Fitted Q-iteration in continuous action-space MDPs

Andras Antos, Rémi Munos, Csaba Szepesvari

Fitted Q-iteration in continuous action-space MDPs

András Antos
Computer and Automation Research Inst.
of the Hungarian Academy of Sciences
Kende u. 13-17, Budapest 1111, Hungary
antos@sztaki.hu

Rémi Munos
SequeL team, INRIA Futurs
University of Lille
59653 Villeneuve d'Ascq, France
remi.munos@inria.fr

Csaba Szepesvári*
Department of Computing Science
University of Alberta
Edmonton T6G 2E8, Canada
szepesva@cs.ualberta.ca

Abstract

We consider continuous state, continuous action batch reinforcement learning where the goal is to learn a good policy from a sufficiently rich trajectory generated by some policy. We study a variant of fitted Q-iteration, where the greedy action selection is replaced by searching for a policy in a restricted set of candidate policies by maximizing the average action values. We provide a rigorous analysis of this algorithm, proving what we believe is the first finite-time bound for value-function based algorithms for continuous state and action problems.

1 Introduction

Batch reinforcement learning (RL) refers to the problem of finding a good policy given some fixed set of samples. This problem is highly relevant for industrial applications where data is gathered by following a fixed controller, after which there is no further opportunity to interact with the system. Another very common characteristics of industrial problems is that the space of states and actions is continuous (or has continuous components). In this paper we study such problems. Intriguingly, to our best knowledge there is no known theoretically sound approach to this problem.

Continuous action problems are most often tackled by means of policy search algorithms (e.g. [1, 2, 3]). Although these algorithms tend to assume that they can obtain new samples (either by interacting with the environment or in a simulated environment) they could be adopted to the batch setting.

In this paper, however, we start from value-function based methods [4]. The potential advantage of these methods is that they may make a better use of the structure of the problem by exploiting the recursive nature of the value functions. However, for this they need efficient function approximation algorithms and learning algorithms that are capable of dealing with the imprecisions introduced by the approximate representation of the value functions. This approach can be criticized on the basis that it may run into trouble when the value functions are difficult to represent. Nevertheless, we find it more interesting to study these methods because at least they carry the promise of being more efficient than brute-force search methods.

Value-function based RL is deeply connected to dynamic programming (DP). Most algorithms are relatives to one of the basic DP methods. *Policy iteration* based algorithms keep a policy, compute its

*Also with: Computer and Automation Research Inst. of the Hungarian Academy of Sciences Kende u. 13-17, Budapest 1111, Hungary.

action-value function and then compute a new policy based on the obtained value function. Least-squares policy iteration (LSPI) is a recent algorithm that uses least-squares temporal difference learning to evaluate policies [5]. Another example is the algorithm proposed in [6] that employs a modified Bellman-residual minimization (BRM) criterion. Interestingly, LSPI can be shown to be a special case of this algorithm [7]. The BRM algorithm comes with a finite time performance bound that (under appropriate "richness" conditions on the data and "smoothness" conditions on the problem) shows that in the limit of an indefinite number of samples the loss compared to an optimal policy can be made to converge to zero, i.e., the algorithm is consistent. Since LSPI is a special case of the BRM algorithm, the theoretical results show that the LSPI can also be made consistent. These results, however, heavily exploits that the number of actions is finite: The bounds explicitly depend on the number of actions (in a quadratic fashion) and therefore it is not evident if these algorithms would work without any modifications when the actions-space is infinite.

Another recent algorithm is fitted Q-iteration (FQI) due to Ernst et al. [8], which is a variant of fitted value iteration [9] applied to action-value functions. Fitted value iteration algorithms come with convergence guarantees but only when the function approximator employed is restricted to "averagers" [10, 9]. These guarantees hold for infinite action spaces, as well. However, to our best knowledge there are no results that would characterize the finite-sample performance of these algorithms and hence no guidance is available for how to make the best use of the available samples (for a related asymptotic result see [11]). Nevertheless, FQI has been applied successfully in a number of challenging domains and was found to perform well even when used with non-averager function approximation methods, such as neural networks [12, 13].

In this paper we study FQI in continuous state and action spaces. We propose a modification of the basic algorithm and prove bounds on its finite-sample performance that can be used to show the consistency of the modified algorithm. The modification concerns the selection of greedy actions: In the modified algorithm the exact, pointwise optimization is replaced by searching for a policy in a restricted policy class that maximizes the sum of action-values over the sampled states. Although this step is not necessarily cheaper than the original pointwise optimization, we will argue that it improves across state generalization and is essential to prevent overfitting which might happen when using the unmodified updates.

2 Preliminaries

We will build on the results from [6, 7, 14] and for this reason we use the same notation as these papers. The unattributed results cited in this section can be found in the book [15].

First, we need to fix some technical notations: For a measurable space with domain $\mathcal{X}$ we let $M(\mathcal{X})$ denote the set of all probability measures over $\mathcal{X}$. For $\nu \in M(\mathcal{X})$ and $f : \mathcal{X} \to \mathbb{R}$ measurable we let $\|f\|_{p,\nu}$ ($p \geq 1$) denote the $L^p(\nu)$-norm of f: $\|f\|_{p,\nu}^p = \int |f(s)|^p \nu(ds)$. We simply write $\|f\|_\nu$ for the L^2-norm of f. We shall use the shorthand notation νf to denote the integral $\int f(s)\nu(ds)$. We denote the space of bounded measurable functions with domain $\mathcal{X}$ by $B(\mathcal{X})$. Further, the space of measurable functions bounded by $0 < K < \infty$ shall be denoted by $B(\mathcal{X}; K)$. We let $\|f\|_\infty$ denote the supremum norm: $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. $\mathbb{I}_E$ denotes the indicator of event E, $\mathbf{1}$ denotes the function that takes on the constant value one everywhere over its domain. The Lebesgue measure shall be denoted by λ.

A discounted MDP is defined by a quintuple $(\mathcal{X}, \mathcal{A}, P, S, \gamma)$, where $\mathcal{X}$ is the (possible infinite) *state space*, $\mathcal{A}$ is the set of *actions*, $P : \mathcal{X} \times \mathcal{A} \to M(\mathcal{X})$ is the *transition probability kernel* with $P(\cdot|x, a)$ defining the next-state distribution upon taking action a from state x, $S(\cdot|x, a)$ gives the corresponding distribution of *immediate rewards*, and $\gamma \in (0, 1)$ is the discount factor. Here $M(\mathcal{X})$ denotes the space of probability measures over $\mathcal{X}$.

We start with the following mild assumption on the MDP:

Assumption A1 (MDP Regularity) $\mathcal{X}$ is a compact subset of the $d_\mathcal{X}$-dimensional Euclidean space, $\mathcal{A} \subset [-A_\infty, A_\infty]^{d_\mathcal{A}}$. We assume that the random immediate rewards are bounded by $\hat{R}_{\max}$, and that the expected immediate reward function, $r(x, a) = \int rS(dr|x, a)$, is uniformly bounded by $R_{\max}$: $\|r\|_\infty \leq R_{\max}$, where $\|\cdot\|_\infty$ denotes the supremum norm.

A *policy* determines the next action given the past observations. The action can be chosen stochastically. Formally, a policy thus maps past observations to a distribution over the set of actions.[1] A policy is deterministic if the probability distribution concentrates on a single action for all histories. A policy is called *(non-stationary) Markovian* if the distribution depends only on the last state of the observation sequence and the length of the history. A policy is called *stationary (Markovian)* if the distribution depends only on the last state of the observation sequence (and not on the length of the history).

The *value* of a policy π when it is started from a state x is defined as the total expected discounted reward that is encountered while the policy is executed: $V^\pi(x) = \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t R_t | X_0 = x\right]$. Here $R_t \sim S(\cdot|X_t, A_t)$ is the reward received at time step t, the state, X_t, evolves according to $X_{t+1} \sim P(\cdot|X_t, A_t)$, where A_t is sampled from the distribution determined by π. We use $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ to denote the *action-value function* of policy π: $Q^\pi(x,a) = \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t R_t | X_0 = x, A_0 = a\right]$.

The goal is to find a policy that attains the best possible values, $V^*(x) = \sup_\pi V^\pi(x)$, at all states $x \in \mathcal{X}$. Here V^* is called the *optimal value function* and a policy π^* that satisfies $V^{\pi^*}(x) = V^*(x)$ for all $x \in \mathcal{X}$ is called *optimal*. The *optimal action-value function* $Q^*(x,a)$ is $Q^*(x,a) = \sup_\pi Q^\pi(x,a)$. We say that a (deterministic stationary) policy π is *greedy* w.r.t. an action-value function $Q \in B(\mathcal{X} \times \mathcal{A})$, and we write $\pi = \hat{\pi}(\cdot; Q)$, if, for all $x \in \mathcal{X}$, $\pi(x) \in \operatorname{argmax}_{a \in \mathcal{A}} Q(x,a)$. Under mild technical assumptions, such a greedy policy always exists. Further, the greedy policy w.r.t. Q^* is optimal. For a deterministic stationary policy π, we define its *evaluation operator*, $T^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$, by $(T^\pi Q)(x,a) = r(x,a) + \gamma \int Q(y, \pi(y)) P(dy|x,a)$. It is known that $Q^\pi = T^\pi Q^\pi$. Further, if we let the *Bellman operator*, $T : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$, defined by $(TQ)(x,a) = r(x,a) + \gamma \int \sup_{a \in \mathcal{A}} Q(y,a) P(dy|x,a)$ then $Q^* = TQ^*$. It is known that V^π and Q^π are bounded by $R_{\max}/(1-\gamma)$, just like Q^* and V^*.

We assign to any deterministic stationary policy $\pi : \mathcal{X} \to \mathcal{A}$ the operator $E^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X}))$ defined by $(E^\pi Q)(x) = Q(x, \pi(x))$ and define $E : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X})$ by $(EQ)(x) = \sup_{a \in \mathcal{A}} Q(x,a)$. These operators are useful in connecting action-value functions with value functions: If Q^π is the action-value function of policy π, $E^\pi Q^\pi$ gives its value function, V^π. Further, $V^* = EQ^*$ and π is a greedy policy w.r.t. Q if and only if $E^\pi Q = EQ$. Moreover, T^π and T can be written as $(T^\pi Q)(x,a) = r(x,a) + \gamma \int E^\pi Q dP(\cdot|x,a)$ and $(TQ)(x,a) = r(x,a) + \gamma \int EQ dP(\cdot|x,a)$.

We shall also need two operators corresponding to the transition probability kernel P that we define now. A right-linear operator, $P\cdot : B(\mathcal{X}) \to B(\mathcal{X} \times \mathcal{A})$, is defined by $(PV)(x,a) = \int V(y) P(dy|x,a)$, i.e., PV just gives the one-step lookahead values with no discounting and no rewards. The left-linear operator, $\cdot P : M(\mathcal{X} \times \mathcal{A}) \to M(\mathcal{X})$, is defined with $(\rho P)(dy) = \int P(dy|x,a) \rho(dx, da)$. Intuitively, ρP is the distribution of states obtained after picking a state-action pair (X, A) randomly according to ρ and then executing action A in state X. This operator is also extended to act on measures over $\mathcal{X}$ with the definition $(\rho P)(dy) = \int P(dy|x,a) \rho(dx) d\lambda_{\mathcal{A}}(a)$, where $\lambda_{\mathcal{A}}$ is the uniform distribution on $\mathcal{A}$.[2] This corresponds to the distribution of states obtained by executing a random action from a state sampled from ρ. By composing P and E^π, we define $P^\pi = PE^\pi$. Note that this equation defines two operators: a right- and a left-linear one and with their help the operator T^π can be succinctly written as $T^\pi Q = r + \gamma P E^\pi Q$.

Throughout the paper $\mathcal{F} \subset \{f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$ will denote a subset of real-valued functions over the state-action space $\mathcal{X} \times \mathcal{A}$ and $\Pi \subset \mathcal{A}^{\mathcal{X}}$ will be a set of policies. For $\nu \in M(\mathcal{X})$, we extend $\|\cdot\|_{p,\nu}$ ($p \geq 1$) to $\mathcal{F}$ by $\|f\|_{p,\nu}^p = \int_{\mathcal{X} \times \mathcal{A}} |f|^p (x,a) d(\nu \times \lambda_{\mathcal{A}})(x,a)$, where $\nu \times \lambda_{\mathcal{A}}$ denotes the product measure over $\mathcal{X} \times \mathcal{A}$ obtained from ν and $\lambda_{\mathcal{A}}$.

3 Fitted Q-iteration with approximate policy maximization

We assume that we are given a finite trajectory, $\{(X_t, A_t, R_t)\}_{1 \leq t \leq N}$, generated by some stochastic stationary policy π_b, called the *behavior policy*: $A_t \sim \pi_b(\cdot|X_t)$, $X_{t+1} \sim P(\cdot|X_t, A_t)$, $R_t \sim S(\cdot|X_t, A_t)$, where $\pi_b(\cdot|x)$ is a density with $\pi_0 \stackrel{\text{def}}{=} \inf_{(x,a) \in \mathcal{X} \times \mathcal{A}} \pi_b(\cdot|x) > 0$.

[1] The policy should be a measurable function. However, in this paper, by assuming sufficient regularity, we will disregard measurability issues.

[2] For any measurable subset U of $\mathcal{A}$, $\lambda_{\mathcal{A}}(U) = \lambda(U)/\lambda(\mathcal{A})$.

The generic recipe for fitted Q-iteration (FQI) [8] is

$$Q_{k+1} = \mathrm{Regress}(D_k(Q_k)), \tag{1}$$

where $\mathrm{Regress}$ is an appropriate regression procedure and $D_k(Q_k)$ is a dataset defining a regression problem in the form of a list of data-point pairs:

$$D_k(Q_k) = \left\{ \left[(X_t, A_t), R_t + \gamma \max_{b \in \mathcal{A}} Q_k(X_{t+1}, b) \right]_{1 \le t \le N} \right\}.^3$$

Fitted Q-iteration can be viewed as approximate value iteration applied to action-value functions. To see this note that value iteration would assign the value $(TQ_k)(x,a) = r(x,a) + \gamma \int \max_{b \in \mathcal{A}} Q_k(y,b)\, P(dy|x,a)$ to $Q_{k+1}(x,a)$ [4]. Now, remember that the regression function for the jointly distributed random variables (Z, Y) is defined by the conditional expectation of Y given Z: $m(Z) = \mathbb{E}[Y|Z]$. Since for any *fixed* function Q, $\mathbb{E}[R_t + \max_{b \in \mathcal{A}} Q(X_{t+1}, b)|X_t, A_t] = (TQ)(X_t, A_t)$, thus the regression function corresponding to the dataset $D_k(Q)$ is indeed TQ and hence if FQI could solve the regression problem defined by Q_k exactly, it would simulate value iteration exactly.

However, this argument itself does not directly lead to a rigorous analysis of FQI: Since Q_k is obtained based on the data, it is itself a random function. Hence, after the first iteration, the "target" function in FQI becomes random. Furthermore, this function depends on the same data that is used to define the regression problem. Will FQI still work despite these issues? To illustrate the potential difficulties consider a dataset where $X_1, \ldots, X_N$ is a sequence of independent random variables, which are all distributed uniformly at random in $[0,1]$. Further, let M be a random integer greater than N which is independent of the dataset $(X_t)_{t=1}^N$. Let U be another random variable, uniformly distributed in $[0,1]$. Now define the regression problem by $Y_t = f_{M,U}(X_t)$, where $f_{M,U}(x) = \mathrm{sgn}(\sin(2^M 2\pi(x+U)))$. Then it is not hard to see that no matter how big N is, no procedure can estimate the regression function $f_{M,U}$ with a small error (in expectation, or with high probability), even if the procedure could exploit the knowledge of the specific form of f. On the other hand, if we restricted M to a finite range then the estimation problem could be solved successfully. The example shows that if the complexity of the random functions defining the regression problem is uncontrolled then successful estimation might be impossible.

Amongst the many regression methods in this paper we have chosen to work with least-squares methods. In this case Equation (1) takes the form

$$Q_{k+1} = \operatorname*{argmin}_{Q \in \mathcal{F}} \sum_{t=1}^{N} \frac{1}{\pi_b(A_t|X_t)} \left(Q(X_t, A_t) - \left[R_t + \gamma \max_{b \in \mathcal{A}} Q_k(X_{t+1}, b) \right] \right)^2. \tag{2}$$

We call this method the least-squares fitted Q-iteration (LSFQI) method. Here we introduced the weighting $1/\pi_b(A_t|X_t)$ since we do not want to give more weight to those actions that are preferred by the behavior policy. Other weighting factors, expressing ones' beliefs about the behavior policy is, are also possible.

Besides this weighting, the only parameter of the method is the function set $\mathcal{F}$. This function set should be chosen carefully, to keep a balance between the representation power and the number of samples. As a specific example for $\mathcal{F}$ consider neural networks with some fixed architecture. In this case the function set is generated by assigning weights in all possible ways to the neural net. Then the above minimization becomes the problem of tuning the weights. Another example is to use linearly parameterized function approximation methods with appropriately selected basis functions. In this case the weight tuning problem would be less demanding. Yet another possibility is to let $\mathcal{F}$ be an appropriate restriction of a Reproducing Kernel Hilbert Space (e.g., in a ball). In this case the training procedure becomes similar to LS-SVM training [16].

As indicated above, the analysis of this algorithm is complicated by the fact that the new dataset is defined in terms of the previous iterate, which is already a function of the dataset. Another complication is that the samples in a trajectory are in general correlated and that the bias introduced by the imperfections of the approximation architecture may yield to an explosion of the error of the procedure, as documented in a number of cases in, e.g., [9].

[3] Since the designer controls Q_k, we may assume that it is continuous, hence the maximum exists.

Nevertheless, at least for finite action sets, the tools developed in [6, 14, 7] look suitable to show that under appropriate conditions these problems can be overcome if the function set is chosen in a judicious way. However, the results of these works would become essentially useless in the case of an infinite number of actions since these previous bounds grow to infinity with the number of actions. Actually, we believe that this is not an artifact of the proof techniques of these works, as suggested by the counterexample that involved random targets. The following result elaborates this point further:

Proposition 3.1. *Let $\mathcal{F} \subset B(\mathcal{X} \times \mathcal{A})$. Then even if the pseudo-dimension of $\mathcal{F}$ is finite, the fat-shattering function of*

$$\mathcal{F}_{\max}^{\vee} = \left\{ V_Q \,:\, V_Q(\cdot) = \max_{a \in \mathcal{A}} Q(\cdot, a), Q \in \mathcal{F} \right\}$$

can be infinite over $(0, 1/2)$.

Without going into further details, let us just note that the finiteness of the fat-shattering function is a sufficient and necessary condition for learnability and the finiteness of the fat-shattering function is implied by the finiteness of the pseudo-dimension [17].The above proposition thus shows that without imposing further special conditions on $\mathcal{F}$, the learning problem may become infeasible.

One possibility is of course to discretize the action space, e.g., by using a uniform grid. However, if the action space has a really high dimensionality, this approach becomes unfeasible (even enumerating $2^{d_{\mathcal{A}}}$ points could be impossible when $d_{\mathcal{A}}$ is large). Therefore we prefer alternate solutions.

Another possibility is to make the functions in $\mathcal{F}$ e.g. uniformly Lipschitz in their state coordinates. Then the same property will hold for functions in $\mathcal{F}_{\max}^{\vee}$ and hence by a classical result we can bound the capacity of this set (cf. pp. 353–357 of [18]). One potential problem with this approach is that this way it might be difficult to get a fine control of the capacity of the resulting set.

In the approach explored here we modify the fitted Q-iteration algorithm by introducing a policy set Π and a search over this set for an approximately greedy policy in a sense that will be made precise in a minute. Our algorithm thus has four parameters: $\mathcal{F}, \Pi, K, Q_0$. Here $\mathcal{F}$ is as before, Π is a user-chosen set of policies (mappings from $\mathcal{X}$ to $\mathcal{A}$), K is the number of iterations and Q_0 is an initial value function (a typical choice is $Q_0 \equiv 0$). The algorithm computes a sequence of iterates $(Q_k, \hat{\pi}_k)$, $k = 0, \ldots, K$, defined by the following equations:

$$\begin{aligned}
\hat{\pi}_0 &= \operatorname*{argmax}_{\pi \in \Pi} \sum_{t=1}^{N} Q_0(X_t, \pi(X_t)). \\
Q_{k+1} &= \operatorname*{argmin}_{Q \in \mathcal{F}} \sum_{t=1}^{N} \frac{1}{\pi_b(A_t|X_t)} \Big(Q(X_t, A_t) - \big[R_t + \gamma Q_k(X_{t+1}, \hat{\pi}_k(X_{t+1})) \big] \Big)^2, \qquad (3) \\
\hat{\pi}_{k+1} &= \operatorname*{argmax}_{\pi \in \Pi} \sum_{t=1}^{N} Q_{k+1}(X_t, \pi(X_t)). \qquad (4)
\end{aligned}$$

Thus, (3) is similar to (2), while (4) defines the policy search problem. The policy search will generally be solved by a gradient procedure or some other appropriate method. The cost of this step will be primarily determined by how well-behaving the iterates Q_{k+1} are in their action arguments. For example, if they were quadratic and if π was linear then the problem would be a quadratic optimization problem. However, except for special cases[4] the action value functions will be more complicated, in which case this step can be expensive. Still, this cost could be similar to that of searching for the maximizing actions for each $t = 1, \ldots, N$ if the approximately maximizing actions are similar across similar states.

This algorithm will be shown to overcome the above mentioned complexity control problem provided that the complexity of Π is controlled appropriately. Indeed, in this case set of possible regression problems is determined by the set

$$\mathcal{F}_{\Pi}^{\vee} = \{ f \,:\, f(\cdot) = Q(\cdot, \pi(\cdot)), Q \in \mathcal{F}, \pi \in \Pi \},$$

and the proof will rely on controlling the complexity of $\mathcal{F}_{\Pi}^{\vee}$ by selecting $\mathcal{F}$ and Π appropriately.

[4]Linear quadratic regulation is such a nice case. It is interesting to note that in this special case the obvious choices for $\mathcal{F}$ and Π yield zero error in the limit, as can be proven based on the main result of this paper.

4 The main theoretical result

4.1 Outline of the analysis

In order to gain some insight into the behavior of the algorithm, we provide a brief summary of its error analysis. The main result will be presented subsequently.

For f,$Q \in \mathcal{F}$ and a policy $\hat{\pi}$, we define the t^{th} TD-error as follows:

$$d_t(f; Q, \hat{\pi}) = R_t + \gamma Q(X_{t+1}, \hat{\pi}(X_{t+1})) - f(X_t, A_t).$$

Further, we define the empirical loss function by

$$\hat{L}_N(f; Q, \hat{\pi}) = \frac{1}{N} \sum_{t=1}^{N} \frac{d_t^2(f; Q, \hat{\pi})}{\lambda(\mathcal{A})\pi_b(A_t|X_t)}, \tag{5}$$

where the normalization with $\lambda(\mathcal{A})$ is introduced for mathematical convenience. Then (3) can be written compactly as

$$Q_{k+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{L}_N(f; Q_k, \hat{\pi}_k).$$

The algorithm can then be motivated by the observation that for any f,Q and $\hat{\pi}$, $\hat{L}_N(f; Q, \hat{\pi})$ is an unbiased estimate of

$$L(f; Q, \hat{\pi}) \stackrel{\text{def}}{=} \left\| f - T^{\hat{\pi}} Q \right\|_\nu^2 + L^*(Q, \hat{\pi}), \tag{6}$$

where the first term is the error we are interested in and the second term captures the variance of the random samples:

$$L^*(Q, \hat{\pi}) = \int_{\mathcal{A}} \mathbb{E}\left[\mathrm{Var}\left[R_1 + \gamma Q(X_2, \hat{\pi}(X_2))|X_1, A_1 = a\right]\right] d\lambda_{\mathcal{A}}(a).$$

This result is stated formally in the following lemma:

Lemma 4.1 (Unbiased Loss Approximation)**.** *Assume that $\pi_0 > 0$. Then for any f,$Q \in \mathcal{F}$, policy $\hat{\pi}$, $\hat{L}_N(f; Q, \hat{\pi})$ as defined by (5) provides an unbiased estimate to $L(f; Q, \hat{\pi})$:*

$$\mathbb{E}\left[\hat{L}_N(f; Q, \hat{\pi})\right] = L(f; Q, \hat{\pi}). \tag{7}$$

Proof. Let us define $\hat{Q}_t = R_t + \gamma Q(X_{t+1}, \hat{\pi}(X_{t+1}))$. Then, by (5), the tth term of $\hat{L}_N(f; Q, \hat{\pi})$ can be written as

$$L^{(t)} = \frac{1}{\lambda(\mathcal{A})\pi_b(A_t|X_t)} (f(X_t, A_t) - \hat{Q}_t)^2. \tag{8}$$

Note that $\mathbb{E}\left[R_t|X_t, A_t\right] = r(X_t, A_t)$ and

$$\mathbb{E}\left[\hat{Q}_t \middle| X_t, A_t\right] = r(X_t, A_t) + \gamma \int_y Q(y, \hat{\pi}(y))\, dP(y|X_t, A_t) = (T^{\hat{\pi}} Q)(X_t, A_t). \tag{9}$$

Taking expectations,

$$\mathbb{E}\left[L^{(1)}\right] = \mathbb{E}\left[\mathbb{E}\left[L^{(1)} \middle| X_1, A_1\right]\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[(f(X_1, A_1) - \hat{Q}_1)^2|X_1, A_1\right]}{\lambda(\mathcal{A})\pi_b(A_1|X_1)}\right].$$

Now since all actions are sampled with positive probability in any state, we get

$$\begin{aligned}
&\mathbb{E}\left[(f(X_1, A_1) - \hat{Q}_1)^2|X_1, A_1\right] \\
&= \mathrm{Var}\left[\hat{Q}_1|X_1, A_1\right] + \left(f(X_1, A_1) - \mathbb{E}\left[\hat{Q}_1|X_1, A_1\right]\right)^2 \\
&= \mathrm{Var}\left[\hat{Q}_1|X_1, A_1\right] + (f(X_1, A_1) - (T^{\hat{\pi}} Q)(X_1, A_1))^2 \qquad \text{(by (9))}.
\end{aligned}$$

Taking expectations of both sides we get that

$$\begin{aligned}\mathbb{E}\left[L^{(1)}\right] &= \mathbb{E}\left[\frac{\mathrm{Var}\left[\hat{Q}_1|X_1,A_1\right]+(f(X_1,A_1)-(T^{\hat{\pi}}Q)(X_1,A_1))^2}{\lambda(\mathcal{A})\pi_b(A_1|X_1)}\right]\\ &= L^*(Q,\hat{\pi})+\left\|f-T^{\hat{\pi}}Q\right\|_\nu^2 \qquad (10)\\ &= L(f;Q,\hat{\pi}).\end{aligned}$$

Because of stationarity this holds for $\mathbb{E}\left[L^{(t)}\right]$ for any t, thus finishing the proof of (7). □

Since the variance term in (6) is independent of f, $\mathrm{argmin}_{f\in\mathcal{F}} L(f;Q,\hat{\pi}) = \mathrm{argmin}_{f\in\mathcal{F}} \left\|f-T^{\hat{\pi}}Q\right\|_\nu^2$. Thus, if $\hat{\pi}_k$ were greedy w.r.t. Q_k then $\mathrm{argmin}_{f\in\mathcal{F}} L(f;Q_k,\hat{\pi}_k) = \mathrm{argmin}_{f\in\mathcal{F}} \|f-TQ_k\|_\nu^2$. Hence we can still think of the procedure as approximate value iteration over the space of action-value functions, projecting TQ_k using empirical risk minimization on the space $\mathcal{F}$ w.r.t. $\|\cdot\|_\nu$ distances in an approximate manner. Since $\hat{\pi}_k$ is only approximately greedy, we will have to deal with both the error coming from the approximate projection and the error coming from the choice of $\hat{\pi}_k$. To make this clear, we write the iteration in the form

$$\begin{aligned}Q_{k+1} &= T^{\hat{\pi}_k}Q_k+\varepsilon'_k\\ &= TQ_k+\varepsilon'_k+(T^{\hat{\pi}_k}Q_k-TQ_k)\\ &= TQ_k+\varepsilon_k, \qquad (11)\end{aligned}$$

where ε'_k is the error committed while computing $T^{\hat{\pi}_k}Q_k$, $\varepsilon''_k \stackrel{\text{def}}{=} T^{\hat{\pi}_k}Q_k - TQ_k$ is the error committed because the greedy policy is computed approximately and $\varepsilon_k = \varepsilon'_k+\varepsilon''_k$ is the total error of step k. Hence, in order to show that the procedure is well behaved, one needs to show that both errors are controlled and that when the errors are propagated through these equations, the resulting error stays controlled, too. Since we are ultimately interested in the performance of the policy obtained, we will also need to show that small action-value approximation errors yield small performance losses. For these we need a number of assumptions that concern either the training data, the MDP, or the function sets used for learning.

4.2 Assumptions

4.2.1 Assumptions on the training data

We shall assume that the data is rich, is in a steady state and is fast-mixing, where, informally, mixing means that future depends weakly on the past. More formally, we use β-mixing, which is one of the weakest mixing concepts:

Definition 4.2 (β-mixing). *Let $\{Z_t\}_{t=1,2,\ldots}$ be a stochastic process. Denote by $Z^{1:n}$ the collection $(Z_1,\ldots,Z_n)$, where we allow $n=\infty$. Let $\sigma(Z^{i:j})$ denote the sigma-algebra generated by $Z^{i:j}$ ($i\le j$). The m-th β-mixing coefficient of $\{Z_t\}$, β_m, is defined by*

$$\beta_m = \sup_{t\ge 1}\mathbb{E}\left[\sup_{B\in\sigma(Z^{t+m:\infty})}|P(B|Z^{1:t})-P(B)|\right].$$

$\{Z_t\}$ is said to be β-mixing if $\beta_m\to 0$ as $m\to\infty$. In particular, we say that a β-mixing process mixes at an exponential *rate with parameters $\overline{\beta}$,b,$\kappa>0$ if $\beta_m\le\overline{\beta}\exp(-bm^\kappa)$ holds for all $m\ge 0$.*

Now we are ready to state our assumptions on the data (cf. [7]):

Assumption A2 (Sample Path Properties) Assume that

$$\{(X_t,A_t,R_t)\}_{t=1,\ldots,N}$$

is the sample path of π_b, a stochastic stationary policy. Further, assume that $\{X_t\}$ is strictly stationary ($X_t\sim\nu\in M(\mathcal{X})$) and exponentially β-mixing with the actual rate given by the parameters $(\overline{\beta},b,\kappa)$.We further assume that the sampling policy π_b satisfies $\pi_0\stackrel{\text{def}}{=}\inf_{(x,a)\in\mathcal{X}\times\mathcal{A}}\pi_b(a|x)>0$.

The β-mixing property will be used to establish tail inequalities for certain empirical processes. Note that the mixing coefficients do not need to be known. In the case when no mixing condition is satisfied, learning might be impossible. To see this just consider the case when $X_1 = X_2 = \ldots = X_N$. Thus, in this case the learner has many copies of the same random variable and successful generalization is thus impossible. We believe that the assumption that the process is in a steady state is not essential for our result, as when the process reaches its steady state quickly then (at the price of a more involved proof) the result would still hold.

4.2.2 Assumptions on the MDP

In order to prevent the uncontrolled growth of the errors as they are propagated through the updates, we shall need some assumptions on the MDP. A convenient assumption is the following one [19]:

Assumption A3 (Uniformly stochastic transitions) For all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, assume that $P(\cdot|x,a)$ is absolutely continuous w.r.t. ν and the Radon-Nikodym derivative of P w.r.t. ν is bounded uniformly with bound C_ν:

$$C_\nu \stackrel{\text{def}}{=} \sup_{x\in\mathcal{X},a\in\mathcal{A}} \left\| \frac{dP(\cdot|x,a)}{d\nu} \right\|_\infty < +\infty.$$

Note that by the definition of measure differentiation, Assumption A3 means that $P(\cdot|x,a) \le C_\nu \nu(\cdot)$, where the inequality holds for every measurable set. This assumption essentially requires the transitions to be noisy. We will also prove (weaker) results under the following, *weaker* assumption:

Assumption A4 (Discounted-average concentrability of future-state distributions) Given ρ, ν, $m \ge 1$ and an arbitrary sequence of stationary policies $\{\pi_m\}_{m\ge 1}$, assume that the future-state distribution $\rho P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m}$ is absolutely continuous w.r.t. ν. Assume that

$$c(m) \stackrel{\text{def}}{=} \sup_{\pi_1,\ldots,\pi_m} \left\| \frac{d(\rho P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m})}{d\nu} \right\|_\infty \tag{12}$$

satisfies

$$C_{\rho,\nu} \stackrel{\text{def}}{=} (1-\gamma)^2 \sum_{m\ge 1} m\gamma^{m-1} c(m) < +\infty.$$

We shall call $c(m)$ the m-step concentrability of a future-state distribution, while we call $C_{\rho,\nu}$ the *discounted-average concentrability coefficient* of the future-state distributions.

The number $c(m)$ measures how much ρ can get amplified in m steps as compared to the reference distribution ν. Hence, in general we expect $c(m)$ to grow with m. In fact, the condition that $C_{\rho,\mu}$ is finite is a growth rate condition on $c(m)$. Thanks to discounting, $C_{\rho,\mu}$ is finite for a reasonably large class of systems (see the discussion in [19]).

A related assumption is needed in the error analysis of the approximate greedy step of the algorithm:

Assumption A5 (The random policy "goes everywhere") Consider the distribution $\mu = (\nu \times \lambda_{\mathcal{A}})P$ which is the distribution of a state that results from sampling an initial state according to ν and then executing an action which is selected uniformly at random.[5] Then

$$\Gamma_\nu = \left\| \frac{d\mu}{d\nu} \right\|_\infty < +\infty.$$

Note that under Assumption A3 we have $\Gamma_\nu \le C_\nu$. This (very mild) assumption means that after one step, starting from ν the random policy cannot avoid some part of the state space.

Besides, we assume that $\mathcal{A}$ has the following regularity property:

[5] Remember that $\lambda_{\mathcal{A}}$ denotes the uniform distribution over the action set $\mathcal{A}$.

Assumption A6 (Regularity of the action space) Let λ denote the Lebesgue-measure. For any $a \in \mathcal{A}$, write $B(a,\rho) = \left\{a' \in \mathbb{R}^{d_{\mathcal{A}}}\right\} \|a - a'\|_1 \leq \rho$ for the ball centered at a. We assume that there exists $\alpha > 0$, such that for all $a \in \mathcal{A}$, for all $\rho > 0$,

$$\lambda(B(a,\rho) \cap \mathcal{A}) \geq \min\left(\alpha\lambda(B(a,\rho)), \lambda(\mathcal{A})\right).$$

For example, if $\mathcal{A}$ is an L^1-ball itself, then this assumption will be satisfied with $\alpha = 2^{-d_{\mathcal{A}}}$. In general, this assumption is satisfied whenever $\mathcal{A}$ has a non-degenerated boundary.

Without assuming any smoothness of the MDP, learning in *infinite* MDPs looks hard (see, e.g., [1, 20]). Here we employ the following extra condition:

Assumption A7 (Lipschitzness of the MDP in the actions) Assume that the transition probabilities and rewards are Lipschitz w.r.t. their action variable, i.e., there exists $L_P, L_r > 0$ such that for all $(x, a, a') \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}$ and measurable set B of $\mathcal{X}$,

$$\begin{aligned} |P(B|x,a) - P(B|x,a')| &\leq L_P \|a - a'\|_1, \\ |r(x,a) - r(x,a')| &\leq L_r \|a - a'\|_1. \end{aligned}$$

Note that previously Lipschitzness w.r.t. the *state* variables was used e.g. in [19] to construct consistent planning algorithms.

4.2.3 Assumptions on the function sets used by the algorithm

These assumptions are less demanding since they are under the control of the user of the algorithm. However, the choice of these function sets will greatly influence the performance of the algorithm, as we shall see it from the bounds. The first assumption concerns the class $\mathcal{F}$:

Assumption A8 (Lipschitzness of candidate action-value functions) Assume $\mathcal{F} \subset B(\mathcal{X} \times \mathcal{A})$ and that any elements of $\mathcal{F}$ is uniformly Lipschitz in its action-argument in the sense that

$$|Q(x,a) - Q(x,a')| \leq L_{\mathcal{A}} \|a - a'\|_1$$

holds for any $x \in \mathcal{X}$, $a, a' \in \mathcal{A}$ and $Q \in \mathcal{F}$.

We shall also need to control the capacity of our function sets. We assume that the reader is familiar with the concept of VC-dimension.[6] Here we use the *pseudo-dimension* of function sets that builds upon the concept of VC-dimension:

Definition 4.3 (Pseudo-dimension)**.** *The* pseudo-dimension $V_{\mathcal{F}^+}$ *of* $\mathcal{F}$ *is defined as the* VC*-dimension of the subgraphs of functions in* $\mathcal{F}$ *(hence it is also called the* VC-subgraph dimension *of* $\mathcal{F}$*).*

Since $\mathcal{A}$ is multidimensional, we define V_{Π^+} to be the sum of the pseudo-dimensions of the coordinate projection spaces, Π_k of Π:

$$V_{\Pi^+} = \sum_{k=1}^{d_{\mathcal{A}}} V_{\Pi_k^+}, \quad \Pi_k = \left\{ \pi_k : \mathcal{X} \to \mathbb{R} \,:\, \pi = (\pi_1, \ldots, \pi_k, \ldots, \pi_{d_{\mathcal{A}}}) \in \Pi \right\}.$$

See Lemma E.4 for the motivation of this definition.

Now we are ready to state our assumptions on our function sets:

Assumption A9 (Capacity of the function and policy sets) Assume that $\mathcal{F} \subset B(\mathcal{X} \times \mathcal{A}; Q_{\max})$ for $Q_{\max} > 0$ and $V_{\mathcal{F}^+} < +\infty$. Also, $V_{\Pi^+} < +\infty$.

[6] Readers not familiar with VC-dimension are suggested to consult a book, such as the one by Anthony and Bartlett [21].

Besides their capacity, one shall also control the approximation power of the function sets involved. Let us first consider the policy set Π. Introduce

$$e^*(\mathcal{F},\Pi) = \sup_{Q\in\mathcal{F}} \inf_{\pi\in\Pi} \nu(EQ - E^\pi Q).$$

Note that $\inf_{\pi\in\Pi} \nu(EQ - E^\pi Q)$ measures the quality of approximating νEQ by $\nu E^\pi Q$. Hence, $e^*(\mathcal{F},\Pi)$ measures the worst-case approximation error of νEQ as Q is changed within $\mathcal{F}$. This can be made small by choosing Π large.

Another related quantity is the *one-step Bellman-error of $\mathcal{F}$ w.r.t. Π*. This is defined as follows: For a fixed policy $\hat{\pi}$, the one-step Bellman-error of $\mathcal{F}$ w.r.t. $T^{\hat{\pi}}$ is defined as

$$E_1(\mathcal{F};\hat{\pi}) = \sup_{Q\in\mathcal{F}} \inf_{Q'\in\mathcal{F}} \left\|Q' - T^{\hat{\pi}}Q\right\|_\nu .$$

Taking again a pessimistic approach, the one-step Bellman-error of $\mathcal{F}$ is defined as

$$E_1(\mathcal{F},\Pi) = \sup_{\hat{\pi}\in\Pi} E_1(\mathcal{F};\hat{\pi}).$$

Typically by increasing $\mathcal{F}$, $E_1(\mathcal{F},\Pi)$ can be made smaller (this is discussed at some length in [14]). However, it also holds for both Π and $\mathcal{F}$ that making them bigger will increase their capacity (pseudo-dimensions) which leads to an increase of the estimation errors. Hence, $\mathcal{F}$ and Π must be selected to balance the approximation and estimation errors, just like in supervised learning.

For $p \geq 1$, let $\|V\|_{p,\rho}^p = \int_{\mathcal{X}} |f(x)|^p d\rho(x)$. Our main result is the following theorem(the theorem is restated as Theorem B.1 in the Appendix with more details) :

Theorem 4.4. *Under Assumptions A1, A2, and A5–A9, for all $\delta > 0$ we have with probability at least $1-\delta$: given Assumption A3 (respectively A4), $\|V^* - V^{\pi_K}\|_\infty$ (resp. $\|V^* - V^{\pi_K}\|_{1,\rho}$), is bounded by*

$$C\left\{\left(E_1(\mathcal{F},\Pi) + e^*(\mathcal{F},\Pi) + \frac{(\log N + \log(K/\delta))^{\frac{\kappa+1}{4\kappa}}}{N^{1/4}}\right)^{\frac{1}{d_{\mathcal{A}}+1}} + \gamma^K\right\}.$$

where C depends on $d_{\mathcal{A}}$, $V_{\mathcal{F}^+}$, $(V_{\Pi_k^+})_{k=1}^{d_{\mathcal{A}}}$, γ, κ, b, $\overline{\beta}$, C_ν (resp. $C_{\rho,\nu}$), Γ_ν, $L_{\mathcal{A}}$, L_P,L_r, α, $\lambda(\mathcal{A})$, π_0, $Q_{\max}$, $R_{\max}$, $\hat{R}_{\max}$, and A_∞. In particular, C scales with $V^{\frac{\kappa+1}{4\kappa(d_{\mathcal{A}}+1)}}$, where $V = 2V_{\mathcal{F}^+} + V_{\Pi^+}$ plays the role of the "combined effective" dimension of $\mathcal{F}$ and Π.

5 Discussion

We have presented what we believe is the first finite-time bounds for continuous-state and action-space RL that uses value functions. Further, this is the first analysis of fitted Q-iteration, an algorithm that has proved to be useful in a number of cases, even when used with non-averagers for which no previous theoretical analysis existed (e.g., [12, 13]). In fact, our main motivation was to show that there is a systematic way of making these algorithms work and to point at possible problem sources the same time. We discussed why it can be difficult to make these algorithms work in practice. We suggested that either the set of action-value candidates has to be carefully controlled (e.g., assuming uniform Lipschitzness w.r.t. the state variables), or a policy search step is needed, just like in actor-critic algorithms. The bound in this paper is similar in many respects to a previous bound of a Bellman-residual minimization algorithm [7]. It looks that the techniques developed here can be used to obtain results for that algorithm when it is applied to continuous action spaces. Finally, although we have not explored them here, consistency results for FQI can be obtained from our results using standard methods, like the methods of sieves. We believe that the methods developed here will eventually lead to algorithms where the function approximation methods are chosen based on the data (similar to adaptive regression methods) so as to optimize performance, which in our opinion is one of the biggest open questions in RL. Currently we are exploring the possibility of this.

A Proof of Proposition 3.1

Proof. We give such an $\mathcal{F}$ the following way: Choose $\mathcal{X} = \mathcal{A} = \{1, 2, \ldots\}$. Enumerate all the finite subsets of $\mathcal{X}$ as S_1,S_2,.... Let $Q_i(x, a) = \mathbb{I}_{\{x \in S_i, a=i\}}$ and $\mathcal{F} = \{Q_i\}\, i = 1, 2, \ldots$. Then

$$V_{Q_i}(x) = \max_{a \in \mathcal{A}} Q_i(x, a) = \mathbb{I}_{\{x \in S_i\}},$$

hence the subgraph system of $\mathcal{F}^{\vee}_{\max}$ shatters arbitrary large set of finite number of points with any positive fat-shattering less than $1/2$. Thus the fat-shattering function of $\mathcal{F}^{\vee}_{\max}$ is not finite over $(0, 1)$. On the other hand, it is easy to see that the subgraph system of $\mathcal{F}$ does not shatter even two points, hence $V_{\mathcal{F}^+} = 1$. □

B The main theorem

Here we present the main theorem with the precise constants.

Theorem B.1. *Under Assumptions A1, A2, A5, A6, A7, A8, and A9, for all $\delta > 0$ we have with probability at least $1 - \delta$:*

- *Given Assumption A3:* $\|V^* - V^{\pi_K}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2}\left\{C_\nu\, w + \frac{2R_{\max}}{1-\gamma}\gamma^K\right\}$.
- *Given Assumption A4:* $\|V^* - V^{\pi_K}\|_{1,\rho} \le \frac{2\gamma}{(1-\gamma)^2}\left\{C_{\rho,\nu}\, w + \frac{2R_{\max}}{1-\gamma}\gamma^K\right\}$.

where

$$w = \max\left(\left[\frac{\lambda(\mathcal{A})(d_{\mathcal{A}}+1)!}{\alpha(2/(L_{\mathcal{A}} + L_r + \gamma Q_{\max} L_P))^{d_{\mathcal{A}}}}\varepsilon\right]^{1/(d_{\mathcal{A}}+1)}, (d_{\mathcal{A}}+1)\varepsilon\right),$$

with $\varepsilon = \varepsilon' + \varepsilon''$,

$$\begin{aligned}(\varepsilon')^2 &= E_1^2(\mathcal{F}, \Pi) + \sqrt{\frac{\Lambda_N(\delta/(2K))\,[\Lambda_N(\delta/(2K))/b \vee 1]^{1/\kappa}}{C_2 N}},\\ \varepsilon'' &= \gamma\Gamma_\nu\left(e^*(\mathcal{F}, \Pi) + 2\sqrt{\frac{\Lambda'_N(\delta/(2K))\,[\Lambda'_N(\delta/(2K))/b \vee 1]^{1/\kappa}}{C'_2 N}}\right).\end{aligned}$$

(By definition, $a \vee b = \max(a, b)$). Here $\Lambda_N(\delta)$ and $\Lambda'_N(\delta)$ quantify the dependence of the estimation error on N, δ, and the capacities of the sets $\mathcal{F}$ and Π:

$$\Lambda_N(\delta) = \tfrac{V}{2}\log N + \log(e/\delta) + \log^+(C_1 C_2^{V/2} \vee \overline{\beta}),$$
$$\Lambda'_N(\delta) = \tfrac{V'}{2}\log N + \log(e/\delta) + \log^+(C'_1 {C'_2}^{V'/2} \vee \overline{\beta}),$$

V,V' playing the role of the "combined effective" dimensions of $\mathcal{F}$ and Π:

$$V = 2V_{\mathcal{F}^+} + V_{\Pi^+}, \qquad V' = V_{\mathcal{F}^+} + V_{\Pi^+},$$

$$\begin{aligned}\log C_1 &= V\log\left(\frac{128 e \tilde{R}_{\max}(1 + \gamma(L_{\mathcal{A}}+1))}{\lambda(\mathcal{A})\pi_0}\right) + 2V_{\mathcal{F}^+}\log Q_{\max} + V_{\Pi^+}\log(d_{\mathcal{A}}A_\infty)\\ &\quad + \sum_{k=1}^{d_{\mathcal{A}}}\log(e(V_{\Pi_k^+}+1)) + 2\log(V_{\mathcal{F}^+}+1) + 2\log(4e),\\ \log C'_1 &= V'\log(32e(L_{\mathcal{A}}+1)) + V_{\mathcal{F}^+}\log Q_{\max} + V_{\Pi^+}\log(d_{\mathcal{A}}A_\infty)\\ &\quad + \sum_{k=1}^{d_{\mathcal{A}}}\log(e(V_{\Pi_k^+}+1)) + \log(V_{\mathcal{F}^+}+1) + \log(16e),\end{aligned}$$

$$C_2 = \frac{1}{2}\left(\frac{\lambda(\mathcal{A})\pi_0}{32\tilde{R}^2_{\max}}\right)^2, \qquad C'_2 = \frac{1}{2}(16Q_{\max})^{-2},$$

and

$$\tilde{R}_{\max} = (1+\gamma)Q_{\max} + \hat{R}_{\max}.$$

C Some definitions

To avoid any confusions we introduce the definition of covering numbers:

Definition C.1 (Covering Numbers)**.** *Fix $\varepsilon > 0$ and a semi-metric space $\mathcal{M} = (\mathcal{M}, d)$. We say that $\mathcal{M}$ is covered by m discs $D_1, \ldots, D_m$ if $\mathcal{M} \subset \cup_j D_j$. We define the* covering number $\mathcal{N}(\varepsilon, \mathcal{M}, d)$ *of $\mathcal{M}$ as the smallest integer m such that $\mathcal{M}$ can be covered by m discs each of which having a radius less than ε. If no such finite m exists then we let $\mathcal{N}(\varepsilon, \mathcal{M}, d) = \infty$.*

In particular, for a class $\mathcal{F}$ of real-valued functions with domain $\mathcal{X}$ and points $x^{1:N} \stackrel{\text{def}}{=} (x_1, x_2, \ldots, x_N)$ in $\mathcal{X}$, we use the *empirical covering numbers*, i.e., the covering number of $\mathcal{F}$ equipped with the empirical L^1 semi-metric

$$l_{x^{1:N}}(f, g) = \frac{1}{N} \sum_{t=1}^{N} d(f(x_t), g(x_t)),$$

where d is a distance function on the range of functions in $\mathcal{F}$. When this range is the reals then we use $d(a, b) = |a - b|$. If we define $\mathcal{F}(x^{1:N})$ as $\left\{ f(x^{1:N}) : f \in \mathcal{F} \right\}$ then we see that the empirical covering number of $\mathcal{F}$ can be equivalently defined as the covering number of $\mathcal{F}(x^{1:N})$ when this latter set is equipped with the ℓ^1 distance normalized by N. (Here and in what follows $f(x^{1:N}) \stackrel{\text{def}}{=} (f(x_1), \ldots, f(x_N))$.) For brevity we shall denote $\mathcal{N}(\varepsilon, \mathcal{F}, l_{x^{1:N}})$ by $\mathcal{N}_1(\varepsilon, \mathcal{F}, x^{1:N})$.

The concept of pseudo-dimensions has been introduced earlier. The "scale-sensitive" counterpart of pseudo-dimension is the fat-shattering function, which is defined as follows: Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to (say) $[0, 1]$. Let $\gamma > 0$. We say that $x^{1:N} \in \mathcal{X}^N$ is γ-shattered if there is $r \in [0, 1]^N$ such that for any binary sequence b of length N there is a function $f \in \mathcal{F}$ such that $f(x_i) \geq r_i + \gamma$ if $b_i = 1$ and $f(x_i) \leq r_i - \gamma$ if $b_0 = 0$. Thus, $x^{1:N}$ is γ-shattered by the function set $\mathcal{F}$ if it is shattered with a "width of shattering" of at least γ. The fat-shattering function, $\mathrm{fat}_{\mathcal{F}}$, takes a shattering width γ and returns the largest integer N such that some $x^{1:N} \in \mathcal{X}^N$ is γ-shattered with the understanding that if no such upper bound exists then the function returns infinite. We shall say that $\mathcal{F}$ has finite fat-shattering function whenever it is the case that for all $\gamma \in (0, 1)$, $\mathrm{fat}_{\mathcal{F}}(\gamma) < +\infty$.

D Error propagation during the updates

D.1 Error propagation for value functions

The update rule $Q_{k+1} = TQ_k + \varepsilon_k$, (for $\varepsilon_k \in B(\mathcal{X} \times \mathcal{A})$) when T is expanded take the form

$$Q_{k+1}(x, a) = r(x, a) + \gamma \int P(dy|x, a) \max_{b \in \mathcal{A}} Q_k(y, b) + \varepsilon_k(x, a).$$

Defining $V_k(x) = \max_{a \in \mathcal{A}} Q_k(x, a) \in B(\mathcal{X})$, we have

$$\begin{aligned}
V_{k+1}(x) &= \max_{a \in \mathcal{A}} \left[r(x, a) + \gamma \int P(dy|x, a) V_k(y) + \varepsilon_k(x, a) \right] \\
&= r(x, \pi_{k+1}(x)) + \gamma \int P(dy|x, \pi_{k+1}(x)) V_k(y) + \varepsilon_k(x, \pi_{k+1}(x)) \\
&\leq \sup_{a \in \mathcal{A}} \left[r(x, a) + \gamma \int P(dy|x, a) V_k(y) \right] + \varepsilon_k(x, \pi_{k+1}(x)) \\
&= TV_k(x) + \varepsilon_k(x, \pi_{k+1}(x)),
\end{aligned}$$

where we wrote $\pi_{k+1}(x) = \mathrm{argmax}_{a \in \mathcal{A}} Q_{k+1}(x, a)$ for the greedy policy w.r.t. Q_{k+1}, and T for the Bellman operator applied to functions in $B(\mathcal{X})$.

Now, writing $\bar{\pi}_k(x)$ the greedy policy w.r.t. V_k, i.e.,

$$\bar{\pi}_k(x) = \operatorname*{argmax}_{a \in \mathcal{A}} \left[r(x, a) + \gamma \int P(dy|x, a) V_k(y) \right],$$

we have:

$$\begin{aligned} V_{k+1}(x) &\geq r(x,\bar{\pi}_k(x)) + \gamma \int P(dy|x,\bar{\pi}_k(x))V_k(y) + \varepsilon_k(x,\bar{\pi}_k(x)) \\ &= TV_k(x) + \varepsilon_k(x,\bar{\pi}_k(x)). \end{aligned}$$

Thus we have:

$$|V_{k+1}(x) - TV_k(x)| \leq \bar{\varepsilon}_k(x),$$

where $\bar{\varepsilon}_k(x) = \max\{|\varepsilon_k(x,\bar{\pi}_k(x))|, |\varepsilon_k(x,\pi_{k+1}(x))|\}$.

We now would like to apply Lemma 4 of [19] to obtain a bound on a L^p norm of $V^* - V^{\pi_K}$ in terms of the $\|\bar{\varepsilon}_k\|_\nu$, which in turn, may be bounded by $\|\varepsilon_k\|_\nu$.

D.2 Bound on $\|\bar{\varepsilon}_k\|_{p,\nu}$

As a first step, now we bound the L^p norm (with weight $\nu \in M(\mathcal{X})$) of $\bar{\varepsilon}_k$ in terms of the L^p norm (with weight $(\nu \times \lambda_\mathcal{A}) \in M(\mathcal{X} \times \mathcal{A})$) of $\varepsilon_k = Q_{k+1} - TQ_k$. For that we use the Lipschitz property of ε_k w.r.t. the action variable.

Lemma D.1. *Under Assumptions A7 and A8, for all $k \geq 0$, ε_k is L-Lipschitz w.r.t. its action variable, with $L = L_\mathcal{A} + L_r + \gamma Q_{\max} L_P$.*

Proof. From Assumption A8, we know that $Q_{k+1} \in \mathcal{F}$ is $L_\mathcal{A}$-Lipschitz. Now, from A7, for any function $Q \in B(\mathcal{X} \times \mathcal{A}, Q_{\max})$, TQ is $(L_r + \gamma Q_{\max} L_P)$-Lipschitz since:

$$\begin{aligned} |TQ(x,a) - TQ(x,a')| &= r(x,a) - r(x,a') + \gamma \int [P(dy|x,a) - P(dy|x,a')] \max_{b \in \mathcal{A}} Q(y,b) \\ &\leq (L_r + \gamma Q_{\max} L_P) \|a - a'\|_1 , \end{aligned}$$

thus $\varepsilon_k = Q_{k+1} - TQ_k$ is $(L_\mathcal{A} + L_r + \gamma Q_{\max} L_P)$-Lipschitz. □

Lemma D.2. *Under Assumptions A6, A7, and A8, for all $p \geq 1$, we have*

$$\|\bar{\varepsilon}_k\|_{p,\nu} \leq \max\left(\left[\frac{\lambda(\mathcal{A})(d_\mathcal{A}+1)!}{\alpha(2/L)^{d_\mathcal{A}}} \|\varepsilon_k\|_{p,\nu} \right]^{1/(d_\mathcal{A}+1)}, (d_\mathcal{A}+1)\|\varepsilon_k\|_{p,\nu} \right),$$

where $L = L_\mathcal{A} + L_r + \gamma Q_{\max} L_P$.

Notice that the left-hand side of the bound makes use of a L^p norm (weighted by $\nu \in M(\mathcal{X})$) for functions defined on $\mathcal{X}$ whereas the right-hand side uses L^p norm (weighted by $\nu \times \lambda_\mathcal{A}$) for functions defined on $\mathcal{X} \times \mathcal{A}$.

Proof. Let f be an L-Lipschitz function defined over $\mathcal{A}$: $|f(a) - f(a')| \leq L\|a - a'\|_1$ $(a, a' \in \mathcal{A})$. We first want to bound its L^∞ norm, $\|f\|_\infty = \sup_{a \in \mathcal{A}} |f(a)|$, in terms of its L^1-norm, $\|f\|_{\lambda_\mathcal{A},1} = \frac{1}{\lambda(\mathcal{A})} \int_{a \in \mathcal{A}} |f(a)| da$. We may assume without the loss of generality that $f \geq 0$. Now, given a function f and $a \in \mathcal{A}$ such that $f(a) > 0$, from the Lipschitz property of f, it follows that the function cannot go below the surface of the pyramid $\text{Py}(a, f(a), \rho)$ centered at a with height $f(a)$ and with basis $B(a,\rho)$ for $\rho = f(a)/L$. Thus, we have that for all $a \in \mathcal{A}$,

$$\|f\|_{\lambda_\mathcal{A},1} = \frac{1}{\lambda(\mathcal{A})} \int_{a' \in \mathcal{A}} |f(a')| da' \geq \frac{\lambda(\text{Py}(a, f(a), f(a)/L) \cap (\mathcal{A} \times \mathbb{R}))}{\lambda(\mathcal{A})},$$

But, since the volume of any pyramid with base B and height h is $\lambda(B)\int_0^h (1 - \frac{x}{h})^{d_\mathcal{A}} dx = \lambda(B)h/(d_\mathcal{A}+1)$, and using Assumption A6, we deduce that:

$$\begin{aligned} \|f\|_{\lambda_\mathcal{A},1} &\geq \frac{1}{\lambda(\mathcal{A})} \lambda(B(a, f(a)/L) \cap \mathcal{A}) \frac{f(a)}{d_\mathcal{A}+1} \\ &\geq \min\left(\frac{\alpha}{\lambda(\mathcal{A})} \lambda(B(a, f(a)/L)), 1 \right) \frac{f(a)}{d_\mathcal{A}+1} \\ &= \min\left(\frac{\alpha}{\lambda(\mathcal{A})} \frac{(2/L)^{d_\mathcal{A}}}{(d_\mathcal{A}+1)!} f(a)^{d_\mathcal{A}+1}, \frac{f(a)}{d_\mathcal{A}+1} \right), \end{aligned}$$

where we used the fact that $\lambda(B(a,\rho)) \geq \lambda(\{a' \in \mathcal{A}\} \|a - a'\|_1 \leq \rho) = (2\rho)^{d_{\mathcal{A}}}/(d_{\mathcal{A}}!)$. Since this holds for all $a \in \mathcal{A}$, we have

$$\|f\|_\infty \leq \max\left(\left[\frac{\lambda(\mathcal{A})(d_{\mathcal{A}}+1)!}{\alpha(2/L)^{d_{\mathcal{A}}}}\|f\|_{\lambda_{\mathcal{A}},1}\right]^{1/(d_{\mathcal{A}}+1)}, (d_{\mathcal{A}}+1)\|f\|_{\lambda_{\mathcal{A}},1}\right).$$

Now, given that $\lambda_{\mathcal{A}}$ is a uniform distribution over $\mathcal{A}$, we apply this property to $\varepsilon_k = Q_{k+1} - TQ_k$, which is an L-Lipschitz function w.r.t. its action variable, thanks to Lemma D.1, giving the desired result with $p = 1$ (i.e., for the L^1-norm). This is extended to L^p-norms since $\|f\|_{\lambda_{\mathcal{A}},1} \leq \|f\|_{\lambda_{\mathcal{A}},p}$. □

D.3 Error propagation for action value functions

We now deduce the following L^p performance bound on $V^* - V^{\pi_K}$ in terms of the L^p-norm of ε_k, with weight $\nu \times \lambda_{\mathcal{A}}$.

Lemma D.3. *Under Assumptions A6, A7, and A8 the followings hold:*

- *Given Assumption A3 we have*

$$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\left\{C_\nu^{1/p}\max_{0\leq k<K} w_k + \frac{2R_{\max}}{1-\gamma}\gamma^{K/p}\right\}. \tag{13}$$

- *Given Assumption A4 we have*

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2}\left\{C_{\rho,\nu}^{1/p}\max_{0\leq k<K} w_k + \frac{2R_{\max}}{1-\gamma}\gamma^{K/p}\right\}. \tag{14}$$

Here

$$w_k = \max\left(\left[\frac{\lambda(\mathcal{A})(d_{\mathcal{A}}+1)!}{\alpha(2/L)^{d_{\mathcal{A}}}}\|\varepsilon_k\|_{p,\nu}\right]^{1/(d_{\mathcal{A}}+1)}, (d_{\mathcal{A}}+1)\|\varepsilon_k\|_{p,\nu}\right)$$

and $L = L_{\mathcal{A}} + L_r + \gamma Q_{\max} L_P$.

Proof. This directly follows from Lemma 4 of [19] and Lemma D.2. □

E Controlling the error of the individual updates

E.1 Some definitions and technical results

Since we are dealing with β-mixing process, we need an extension of Pollard's tail inequality for this case.

Lemma E.1 ([7], Lemma 4, see also [22])**.** *Suppose that $Z_1, \ldots, Z_N \in \mathcal{Z}$ is a stationary β-mixing process with mixing coefficients $\{\beta_m\}$, $Z'_t \in \mathcal{Z}$ $(t \in H)$ are the block-independent "ghost" samples as done by [23], $H = \{2ik_N + j\}\, 0 \leq i < m_N, 1 \leq j \leq k_N$ and $\mathcal{F}$ is a permissible class of $\mathcal{Z} \to [-K, K]$ functions. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^N f(Z_t) - \mathbb{E}\left[f(Z_1)\right]\right| > \varepsilon\right) \leq 16\mathbb{E}\left[\mathcal{N}_1(\tfrac{\varepsilon}{8}, \mathcal{F}, (Z'_t)_{t\in H})\right] e^{-\frac{m_N\varepsilon^2}{128K^2}} + 2m_N\beta_{k_N+1}.$$

We now state some results that will be used to build estimates of the covering numbers of $\mathcal{F}_\Pi^\vee$. We start by the following observation:

Lemma E.2. *Let Assumption A8 hold for the function space $\mathcal{F}$. Fix $x^{1:N} \in \mathcal{X}^N$. Let $K = \mathcal{N}_1\left(\frac{\alpha\varepsilon}{L_{\mathcal{A}}}, \Pi, x^{1:N}\right)$ and let $(\pi_k)_{k=1,\ldots,K}$ be the corresponding cover. Then for any $\alpha > 0$*

$$\begin{aligned}\mathcal{N}_1(\varepsilon, \mathcal{F}_\Pi^\vee, x^{1:N}) &\leq \sum_{k=1}^K \mathcal{N}_1\left((1-\alpha)\varepsilon, \mathcal{F}, \langle \pi_k(x^{1:N})\rangle\right) \\ &\leq \mathcal{N}_1\left(\frac{\alpha\varepsilon}{L_{\mathcal{A}}}, \Pi, x^{1:N}\right) \sup_{\pi\in\Pi}\mathcal{N}_1\left((1-\alpha)\varepsilon, \mathcal{F}, \langle \pi(x^{1:N})\rangle\right),\end{aligned}$$

where $\langle \pi_k(x^{1:N}) \rangle \stackrel{\text{def}}{=} ((x_1, \pi_k(x_1)), \ldots, (x_N, \pi_k(x_N)))$.

Proof. Let $(\pi_k)_{k=1,\ldots,K}$ be the $\frac{\alpha\varepsilon}{L_{\mathcal{A}}}$-covering of $\Pi(x^{1:N})$ and let $(Q_{kj})_{j=1,\ldots,J(k)}$ be the $(1-\alpha)\varepsilon$-covering of $\mathcal{F}(\langle \pi_k(x^{1:N}) \rangle)$, $k = 1, \ldots, K$. It suffices to show that $(Q_{kj}(\cdot, \pi_k(\cdot)))_{k=1,\ldots,K, j=1,\ldots,J(k)}$ is an ε-covering of $\mathcal{F}_\Pi^\vee(x^{1:N})$.

Pick any pair $(Q, \pi) \in \mathcal{F} \times \Pi$. Let k be such that $l_{x^{1:N}}(\pi, \pi_k) \leq \frac{\alpha\varepsilon}{L_{\mathcal{A}}}$. Further, let j be such that $l_{\langle \pi_k(x^{1:N}) \rangle}(Q, Q_{kj}) \leq (1-\alpha)\varepsilon$. Then

$$\begin{aligned}
&\frac{1}{N}\sum_{t=1}^{N} |Q(x_t, \pi(x_t)) - Q_{kj}(x_t, \pi_k(x_t))| \\
&\quad\leq \frac{1}{N}\sum_{t=1}^{N} (|Q(x_t, \pi(x_t)) - Q(x_t, \pi_k(x_t))| + |Q(x_t, \pi_k(x_t)) - Q_{kj}(x_t, \pi_k(x_t))|) \\
&\quad\leq \frac{L_{\mathcal{A}}}{N}\sum_{t=1}^{N} \|\pi(x_t) - \pi'(x_t)\|_1 + \frac{1}{N}\sum_{t=1}^{N} |Q(x_t, \pi_k(x_t)) - Q_{kj}(x_t, \pi_k(x_t))| \leq \varepsilon,
\end{aligned}$$

proving that $(Q_{kj}(\cdot, \pi_k(\cdot)))_{k=1,\ldots,K, j=1,\ldots,J(k)}$ is an ε-covering of $\mathcal{F}_\Pi^\vee(x^{1:N})$. □

We use the following proposition to further bound these covering numbers:

Proposition E.3 ([24], Corollary 3)**.** *For any set $\mathcal{X}$, any points $x^{1:N} \in \mathcal{X}^N$, any class $\mathcal{F}$ of functions on $\mathcal{X}$ taking values in $[0, K]$ with pseudo-dimension $V_{\mathcal{F}^+} < \infty$, and any $\varepsilon > 0$,*

$$\mathcal{N}_1(\varepsilon, \mathcal{F}, x^{1:N}) \leq e(V_{\mathcal{F}^+} + 1)\left(\frac{2eK}{\varepsilon}\right)^{V_{\mathcal{F}^+}}.$$

Lemma E.4. *We have*

$$\mathcal{N}_1(\varepsilon, \Pi, x^{1:N}) \leq \prod_{k=1}^{d_{\mathcal{A}}} \mathcal{N}_1(\varepsilon/d_{\mathcal{A}}, \Pi_k, x^{1:N}).$$

Proof. The lemma follows directly from

$$\frac{1}{N}\sum_{t=1}^{N} \|\pi(x_t) - \pi'(x_t)\|_1 \leq \sum_{k=1}^{d_{\mathcal{A}}} \frac{1}{N}\sum_{t=1}^{N} |\pi_k(x_t) - \pi'_k(x_t)|.$$

□

Lemma E.5. *Let Assumption A8 and A9 hold for the function space $\mathcal{F}$ and policy set Π. Fix $x^{1:N} \in \mathcal{X}^N$. Then*

$$\begin{aligned}
&\mathcal{N}_1(\varepsilon, \mathcal{F}_\Pi^\vee, x^{1:N}) \\
&\quad\leq e^{d_{\mathcal{A}}+1}(V_{\mathcal{F}^+} + 1)\left(\prod_{k=1}^{d_{\mathcal{A}}}(V_{\Pi_k^+} + 1)\right) {Q_{\max}}^{V_{\mathcal{F}^+}} (d_{\mathcal{A}} A_\infty)^{V_{\Pi^+}} \left(\frac{4e(L_{\mathcal{A}}+1)}{\varepsilon}\right)^{V_{\mathcal{F}^+}+V_{\Pi^+}}.
\end{aligned}$$

Proof. By Lemma E.2 with $\alpha = L_{\mathcal{A}}/(L_{\mathcal{A}}+1)$,

$$\mathcal{N}_1((L_{\mathcal{A}}+1)\varepsilon', \mathcal{F}_\Pi^\vee, x^{1:N}) \leq \mathcal{N}_1(\varepsilon', \Pi, x^{1:N}) \sup_{\pi \in \Pi} \mathcal{N}_1(\varepsilon', \mathcal{F}, \langle \pi(x^{1:N}) \rangle),$$

and by Lemma E.4, the covering number of Π is bounded by $\prod_{k=1}^{d_{\mathcal{A}}} \mathcal{N}_1(\varepsilon'/d_{\mathcal{A}}, \Pi_k, x^{1:N})$. To bound these factors, we use Corollary 3 from [24] that was cited here as Proposition E.3. The pseudo-dimensions of $\mathcal{F}$ and Π_k are $V_{\mathcal{F}^+}$ and $V_{\Pi_k^+}$, respectively, and the ranges of functions from $\mathcal{F}$ and

Π_k have lengths $2Q_{\max}$ and $2A_\infty$, respectively. Thus

$$\begin{aligned}
&\mathcal{N}_1((L_{\mathcal{A}}+1)\varepsilon', \mathcal{F}_\Pi^\vee, x^{1:N}) \\
&\quad\le\quad e(V_{\mathcal{F}^+}+1)\left(\frac{4eQ_{\max}}{\varepsilon'}\right)^{V_{\mathcal{F}^+}} \prod_{k=1}^{d_{\mathcal{A}}}\left(e(V_{\Pi_k^+}+1)\left(\frac{4ed_{\mathcal{A}}A_\infty}{\varepsilon'}\right)^{V_{\Pi_k^+}}\right) \\
&\quad=\quad e^{d_{\mathcal{A}}+1}(V_{\mathcal{F}^+}+1)\left(\prod_{k=1}^{d_{\mathcal{A}}}(V_{\Pi_k^+}+1)\right) {Q_{\max}}^{V_{\mathcal{F}^+}} (d_{\mathcal{A}}A_\infty)^{V_{\Pi^+}} \left(\frac{4e}{\varepsilon'}\right)^{V_{\mathcal{F}^+}+V_{\Pi^+}}.
\end{aligned}$$

Substituting $\varepsilon = (L_{\mathcal{A}}+1)\varepsilon'$ yields the result. □

Finally, we will need the following technical lemma that transforms high probability bounds available for β-mixing processes into deviation size estimates:

Lemma E.6 ([7], Lemma 13)**.** *Let $\beta_m \le \overline{\beta}\exp(-bm^\kappa)$, $N \ge 1$, $k_N = \lceil (C_2 N\varepsilon^2/b)^{\frac{1}{1+\kappa}} \rceil$, $m_N = N/(2k_N)$, $0 < \delta \le 1$, $V \ge 2$, and C_1, C_2, $\overline{\beta}$, b, $\kappa > 0$. Further, define ε and Λ by*

$$\varepsilon = \sqrt{\frac{\Lambda(\Lambda/b \vee 1)^{1/\kappa}}{C_2 N}} \tag{15}$$

with $\Lambda = (V/2)\log N + \log(e/\delta) + \log^+(C_1 C_2^{V/2} \vee \overline{\beta})$. Then

$$C_1\left(\frac{1}{\varepsilon}\right)^V e^{-4C_2 m_N \varepsilon^2} + 2m_N\beta_{k_N} < \delta.$$

E.2 The error of a single update

Lemma E.7 (PAC-bound for the value fitting procedure)**.** *Let Assumption A1 and A2 hold, and fix the set of admissible functions $\mathcal{F}$ satisfying Assumptions A8 and A9 and the set of policies Π satisfying Assumption A9. Let Q be a real-valued random function over $\mathcal{X}\times\mathcal{A}$, $Q(\omega) \in \mathcal{F}$ and $\hat{\pi}$ be a random policy in Π, $\hat{\pi}(\omega) \in \Pi$ (possibly not independent from the sample path). Let f' be defined by*

$$f' = \operatorname*{argmin}_{f\in\mathcal{F}} \hat{L}_N(f; Q, \hat{\pi}).$$

For $0 < \delta \le 1$, $N \ge 1$, with probability at least $1-\delta$,

$$\left\|f' - T^{\hat{\pi}}Q\right\|_\nu^2 \le E_1^2(\mathcal{F};\hat{\pi}) + \sqrt{\frac{\Lambda_N(\delta)(\Lambda_N(\delta)/b \vee 1)^{1/\kappa}}{C_2 N}},$$

where $\Lambda_N(\delta)$ and C_2 are defined as in Theorem B.1.

Proof. We define $\hat{Q}_t = R_t + \gamma Q(X_{t+1}, \hat{\pi}(X_{t+1}))$. Note that, for fixed, deterministic Q and $\hat{\pi}$,

$$\mathbb{E}\left[\hat{Q}_t | X_t, A_t\right] = r(X_t, A_t) + \gamma\int_y Q(y, \hat{\pi}(y))\, dP(y|X_t, A_t) = (T^{\hat{\pi}}Q)(X_t, A_t),$$

that is, $T^{\hat{\pi}}Q$ is the regression function of $\hat{Q}_t$ given (X_t, A_t). What we have to show is that the chosen f' is a good estimate for $T^{\hat{\pi}}Q$ with high probability, noting that Q and $\hat{\pi}$ may not be independent from the sample path.

We can assume that $|\mathcal{F}| \ge 2$ (otherwise the bound is obvious). This implies $V_{\mathcal{F}^+} \ge 1$, and thus $V \ge 2$. Let ε and $\Lambda_N(\delta)$ be chosen as in (15):

$$\varepsilon = \sqrt{\frac{\Lambda_N(\delta)(\Lambda_N(\delta)/b \vee 1)^{1/\kappa}}{C_2 N}}$$

with $\Lambda_N(\delta) = (V/2)\log N + \log(e/\delta) + \log^+(C_1 C_2^{V/2} \vee \overline{\beta}) \ge 1$. Define

$$P_0 \stackrel{\text{def}}{=} \mathbb{P}\left(\left\|f' - T^{\hat{\pi}}Q\right\|_\nu^2 - E_1^2(\mathcal{F};\hat{\pi}) > \varepsilon\right).$$

It follows that it is sufficient to prove that $P_0 < \delta$.

Remember that for $\hat{\pi}$ arbitrary, we defined the following losses:

$$L(f; Q, \hat{\pi}) = L^*(Q, \hat{\pi}) + \left\| f - T^{\hat{\pi}} Q \right\|_\nu^2 .$$

These imply that

$$\left\| f' - T^{\hat{\pi}} Q \right\|_\nu^2 - \inf_{f \in \mathcal{F}} \left\| f - T^{\hat{\pi}} Q \right\|_\nu^2 = L(f'; Q, \hat{\pi}) - \inf_{f \in \mathcal{F}} L(f; Q, \hat{\pi}) = L(f'; Q, \hat{\pi}) - L_{\mathcal{F},Q,\hat{\pi}},$$

where $L_{\mathcal{F},Q,\hat{\pi}} = \inf_{f \in \mathcal{F}} L(f; Q, \hat{\pi})$ is the error of the function with minimum loss in our class.

Now, since $f' = \operatorname{argmin}_{f \in \mathcal{F}} \hat{L}_N(f; Q, \hat{\pi})$,

$$\begin{aligned}
& L(f'; Q, \hat{\pi}) - L_{\mathcal{F},Q,\hat{\pi}} \\
&= L(f'; Q, \hat{\pi}) - \hat{L}_N(f'; Q, \hat{\pi}) + \hat{L}_N(f'; Q, \hat{\pi}) - \inf_{f \in \mathcal{F}} L(f; Q, \hat{\pi}) \\
&\leq |\hat{L}_N(f'; Q, \hat{\pi}) - L(f'; Q, \hat{\pi})| + \inf_{f \in \mathcal{F}} \hat{L}_N(f; Q, \hat{\pi}) - \inf_{f \in \mathcal{F}} L(f; Q, \hat{\pi}) \\
&\quad \text{(by the definition of } f') \\
&\leq 2 \sup_{f \in \mathcal{F}} |\hat{L}_N(f; Q, \hat{\pi}) - L(f; Q, \hat{\pi})| \\
&\leq 2 \sup_{\hat{\pi} \in \Pi, Q, f \in \mathcal{F}} |\hat{L}_N(f; Q, \hat{\pi}) - L(f; Q, \hat{\pi})|.
\end{aligned}$$

Thus we get

$$P_0 \leq \mathbb{P}\left(\sup_{\hat{\pi} \in \Pi, Q, f \in \mathcal{F}} |\hat{L}_N(f; Q, \hat{\pi}) - L(f; Q, \hat{\pi})| > \varepsilon/2 \right).$$

Hence, in the subsequent statements, Q and $\hat{\pi}$ denote an arbitrary (deterministic) function in $\mathcal{F}$ and policy in Π, respectively.

We follow the line of proof of [22]. For any f,$Q \in \mathcal{F}$, $\hat{\pi} \in \Pi$, define the loss function $l_{f,Q,\hat{\pi}} : \mathcal{X} \times \mathcal{A} \times [-\hat{R}_{\max}, \hat{R}_{\max}] \times \mathcal{X} \to \mathbb{R}$ in accordance with (5) as

$$l_{f,Q,\hat{\pi}}(z) = l_{f,Q,\hat{\pi}}(x, a, r, y) \stackrel{\text{def}}{=} \frac{1}{\lambda(\mathcal{A}) \pi_b(a|x)} |f(x, a) - r - \gamma Q(y, \hat{\pi}(y))|^2$$

for $z = (x, a, r, y)$ and $\mathcal{L}_{\mathcal{F}} \stackrel{\text{def}}{=} \{l_{f,Q,\hat{\pi}}\}$ $f, Q \in \mathcal{F}, \hat{\pi} \in \Pi$. Introduce $Z_t = (X_t, A_t, R_t, X_{t+1})$ for $t = 0,\ldots,N$. Note that the process $\{Z_t\}$ is β-mixing with mixing coefficients $\{\beta_{m-1}\}$.

Observe that by (5)

$$l_{f,Q,\hat{\pi}}(Z_t) = \frac{1}{\lambda(\mathcal{A}) \pi_b(A_t|X_t)} |f(X_t, A_t) - \hat{Q}_t|^2 = L^{(t)},$$

hence we have for any f,$Q \in \mathcal{F}$, $\hat{\pi} \in \Pi$,

$$\frac{1}{N} \sum_{t=1}^{N} l_{f,Q,\hat{\pi}}(Z_t) = \hat{L}_N(f; Q, \hat{\pi}),$$

and (by (10))

$$\mathbb{E}\left[l_{f,Q,\hat{\pi}}(Z_t) \right] = \mathbb{E}\left[L^{(t)} \right] = L(f; Q, \hat{\pi})$$

(coincidentally with (7)). This reduces the bound to a uniform tail probability of an empirical process over $\mathcal{L}_{\mathcal{F}}$:

$$P_0 \leq \mathbb{P}\left(\sup_{Q, f \in \mathcal{F}, \hat{\pi} \in \Pi} \left| \frac{1}{N} \sum_{t=1}^{N} l_{f,Q,\hat{\pi}}(Z_t) - \mathbb{E}\left[l_{f,Q,\hat{\pi}}(Z_0) \right] \right| > \varepsilon/2 \right).$$

Since the samples are correlated, Pollard's tail inequality cannot be used directly. Hence we use the method of [23], as mentioned previously in Section E.1. For this we split the N samples into

$2m_N$ blocks which come in pairs (for simplicity we assume that splitting can be done exactly), i.e., $N = 2m_N k_N$. Introduce the following blocks, each having the same length, k_N:

$$\underbrace{Z_1,\dots,Z_{k_N}}_{H_1},\underbrace{Z_{k_N+1},\dots,Z_{2k_N}}_{T_1},\underbrace{Z_{2k_N+1},\dots,Z_{3k_N}}_{H_2},\underbrace{Z_{3k_N+1},\dots,Z_{4k_N}}_{T_2},\dots$$
$$\dots,\underbrace{Z_{(2m_N-2)k_N+1},\dots,Z_{(2m_N-1)k_N}}_{H_{m_N}},\underbrace{Z_{(2m_N-1)k_N+1},\dots,Z_{2m_N k_N}}_{T_{m_N}}.$$

Here $H_i \stackrel{\text{def}}{=} \{2k_N(i-1)+1,\dots,2k_N(i-1)+k_N\}$ and $T_i \stackrel{\text{def}}{=} \{2ik_N-(k_N-1),\dots,2ik_N\}$. Next, we introduce the block-independent "ghost" samples as it was done by [23] and [22]:

$$\underbrace{Z'_1,\dots,Z'_{k_N}}_{H_1},\quad \underbrace{Z'_{2k_N+1},\dots,Z'_{3k_N}}_{H_2},\quad \dots\quad \underbrace{Z'_{(2m_N-2)k_N+1},\dots,Z'_{(2m_N-1)k_N}}_{H_{m_N}},$$

where any particular block has the same marginal distribution as originally, but the m_N blocks are independent of one another. Introduce $H = \bigcup_{i=1}^{m_N} H_i$.

For this ansatz we use Lemma E.1 above with $\mathcal{Z} = \mathcal{X}\times\mathcal{A}\times\mathbb{R}\times\mathcal{X}$, $\mathcal{F} = \mathcal{L}_{\mathcal{F}}$ noting that any $l_{f,Q,\hat\pi} \in \mathcal{L}_{\mathcal{F}}$ is bounded by

$$K = \frac{\tilde R_{\max}^2}{\lambda(\mathcal{A})\pi_0}$$

with $\tilde R_{\max} = (1+\gamma)Q_{\max} + \hat R_{\max}$, to get the bound

$$\mathbb{P}\left(\sup_{Q,f\in\mathcal{F},\hat\pi\in\Pi}\left|\frac{1}{N}\sum_{t=1}^{N} l_{f,Q,\hat\pi}(Z_t) - \mathbb{E}\left[l_{f,Q,\hat\pi}(Z_0)\right]\right| > \varepsilon/2\right)$$
$$\le 16\mathbb{E}\left[\mathcal{N}_1(\varepsilon/16,\mathcal{L}_{\mathcal{F}},(Z'_t;t\in H))\right] e^{-\frac{m_N}{2}\left(\frac{\lambda(\mathcal{A})\pi_0\varepsilon}{16\tilde R_{\max}^2}\right)^2} + 2m_N\beta_{k_N}.$$

By some calculation, the distance in $\mathcal{L}_{\mathcal{F}}$ can be bounded as follows:

$$\begin{aligned}
&\frac{2}{N}\sum_{t\in H}|l_{f,Q,\hat\pi}(Z'_t) - l_{g,\tilde Q,\tilde{\hat\pi}}(Z'_t)| \\
&= \frac{2}{N\lambda(\mathcal{A})}\sum_{t\in H}\frac{1}{\pi_b(A'_t|X'_t)}\Big||f(X'_t,A'_t) - R'_t - \gamma Q(X'_{t+1},\hat\pi(X'_{t+1}))|^2 \\
&\quad - |g(X'_t,A'_t) - R'_t - \gamma\tilde Q(X'_{t+1},\tilde{\hat\pi}(X'_{t+1}))|^2\Big| \\
&\le \frac{2}{N\lambda(\mathcal{A})}\sum_{t\in H}\frac{2\tilde R_{\max}}{\pi_0}\Big(|f(X'_t,A'_t) - g(X'_t,A'_t)| + \gamma|\tilde Q(X'_{t+1},\tilde{\hat\pi}(X'_{t+1})) - Q(X'_{t+1},\hat\pi(X'_{t+1}))|\Big) \\
&\quad \text{(using the identity } a^2-b^2=(a+b)(a-b)\text{, the triangle inequality,} \\
&\quad \text{and the assumed bounds for } \pi_b,\ f,\ g,\ Q,\ \tilde Q,\text{ and } R'_t) \\
&= \frac{2\tilde R_{\max}}{\lambda(\mathcal{A})\pi_0}\left(\frac{2}{N}\sum_{t\in H}|f(X'_t,A'_t) - g(X'_t,A'_t)| + \gamma\frac{2}{N}\sum_{t\in H}|\tilde Q(X'_{t+1},\tilde{\hat\pi}(X'_{t+1})) - Q(X'_{t+1},\hat\pi(X'_{t+1}))|\right).
\end{aligned}$$

Note that the first term is $\mathcal{D}' = ((X'_t,A'_t);t\in H)$-based L^1-distances of functions in $\mathcal{F}$, while the second term is just (γ-times) the $\mathcal{D}'_+ = (X'_{t+1};t\in H)$-based L^1-distance of two functions in $\mathcal{F}_\Pi^\vee$ corresponding to $(Q,\hat\pi)$ and $(\tilde Q,\tilde{\hat\pi})$. This leads to

$$\mathcal{N}_1\left(\frac{2\tilde R_{\max}}{\lambda(\mathcal{A})\pi_0}(1+\gamma\alpha')\varepsilon',\mathcal{L}_{\mathcal{F}},(Z'_t;t\in H)\right) \le \mathcal{N}_1(\varepsilon',\mathcal{F},\mathcal{D}')\mathcal{N}_1(\alpha'\varepsilon',\mathcal{F}_\Pi^\vee,\mathcal{D}'_+)$$

for any $\alpha' > 0$. Applying now Proposition E.3 for the first factor and Lemma E.5 for the second one, with $\alpha' = L_{\mathcal{A}} + 1$, we have

$$\begin{aligned}
&\mathcal{N}_1\left(\frac{2\tilde{R}_{\max}}{\lambda(\mathcal{A})\pi_0}(1+\gamma(L_{\mathcal{A}}+1))\varepsilon', \mathcal{L}_{\mathcal{F}}, (Z_t'; t\in H)\right)\\
&\quad\le\quad e(V_{\mathcal{F}^+}+1)\left(\frac{4eQ_{\max}}{\varepsilon'}\right)^{V_{\mathcal{F}^+}} e^{d_{\mathcal{A}}+1}(V_{\mathcal{F}^+}+1)\left(\prod_{k=1}^{d_{\mathcal{A}}}(V_{\Pi_k^+}+1)\right) {Q_{\max}}^{V_{\mathcal{F}^+}}(d_{\mathcal{A}}A_\infty)^{V_{\Pi^+}}\left(\frac{4e}{\varepsilon'}\right)^{V_{\mathcal{F}^+}+V_{\Pi^+}}\\
&\quad=\quad e^{d_{\mathcal{A}}+2}(V_{\mathcal{F}^+}+1)^2\left(\prod_{k=1}^{d_{\mathcal{A}}}(V_{\Pi_k^+}+1)\right) {Q_{\max}}^{2V_{\mathcal{F}^+}}(d_{\mathcal{A}}A_\infty)^{V_{\Pi^+}}\left(\frac{4e}{\varepsilon'}\right)^{V},
\end{aligned}$$

where $V = 2V_{\mathcal{F}^+} + V_{\Pi^+}$ is the "combined effective" dimension, and thus

$$\begin{aligned}
&\mathcal{N}_1(\varepsilon/16, \mathcal{L}_{\mathcal{F}}, (Z_t'; t\in H))\\
&\quad\le\quad e^{d_{\mathcal{A}}+2}(V_{\mathcal{F}^+}+1)^2\left(\prod_{k=1}^{d_{\mathcal{A}}}(V_{\Pi_k^+}+1)\right) {Q_{\max}}^{2V_{\mathcal{F}^+}}(d_{\mathcal{A}}A_\infty)^{V_{\Pi^+}}\left(\frac{128e\tilde{R}_{\max}(1+\gamma(L_{\mathcal{A}}+1))}{\lambda(\mathcal{A})\pi_0\varepsilon}\right)^{V}\\
&\quad=\quad \frac{C_1}{16}\left(\frac{1}{\varepsilon}\right)^{V},
\end{aligned}$$

with $C_1 = C_1(\lambda(\mathcal{A}), V_{\mathcal{F}^+}, (V_{\Pi_k^+})_{k=1}^{d_{\mathcal{A}}}, Q_{\max}, \hat{R}_{\max}, A_\infty, \gamma, \pi_0, L_{\mathcal{A}}, d_{\mathcal{A}})$.

Putting together the above bounds we get

$$P_0 \le C_1\left(\frac{1}{\varepsilon}\right)^{V} e^{-\frac{\lambda(\mathcal{A})^2\pi_0^2 m_N\varepsilon^2}{512\tilde{R}_{\max}^4}} + 2m_N\beta_{k_N} = C_1\left(\frac{1}{\varepsilon}\right)^{V} e^{-4C_2 m_N\varepsilon^2} + 2m_N\beta_{k_N}, \tag{16}$$

where $C_2 = \frac{1}{2}\left(\frac{\lambda(\mathcal{A})\pi_0}{32\tilde{R}_{\max}^2}\right)^2 = \frac{\lambda(\mathcal{A})^2\pi_0^2}{2048\tilde{R}_{\max}^4}$. Defining $k_N = \lceil (C_2 N\varepsilon^2/b)^{\frac{1}{1+\kappa}}\rceil$ and $m_N = N/(2k_N)$, the proof is finished by Lemma E.6, which, together with (16), implies $P_0 < \delta$. □

E.3 Controlling the error of the approximate greedy step

Remember that $E : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X})$ is defined by $(EQ)(x) = \sup_{a\in\mathcal{A}} Q(x,a)$, while $E^\pi : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X})$ is defined by $(E^\pi Q)(x) = Q(x,\pi(x))$. In this section we are interested in bounding $\left\|TQ - T^{\hat{\pi}}Q\right\|_{1,\nu}$. The proof will consists of several steps. First, we bound $\left\|T^{\hat{\pi}}Q - TQ\right\|_{1,\nu}$ by $\nu(EQ - E^{\hat{\pi}}Q)$. Then we show that $\nu E^{\hat{\pi}}Q$ is close to $\sup_{\pi\in\Pi}\nu E^\pi Q$, from which the result will follow.

Lemma E.8. *Under Assumption A5*

$$\|TQ - T^\pi Q\|_{1,\nu} \le \gamma\Gamma_\nu\,\nu(EQ - E^\pi Q) \tag{17}$$

holds for any $Q \in B(\mathcal{X}\times\mathcal{A})$ and policy π.

Proof. Let $\Delta = EQ - E^\pi Q$, $\mu = (\nu\times\lambda_{\mathcal{A}})P$. By noting that $EQ \ge E^\pi Q$ and so $TQ \ge T^\pi Q$ and $\Delta \ge 0$ we have

$$\|TQ - T^\pi Q\|_{1,\nu} = \left\|\gamma\int \Delta\, dP(\cdot|x,a)\right\|_{1,\nu} = \gamma\int\Delta\, d\mu = \gamma\int\Delta\frac{d\mu}{d\nu}\,d\nu \le \gamma\Gamma_\nu\,\nu\Delta.$$

□

Remember that $\mathcal{F}_\Pi^\vee = \{\, f : f(\cdot) = Q(\cdot,\pi(\cdot)), Q\in\mathcal{F}, \pi\in\Pi\,\}$. Combining Lemmas E.1, E.5, and E.6 we get the following result:

Lemma E.9. *Let Assumptions A2, A8, A9 hold. Then, with probability $1-\delta$,*

$$\sup_{V\in\mathcal{F}_\Pi^\vee}\left|\mathbb{E}\left[V(X_1)\right] - \frac{1}{N}\sum_{t=1}^N V(X_t)\right| \le \sqrt{\frac{\Lambda_N'(\delta)(\Lambda_N'(\delta)/b\vee 1)^{1/\kappa}}{C_2' N}}. \tag{18}$$

where $\Lambda_N'(\delta)$ and C_2' are defined as in Theorem B.1.

Proof. Let ε and $\Lambda'_N(\delta)$ be chosen as in (15):

$$\varepsilon = \sqrt{\frac{\Lambda'_N(\delta)(\Lambda'_N(\delta)/b \vee 1)^{1/\kappa}}{C'_2 N}}$$

with $\Lambda'_N(\delta) = (V'/2)\log N + \log(e/\delta) + \log^+ (C'_1 {C'_2}^{V'/2} \vee \overline{\beta}) \geq 1$. Define

$$P'_0 = \mathbb{P}\left(\sup_{V \in \mathcal{F}_\Pi^\vee} \left| \mathbb{E}\left[V(X_1)\right] - \frac{1}{N}\sum_{t=1}^N V(X_t) \right| > \varepsilon \right).$$

Let $Z'_t = (X'_t, A'_t, R'_t)$, m_N, k_N, and H be as in the proof of Lemma E.7.

We use Lemma E.1 above with $\mathcal{Z} = \mathcal{X}$, $\mathcal{F} = \mathcal{F}_\Pi^\vee$ noting that any $f \in \mathcal{F}_\Pi^\vee$ is bounded by $Q_{\max}$ and that any deterministic image of a β-mixing process is also β-mixing with the same or faster rate, to get the bound

$$P'_0 \leq 16\mathbb{E}\left[\mathcal{N}_1(\varepsilon/8, \mathcal{F}_\Pi^\vee, (X'_t; t \in H))\right] e^{-\frac{m_N \varepsilon^2}{128 {Q_{\max}}^2}} + 2m_N \beta_{k_N+1}.$$

Applying Lemma E.5 we get

$$\begin{aligned}
&\mathcal{N}_1(\varepsilon/8, \mathcal{F}_\Pi^\vee, (X'_t; t \in H)) \\
&\leq \; e^{d_\mathcal{A}+1}(V_{\mathcal{F}^+}+1)\left(\prod_{k=1}^{d_\mathcal{A}}(V_{\Pi_k^+}+1)\right) {Q_{\max}}^{V_{\mathcal{F}^+}} (d_\mathcal{A} A_\infty)^{V_{\Pi^+}} \left(\frac{32e(L_\mathcal{A}+1)}{\varepsilon}\right)^{V_{\mathcal{F}^+}+V_{\Pi^+}} \\
&= \; \frac{C'_1}{16}\left(\frac{1}{\varepsilon}\right)^{V'},
\end{aligned}$$

with $V' = V_{\mathcal{F}^+} + V_{\Pi^+}$ and $C'_1 = C'_1(V_{\mathcal{F}^+}, (V_{\Pi_k^+})_{k=1}^{d_\mathcal{A}}, Q_{\max}, A_\infty, L_\mathcal{A}, d_\mathcal{A})$. Putting together the above bounds we get

$$P'_0 \leq C'_1 \left(\frac{1}{\varepsilon}\right)^{V'} e^{-\frac{m_N \varepsilon^2}{128 {Q_{\max}}^2}} + 2m_N \beta_{k_N+1} = C'_1 \left(\frac{1}{\varepsilon}\right)^{V'} e^{-4C'_2 m_N \varepsilon^2} + 2m_N \beta_{k_N+1}, \tag{19}$$

where $C'_2 = \frac{1}{2}\frac{1}{(16 Q_{\max})^2} = \frac{1}{512 {Q_{\max}}^2}$. Defining $k_N = \lceil (C'_2 N \varepsilon^2/b)^{\frac{1}{1+\kappa}} \rceil$ and $m_N = N/(2k_N)$, the proof is finished by Lemma E.6, which, together with (19) and $\beta_{k_N+1} \leq \beta_{k_N}$ implies $P'_0 < \delta$. ☐

Remember that $e^*(\mathcal{F}, \Pi) = \sup_{f \in \mathcal{F}} \inf_{\pi \in \Pi} \nu(EQ - E^\pi Q)$. We are ready to prove the main result of this section:

Lemma E.10. *Let Assumptions A2, A5, A8, A9 hold. Let $Q \in \mathcal{F}$ be random, $\hat{\pi} = \operatorname{argmax}_{\pi \in \Pi} \sum_{t=1}^N Q(X_t, \pi(X_t))$. Then with probability at least $1-\delta$,*

$$\left\|TQ - T^{\hat{\pi}}Q\right\|_{1,\nu} \leq \gamma \Gamma_\nu \left[e^*(\mathcal{F}, \Pi) + 2\sqrt{\frac{\Lambda'_N(\delta)\,(\Lambda'_N(\delta)/b \vee 1)^{1/\kappa}}{C'_2 N}} \right], \tag{20}$$

where $\Lambda'_N(\delta)$ and C'_2 are defined as in Theorem B.1.

Proof. Let us introduce the empirical measure $\nu_N(\cdot) = \frac{1}{N}\sum_{t=1}^N \delta_{X_t}(\cdot)$, where $\delta_x(\cdot)$ is the counting measure associated with the singleton $\{x\}$. By Lemma E.8, $\left\|TQ - T^{\hat{\pi}}Q\right\|_{1,\nu} \leq \gamma \Gamma_\nu \nu(EQ - E^{\hat{\pi}}Q)$. Now, let us use the error decomposition

$$\begin{aligned}
\nu(EQ - E^{\hat{\pi}}Q) &= \inf_{\pi \in \Pi} \nu(EQ - E^\pi Q) + \nu(EQ - E^{\hat{\pi}}Q) - \inf_{\pi \in \Pi} \nu(EQ - E^\pi Q) \\
&\leq \sup_{f \in \mathcal{F}} \inf_{\pi \in \Pi} \nu(Ef - E^\pi f) + \sup_{\pi \in \Pi} \nu E^\pi Q - \nu E^{\hat{\pi}} Q \\
&= e^*(\mathcal{F}, \Pi) + \sup_{\pi \in \Pi} \nu E^\pi Q - \sup_{\pi \in \Pi} \nu_N E^\pi Q + \sup_{\pi \in \Pi} \nu_N E^\pi Q - \nu E^{\hat{\pi}} Q \\
&\leq e^*(\mathcal{F}, \Pi) + 2 \sup_{\pi \in \Pi, f \in \mathcal{F}} |\nu E^\pi f - \nu_N E^\pi f|.
\end{aligned}$$

Here we used that $\sup_{\pi \in \Pi} \nu_N E^\pi Q = \nu_N E^{\hat{\pi}} Q$ and the elementary inequality $\sup_x f(x) - \sup_y g(y) \leq \sup_x (f(x) - g(x))$. Finally, the right hand side is bounded by Lemma E.9 since $\sup_{\pi \in \Pi, f \in \mathcal{F}} |\nu E^\pi f - \nu_N E^\pi f| = \sup_{g \in \mathcal{F}_\Pi^\vee} |\nu g - \nu_N g|$. Combining these bounds yields the result. ☐

E.4 Proof of the Main Result

Proof. For the proof we write the algorithm in the form $Q_{k+1} = TQ_k + \varepsilon_k$ with $\varepsilon_k = \varepsilon'_k + \varepsilon''_k$ where $\varepsilon'_k = Q_{k+1} - T^{\hat{\pi}_k} Q_k$ is the error while computing $T^{\hat{\pi}_k} Q_k$ and $\varepsilon''_k = T^{\hat{\pi}_k} Q_k - TQ_k$ is the error committed because of using the approximately greedy policy $\hat{\pi}_k$ (see also (11)). The result then follows by Lemma D.3 if we can bound the $\|\cdot\|_{1,\nu}$ error of ε_k, $k = 0, \dots, K-1$. By the triangle inequality and the well-known relation of L^p norms, $\|\varepsilon_k\|_{1,\nu} \le \|\varepsilon'_k\|_{1,\nu} + \|\varepsilon''_k\|_{1,\nu} \le \|\varepsilon'_k\|_{2,\nu} + \|\varepsilon''_k\|_{1,\nu}$. Now, we can use Lemma E.7 to get a bound on $\|\varepsilon'_k\|_{2,\nu}$ that fails with probability at most $\delta/(2K)$, giving ε' defined in the text of the theorem. Similarly, we can use Lemma E.10 to get a bound on $\|\varepsilon''_k\|_{1,\nu}$ that fails with probability at most $\delta/(2K)$, giving rise to ε''. Since there are K iterations, the total failure probability is bounded by δ, thus finishing the proof of the theorem. □

Acknowledgments

András Antos would like to acknowledge support for this project from the Hungarian Academy of Sciences (Bolyai Fellowship). Csaba Szepesvári greatly acknowledges the support received from the Alberta Ingenuity Fund, NSERC, the Computer and Automation Research Institute of the Hungarian Academy of Sciences.

References

[1] A.Y. Ng and M. Jordan. PEGASUS: A policy search method for large MDPs and POMDPs. In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, pages 406–415, 2000. 1, 9

[2] L. Peshkin and C.R. Shelton. Learning from scarce experience. In *ICML*, pages 498–505, 2002. 1

[3] D. Aberdeen. Policy-gradient methods for planning. In *Advances in Neural Information Processing Systems 18*, pages 9–16. 2006. 1

[4] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. Bradford Book. MIT Press, 1998. 1, 4

[5] M. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003. 2

[6] A. Antos, Cs. Szepesvári, and R. Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. In *COLT-19*, pages 574–588, 2006. 2, 5

[7] A. Antos, Cs. Szepesvári, and R. Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2007. (accepted). 2, 5, 7, 10, 14, 16

[8] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005. 2, 4

[9] J.A. Boyan and A.W. Moore. Generalization in reinforcement learning: Safely approximating the value function. In *NIPS-7*, pages 369–376, 1995. 2, 4

[10] Geoffrey J. Gordon. Stable function approximation in dynamic programming. In A. Prieditis and S. Russell, editors, *Proc. of ICML 20*, pages 261–268. Morgan Kaufmann, 1995. 2

[11] D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49:161–178, 2002. 2

[12] M. Riedmiller. Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method. In *16th European Conference on Machine Learning*, pages 317–328, 2005. 2, 10

[13] S. Kalyanakrishnan and P. Stone. Batch reinforcement learning in a complex domain. In *AAMAS-07*, 2007. 2, 10

[14] A. Antos, Cs. Szepesvári, and R. Munos. Value-iteration based fitted policy iteration: learning with a single trajectory. In *IEEE ADPRL*, pages 330–337, 2007. 2, 5, 10

[15] D. P. Bertsekas and S.E. Shreve. *Stochastic Optimal Control (The Discrete Time Case)*. Academic Press, New York, 1978. 2

[16] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines (and other kernel-based learning methods)*. Cambridge University Press, 2000. 4

[17] P.L. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52:434–452, 1996. 5

[18] A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional space. *American Mathematical Society Translations*, 17(2):277–364, 1961. 5

[19] R. Munos and Cs. Szepesvári. Finite time bounds for sampling based fitted value iteration. Technical report, Computer and Automation Research Institute of the Hungarian Academy of Sciences, Kende u. 13-17, Budapest 1111, Hungary, 2006. 8, 9, 13, 14

[20] P.L. Bartlett and A. Tewari. Sample complexity of policy search with known dynamics. In *NIPS-19*. MIT Press, 2007. 9

[21] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. 9

[22] R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, April 2000. 14, 17, 18

[23] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, January 1994. 14, 17, 18

[24] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. 15