \tag{50}$$

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

$$\left|\mathbb{E}\left(\xi \cdot (\widehat{Q} - TQ)\right)\right| \leq \left|\mathbb{E}\left(\xi \cdot (\widehat{Q} - f_{k^*})\right)\right| + \left|\mathbb{E}\left(\xi \cdot (f_{k^*} - TQ)\right)\right| \tag{51}$$

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

$$\left|\mathbb{E}\xi \cdot (\widehat{Q} - f_{k^*})\right| \leq \mathbb{E}\left(\|\xi\| \left\|\widehat{Q} - f_{k^*}\right\|\right) \leq \mathbb{E}(\|\xi\|)\sqrt{n}\delta \leq 2nV_{\max}\delta \tag{52}$$

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

$$\left\|\widehat{Q} - f_{k^*}\right\|^2 = \sum_{i=1}^n (\widehat{Q}(X_i) - f_{k^*}(X_i))^2 \leq \sum_{i=1}^n \delta^2 = n\delta^2 \tag{53}$$

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

$$\|\xi\|^2 = \sum_{i=1}^n (Y_i - TQ(X_i))^2 \leq \sum_{i=1}^n (2V_{\max})^2 = 4V_{\max}^2 n \tag{54}$$

To bound the second term in eq. (51) define

$$Z_j := \xi \cdot (f_j - TQ)\left\|f_j - TQ\right\|^{-1} \tag{55}$$

Note that since ξ_i are centered Z_j. For a sub-σ-algebra Σ define the *sub-gaussian* norm by

Definition 11 (Sub-gaussian norm)**.**

$$\|W\|_{\psi_2,\Sigma} := \sup_{p\geq 1} p^{-1/2} \left(\mathbb{E}_\Sigma |W|^p\right)^{1/p}$$

Because of proposition 4 $\xi_i(f_j(X_i) - TQ(X_i))$ is centered for any $i \in [n]$ and

$$\left\|\xi_i(f_j(X_i) - TQ(X_i))\right\|_{\psi_2,\Sigma} \leq 2V_{\max}\left|f_j(X_i) - TQ(X_i)\right| \tag{56}$$

Therefore by lemma 5

$$\begin{aligned}
\left\|Z_j\right\|_{\psi_2,\Sigma}^2 \leq& \left\|f_j - TQ\right\|^{-2} \left\|\sum_{i=1}^n \xi_i(f_j(X_i) - TQ(X_i))\right\|_{\psi_2,\Sigma}^2 && (57)\\
&\leq \left\|f_j - TQ\right\|^{-2} C_1 \sum_{i=1}^n \left\|\xi_i(f_j(X_i) - TQ(X_i))\right\|_{\psi_2,\Sigma}^2 && (58)\\
&\leq \left\|f_j - TQ\right\|^{-2} C_1 \sum_{i=1}^n 4V_{\max}\left|f_j(X_i) - TQ(X_i)\right|^2 && (59)\\
&= 4V_{\max}^2 C_1 && (60)
\end{aligned}$$

Observe that $\|X\|_p \leq \sqrt{p}\ \sup_{p\geq 1}\|X\|_{\psi_2}$. Thus by [Vershynin 2010, p. 11 and Lemma 5.5]

$$\mathbb{E}\exp\left(cZ_j^2/\left\|Z_j\right\|_{\psi_2}^2\right) \leq e \tag{61}$$

so

$$\begin{aligned}
\mathbb{E}\max_{j\in N_\delta} Z_j^2 &= \frac{\max_{j\in[N_\delta]}\left\|Z_j\right\|_{\psi_2}^2}{c}\mathbb{E}\left(\max_{j\in[N_\delta]}\frac{cZ_j^2}{\max_{k\in[N_\delta]}\|Z_k\|_{\psi_2}}\right) && (62)\\
&\leq \frac{4V_{\max}^2 C_1}{c}\mathbb{E}\left(\max_{j\in N_\delta}\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right) && (63)\\
&\leq \frac{4V_{\max}^2 C_1}{c}\log\left(\mathbb{E}\max_{j\in N_\delta}\exp\left(\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right)\right) && (64)\\
&\leq \frac{4V_{\max}^2 C_1}{c}\log\left(\sum_{j\in[N_\delta]}\mathbb{E}\exp\left(\frac{cZ_j^2}{\left\|Z_j\right\|_{\psi_2}}\right)\right) && (65)\\
&\leq \frac{4V_{\max}^2 C_1}{c}\log\left(eN_\delta\right) && (66)\\
&\leq C_2^2 V_{\max}^2 \log(N_\delta) && (67)
\end{aligned}$$

Where $C_2 := \sqrt{8C_1/c}$. Now we can bound

$$\mathbb{E}\left(\xi \cdot (f_{k^*} - TQ)\right) = \mathbb{E}\left(\|f_{k^*} - TQ\||Z_{k^*}|\right) \tag{68}$$

$$\leq \mathbb{E}\left(\left(\left\|\widehat{Q} - TQ\right\| + \left\|\widehat{Q} - f_{k^*}\right\|\right)|Z_{k^*}|\right) \tag{69}$$

$$\leq \mathbb{E}\left(\left(\left\|\widehat{Q} - TQ\right\| + n\delta\right)|Z_{k^*}|\right) \tag{70}$$

$$\leq \left(\mathbb{E}\left(\left\|\widehat{Q} - TQ\right\| + n\delta\right)^2\right)^{1/2}\left(\mathbb{E}Z_{k^*}^2\right)^{1/2} \tag{71}$$

$$\leq \mathbb{E}\left(\left\|\widehat{Q} - TQ\right\| + n\delta\right)\left(\mathbb{E}Z_{k^*}^2\right)^{1/2} \tag{72}$$

$$\leq \left(\sqrt{\mathbb{E}\left\|\widehat{Q} - TQ\right\|_2^2} + n\delta\right)\left(\mathbb{E}Z_{k^*}^2\right)^{1/2} \tag{73}$$

$$\leq \left(\sqrt{\mathbb{E}\left\|\widehat{Q} - TQ\right\|_2^2} + n\delta\right)C_2^2 V_{\max}^2 \log(N_\delta) \tag{74}$$

Where eq. (68) to eq. (69) is by the triangle inequality and eq. (72) to eq. (73) is proposition 3. Combining eq. (49), eq. (51), eq. (52) and eq. (74)

$$\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 \leq \mathbb{E}\|f - TQ\|^2 + 4nV_{\max}\delta + \left(\sqrt{\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2} + \sqrt{n}\delta\right)C_2 V_{\max}\sqrt{\log(N_\delta)} \tag{75}$$

$$= C_2 V_{\max}\sqrt{n\log(N_\delta)}\sqrt{\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2} + nC_2^2\delta V_{\max}^2\log(N_\delta) + \mathbb{E}\|f - TQ\|^2 \tag{76}$$

Lemma 7. 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. $0 \leq (x-y)^2 = x^2 + y^2 - 2xy \implies 2xy \leq x^2 + y^2$ 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 7 applied to eq. (76)

$$\frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 \leq \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2 V_{\max}^2\log(N_\delta) + (1+\kappa)\left(\delta C_2^2 V_{\max}^2\log(N_\delta) + \frac{1}{n}\mathbb{E}\|f - TQ\|^2\right) \tag{77}$$

Since this holds for any $f \in \mathcal{F}$ we can further say

$$\frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 \leq \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2 V_{\max}^2\log(N_\delta) \tag{78}$$

$$+ (1+\kappa)\left(C_2^2 V_{\max}^2\log(N_\delta) + \inf_{f\in\mathcal{F}}\frac{1}{n}\mathbb{E}\|f - TQ\|^2\right)$$

$$\leq \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2 V_{\max}^2\log(N_\delta) + (1+\kappa)\left(C_2^2 V_{\max}^2\log(N_\delta) + \omega(\mathcal{F})\right) \tag{79}$$

Where we take the supremum over $\mathcal{G}$ (recall $TQ \in \mathcal{G}$).

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

$$\left|\left(\widehat{Q}(x) - TQ(x)\right)^2 - \left(f_{k^*}(x) - TQ(x)\right)^2\right| = \left|\widehat{Q}(x) - f_{k^*}(x)\right| \cdot \left|\widehat{Q}(x) + f_{k^*}(x) - 2TQ(x)\right| \tag{80}$$

$$\leq 4V_{\max}\delta \tag{81}$$

Using this twice we can say

$$(\widehat{Q}(\widehat{X}_i) - TQ(\widehat{X}_i)^2 \tag{82}$$

$$\leq (\widehat{Q}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k^*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 + (f_{k^*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 \tag{83}$$

$$\leq (f_{k^*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 + (\widehat{Q}(X_i) - TQ(X_i))^2 - (\widehat{Q}(X_i) - TQ(X_i))^2$$
$$+ (f_{k^*}(X_i) - TQ(X_i))^2 - (f_{k^*}(X_i) - TQ(X_i))^2 + 4V_{\max}\delta \tag{84}$$

$$\leq (\widehat{Q}(X_i) - TQ(X_i))^2 + (f_{k^*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k^*}(X_i) - TQ(X_i))^2 + 8V_{\max}\delta \tag{85}$$

Thus we get

$$\left\|\widehat{Q} - TQ\right\|_\sigma^2 \tag{86}$$

$$= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(\widehat{Q}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 \tag{87}$$

$$\leq \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\left((\widehat{Q}(X_i) - TQ(X_i))^2 + (f_{k^*}(\widetilde{X}_i) - TQ(\widetilde{X}_i))^2 - (f_{k^*}(X_i) - TQ(X_i))^2\right) + 8V_{\max}\delta \tag{88}$$

$$= \frac{1}{n}\left\|\widehat{Q} - TQ\right\|^2 + \frac{1}{n}\sum_{i=1}^{n} h_{k^*}(X_i, \widetilde{X}_i) + 8V_{\max}\delta \tag{89}$$

Where we define

$$h_j(x, y) := \left(f_j(y) - TQ(y)\right)^2 - \left(f_j(x) - TQ(x)\right)^2 \tag{90}$$

For any $j \in [N_\delta]$. Define $\Upsilon = 2V_{\max}$ and

$$T := \max_{j\in[N_\delta]}\left|\sum_{i=1}^{n} h_j(X_i, \widetilde{X}_i)/\Upsilon\right| \tag{91}$$

Then we can bound the middle term in eq. (89)

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} h_{k^*}(X_i, \widetilde{X}_i)\right) \leq \Upsilon/n\mathbb{E}\max_{j\in[N_\delta]}\left(\left|\sum_{i=1}^{n} h_j(X_i, \widetilde{X}_i)/\Upsilon\right|\right) \tag{92}$$

$$\leq \Upsilon/n\mathbb{E}T \tag{93}$$

We want to use Bernsteins inequality (theorem 4) with $U_i = h_j(X_i, \widetilde{X}_i)$. Therefore notice that $\left|h_j\right| \leq \Upsilon^2$ and

$$\mathrm{V}h_j(X_i, \widetilde{X}_i) = 2\mathrm{V}\left(f_j(X_i) - TQ(X_i)\right)^2 \tag{94}$$

$$\leq 2\mathbb{E}\left(f_j(X_i) - TQ(X_i)\right)^4 \tag{95}$$

$$\leq 2\Upsilon^4 \tag{96}$$

so by union bounding for any $u < 6n\Upsilon$ we have

$$\begin{aligned}
\mathbb{E}T =& \int_0^\infty \mathbb{P}(T \geq t) && (97)\\
\leq& u + \int_u^\infty \mathbb{P}(T \geq t)\mathrm{d}t && (98)\\
\leq& u + \int_u^\infty 2N_\delta \exp\left(\frac{-t^2}{2\Upsilon t/3 + 4n\Upsilon^2}\right)\mathrm{d}t && (99)\\
\leq& u + 2N_\delta \int_u^\infty \exp\left(\frac{-t^2}{2\Upsilon^2(t/(3\Upsilon) + 2n)}\right)\mathrm{d}t && (100)\\
\leq& u + 2N_\delta \left(\int_u^{6n\Upsilon} \exp\left(\frac{-t^2}{8n\Upsilon^2}\right)\mathrm{d}t + \int_{6n\Upsilon}^\infty \exp\left(\frac{-t}{4/3\Upsilon}\right)\mathrm{d}t\right) && (101)\\
\leq& u + 2N_\delta \left(\frac{8n\Upsilon}{2u}\exp\left(\frac{-u^2}{8n\Upsilon}\right) + \frac{4\Upsilon}{3}\exp\left(\frac{-24n\Upsilon}{3\Upsilon}\right)\right) && (102)
\end{aligned}$$

where we use lemma 8 from eq. (101) to eq. (102). Now set $u = \Upsilon\sqrt{8n\log N_\delta}$ continuing from eq. (102) we have

$$\begin{aligned}
\cdots =& \Upsilon\sqrt{8n\log N_\delta} + \frac{\Upsilon^2 8nN_\delta}{\Upsilon\sqrt{8n\log N_\delta}}\exp(-\log N_\delta) + 8/3N_\delta\Upsilon\exp(-9/2n) && (103)\\
=& \Upsilon 2\sqrt{2n}\left(\log N_\delta + \frac{1}{\log N_\delta}\right) + 8/3N_\delta e^{-9/2n} && (104)\\
\leq& 4\sqrt{2}\Upsilon\sqrt{n\log N_\delta} + 8/3\Upsilon && (105)
\end{aligned}$$

Inserting eq. (105) and eq. (79) into eq. (89)

$$\begin{aligned}
\left\|\widehat{Q} - TQ\right\|_\nu^2 \leq& \frac{1}{n}\mathbb{E}\left\|\widehat{Q} - TQ\right\|^2 + 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta} + 8V_{\max}(n^{-1} + \delta) && (106)\\
\leq& \frac{(1+\kappa)^2}{\kappa}\frac{1}{n}C_2^2V_{\max}^2\log(N_\delta) + (1+\kappa)\left(\delta C_2^2V_{\max}^2\log(N_\delta) + \omega(\mathcal{F})\right)\\
&+ 8\sqrt{2}V_{\max}n^{-1/2}\sqrt{\log N_\delta} + 8V_{\max}(n^{-1} + \delta) && (107)
\end{aligned}$$

□

9 Appendices

9.1 Various lemmas

Lemma 8. 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

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

□