Sampling-based Approximations with Quantitative Performance for the Probabilistic Reach-Avoid Problem over General Markov Processes

Sofie Haesaert  
Control Systems Group, Department of Electrical Engineering  
Eindhoven University of Technology  
Eindhoven, The Netherlands

Robert Babuska  
Delft Center of Systems and Control  
Delft University of Technology  
Delft, The Netherlands

Alessandro Abate  
Department of Computer Science  
University of Oxford  
Oxford, United Kingdom

Abstract

This article deals with stochastic processes endowed with the Markov (memoryless) property and evolving over general (uncountable) state spaces. The models further depend on a non-deterministic quantity in the form of a control input, which can be selected to affect the probabilistic dynamics. We address the synthesis of optimal controllers that maximize the probability associated to a rather general property of interest, known as the “reach-avoid” specification. The reach-avoid specification deals with assessing the likelihood that any finite-horizon trajectory of the model enters a given goal set, while avoiding a given set of undesired states (both arbitrary subsets of the state space). Equivalently, the property can be expressed as entering a goal set, while dwelling within a set of allowed states. The reach-avoid property is a well known specification that lies at the core of a number of modal logics used in the field of formal verification, and which is relevant for safety-critical applications ranging from robotics to air traffic management. This article newly provides an approximate computational scheme for the reach-avoid specification based on the Fitted Value Iteration algorithm, which hinges on random sample extractions, and derives new formal probabilistic bounds on the error made by the approximation algorithm: as such, the output of the numerical scheme is quantitatively assessed and thus meaningful for safety-critical applications of the property of interest. Furthermore, this contribution provides tighter sample-based probabilistic error bounds for the computation of the reach-avoid problem based on the Fitted Value Iteration. The overall computational scheme is put in relationship with alternative approximation algorithms in the literature, and finally its performance is practically assessed over a benchmark case study.

Keywords: General state-space processes, reach-avoid problem, dynamic programming, fitted value iteration, computational approximation with error bounds.
1. Introduction

This contribution concerns a problem grounded in concepts from a number of different areas: we deal with probabilistic processes evolving over continuous (and in particular uncountable) state spaces – this leads to the use of measure-theoretical material from Stochastic Processes and Probability Theory (Meyn and Tweedie, 1993); we work with models endowed with a control input and investigate control synthesis, which relate to a broad literature in Control Theory (Kushner, 1967; Bertsekas and Shreve, 1996); furthermore, we are interested in quantifying the probability associated to a dynamical property, known as reach-avoid, which corresponds to a widely used model specification in the field of Formal Verification (Clarke et al., 1999; Baier and Katoen, 2008); and finally we employ a sampling-based algorithm to approximately compute the likelihood associated to the above specification. The algorithm, known as Fitted Value Iteration (FVI) (Munos and Szepesvari, 2008), is an advanced regression scheme developed in Machine Learning.

We focus on stochastic processes endowed with the Markov property (where the future is independent of the past, conditional on the present) and, aiming for generality, we deal with processes evolving over a continuous state space. The presence of an uncountable state space requires dealing with measure-theoretical techniques, which adds in complication over finite state models such as Markov chains (Meyn and Tweedie, 1993). We are in particular interested in a class of such models known as stochastic hybrid systems (SHS) (Abate et al., 2008a), which are endowed with a “hybrid” (that is, both continuous and discrete) state space. The results discussed in this work can be exported to SHS, and are thus relevant for a number of applications in Engineering and the Life Sciences (Blom and Lygeros, 2006; Cassandras and J. Lygeros (Eds.), 2006). This work investigates the problem of controller synthesis over the models at hand, namely the selection of sequences of control inputs (which in particular can be functions of the states of the model) over a finite horizon, in order to maximize or to minimize a given figure of merit (Kushner, 1967).

As for the figure of merit of interest in this work, we choose to go beyond the classical properties investigated in Systems and Control theory, which by and large deal with known and standard problems of stability, regulation, and tracking. Instead, in this work we focus on the reach-avoid specification, a property that is well known and central within the Formal Verification field (Clarke et al., 1999; Baier and Katoen, 2008). Notice that classical results in Formal Verification deal with simple models – usually finite-state transition systems or Markov chains – which allow for the development of decidable and computationally optimized results, and which mostly deal with verification tasks that do not involve policy synthesis. In this work instead we consider reach-avoid specifications over models with continuous stochastic transitions and endowed with control inputs.

The reach-avoid problem deals with computing the likelihood that, within a given fixed time horizon, any finite-horizon trajectory of the model enters a given goal set, while avoiding a given set of undesired states (both sets are arbitrary subsets of the state space). Equivalently, the property can be expressed as the probability that the process enters a goal set while dwelling within a set of allowed states. The reach-avoid property is a generalization of widely used properties such as reachability and invariance, and represents a known specification (denoted as “bounded until”) that lies at the core of a number of modal logics used in the field of formal verification, such as Linear Temporal Logic and Computational...
Quantitative Approximations for Reach-Avoid over General Markov Processes

Time Logic (Clarke et al., 1999; Baier and Katoen, 2008). From a controller synthesis perspective, the goal becomes that of either maximizing or minimizing the above specification, with the allowed control inputs and over the given time horizon. As an example for the reach-avoid property, consider the model of an aircraft and seeking a path to maximize the likelihood of navigating toward a given target in the airspace, while avoiding the presence of a storm.

In the context of probabilistic models evolving over continuous domains and in discrete time (which is the framework considered in this work), the probabilistic reachability and reach-avoid specifications have been investigated in (Abate et al., 2008a; Summers and Lygeros, 2010). These results have recently led to the study of other properties, either richer (Abate et al., 2011) or defined over an unbounded time horizon (Tkachev and Abate, 2014). The above results have focused on the theoretical characterization of the specifications/properties of interest: of course it is also of much interest to provide algorithms that can numerically compute these figures. Computational approaches to probabilistic reachability have been studied in (Abate et al., 2010; Esmaeil Zadeh Soudjani and Abate, 2013): the strength of these results is that the proposed numerical schemes have explicitly quantified error bounds. This is unlike other, known approximation schemes in the literature (Kushner and Dupuis, 2001; Prandini and Hu, 2006; Koutsoukos and Riley, 2006), which provide results with properties that are only known asymptotically. Further, since the aforementioned formal approximation approaches are based on discretization techniques that result in finite-space models, they are also promising as automatic software tools for verification and synthesis, known in Formal Verification as model checkers (Hinton et al., 2006; Katoen et al., 2005).

This article provides a new approximate computational scheme for the reach-avoid specification based on the Fitted Value Iteration algorithm, which hinges on random sample extractions. Further, this work originally derives formal probabilistic bounds on the error made by the approximation algorithm. In order to do so, the FVI scheme is tailored to the characterization of the reach-avoid problem, which leads to Dynamic Programming (DP) recursions based on a sum-multiplicative form that is non-standard since it departs from the classical additive (possibly discounted) cost functions (Bertsekas and Shreve, 1996). Starting from the regression bounds in Munos and Szepesvari (2008), this work includes new results on the error for the FVI approximation and a-priori performance guaranties. Additionally, novel and tighter probabilistic error bounds for dynamic programming solutions of the reach-avoid problem based on samples extraction have been developed. It is of interest to benchmark the outcomes of the proposed approach with those of the alternative techniques in the literature (Abate et al., 2010; Esmaeil Zadeh Soudjani and Abate, 2013): we will see how the related techniques provide bounds that are valid deterministically, whereas the proposed result yields in general tighter results that are valid with a certain (tunable) confidence. The outcomes lead to an approach providing controller synthesis with a certified performance, which is relevant in a number of safety-critical applications (Blom and Lygeros, 2006).

The article is structured as follows. Section 2 introduces the model and provides a theoretical characterization of the reach-avoid problem – it also puts forward the two alternative approaches to compute the quantities of interest. Among the two approaches, in Section 3 we introduce the FVI algorithm and tailor it to the reach-avoid problem under study, discussing in detail its implementation. Probabilistic bounds on the error of the FVI
scheme versus the theoretical characterization of the reach-avoid problem are discussed in Section 4. The a-priori, computable probabilistic bounds of Section 4 are complemented with a-posteriori, sample-based probabilistic bounds in Section 5. Section 6 details the application of the FVI algorithm to a case study, and Section 7 concludes the work. The proofs of the statements are included in the Appendix.

2. Probabilistic Reach-Avoid Problem over General Markov Processes

2.1 Models

Definition 1 (General Markov process) A discrete-time general Markov process is comprised of:

- A continuous (uncountable) state space $\mathcal{X} \subset \mathbb{R}^n$;
- An action space $\mathcal{A} = \{a_1, \ldots, a_m\}$ consisting of a finite number of actions;
- A Borel-measurable stochastic kernel $T_x$, which assigns to each state-action pair $x \in \mathcal{X}$ and $a \in \mathcal{A}$ a probability distribution $T_x(\cdot | x, a)$ over $\mathcal{X}$.

We denote with $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P)$ a probability structure on $\mathcal{X}$, where $\mathcal{B}(\mathcal{X})$ is the $\sigma$-algebra associated to $\mathcal{X}$ and $P$ is characterized as $P(y \in A | x \in \mathcal{X}, a \in \mathcal{A}) = \int_A T_x(dy | x, a)$. We assume that the stochastic kernels admit densities so that $\int_A T_x(dy | x, a) = \int_A t_x(y | x, a) dy$.

The notion of a Markov policy is given as follows.

Definition 2 (Markov policy) A Markov policy $\mu$ over the horizon $[0, N_t]$ is a sequence $\mu = (\mu_0, \mu_1, \ldots, \mu_{N_t-1})$ of universally measurable maps, $\mu_k : \mathcal{X} \to \mathcal{A}$, $k = 0, 1, \ldots, N_t - 1$, from the state space $\mathcal{X}$ to the action space $\mathcal{A}$. The set of Markov policies is denoted as $\mathcal{M}$.

The evolution of the general Markov process is considered over a finite horizon $k = 0, 1, \ldots, N_t$, with $N_t \in \mathbb{N}$. Consider a discrete-time general Markov process, a Markov policy $\mu$, a deterministic initial state $x_0 \in \mathcal{X}$ and a finite time horizon $N_t$: an execution of the process characterizes a state trajectory given as $\{x_k | k = 0, 1, \ldots, N_t\}$. The process evolves over the product space $(\mathcal{X})^{N_t+1}$, which is again endowed with a (product) $\sigma$-algebra and allows computing probability associated to events over trajectories – we denote this probability by $\mathbb{P}$, and further define the probabilities $\mathbb{P}_{x_0}, \mathbb{P}^\mu_{x_0}$ as $\mathbb{P}$ conditioned on an initial state and on an initial state and a policy, respectively. The state at the $(k + 1)$-st time instant, $x_{k+1}$, is obtained as a realization of the controlled Borel-measurable stochastic kernel $T_x (\cdot | x_k, \mu_k(x_k))$. As a trivial generalization, the model can be initialized according to an initial probability measure $P_0 \in M(\mathcal{X})$ (cf. Remarks 5, 6), where $M(\cdot)$ denotes the set of probability distributions over a given set.

2.2 The Probabilistic Reach-Avoid Problem: Definition and Characterization

Let us introduce the probabilistic reach-avoid problem, also known as constrained reachability (Baier and Katoen, 2008), and provide its characterization. Consider a finite-horizon
reach-avoid property over a safe set $A \in \mathcal{B}(\mathcal{X})$, a target set $K \in \mathcal{B}(\mathcal{X})$, and a time horizon $N_t \in \mathbb{N}$. A given state trajectory $\{x_k | k = 0, 1, \ldots, N_t\}$ verifies the reach-avoid property if the target set $K$ is reached within the time horizon, while staying inside the safe set $A$. This property can be expressed logically as

$$\exists j \in [0, N_t] : x_j \in K \land \forall i \in [0, j - 1] : x_i \in A \setminus K.$$

Let us now consider the probabilistic reach-avoid property for a general stochastic system, defined as the probability that an execution associated with a fixed Markov policy $\mu \in \mathcal{M}$ and an initial condition $x_0 \in \mathcal{X}$ reaches the target set $K$ while avoiding $\mathcal{X} \setminus A$. This is formalized as

$$r^\mu_{x_0}(K, A) = \mathbf{P}^\mu_{x_0} \left\{ \exists j \in [0, N_t] : x_j \in K \land \forall i \in [0, j - 1] : x_i \in A \setminus K \right\}, \quad (1)$$

where the consecutive states $x_0, x_1, \ldots, x_{N_t} \in \mathcal{X}$ are sampled via the stochastic kernel $T_x$ under policy $\mu$. The logical formula contained in (1) can be written as a boolean expression by using indicator functions, which leads to its expression as an expectation over the state trajectories as

$$r^\mu_{x_0}(K, A) = \mathbf{E}_{x_0}^\mu \left[ \sum_{j \in [0, N_t]} \left( \prod_{i=0}^{j-1} 1_{A \setminus K}(x_i) \right) 1_K(x_j) \right],$$

where $1_B(x) = 1$ if $x \in B$, else it is equal to 0. The reach-avoid problem subsumes other known problems widely studied in System and Control Theory and in Formal Verification, such as that of reachability of set $K$, which is simply obtained by selecting $A = \mathcal{X}$, or that of invariance within a set $B$, which is characterized as the dual of the reachability problem of set $\mathcal{X} \setminus B$.

For a given policy $\mu$, the time-dependent value function $W_k : \mathcal{X} \rightarrow [0, 1]$, defined as

$$W^\mu_k(x) = \mathbf{E}^\mu \left[ \sum_{j \in [k+1, N_t]} \left( \prod_{i=k+1}^{j-1} 1_{A \setminus K}(x_i) \right) 1_K(x_j) \bigg| x_k = x \right],$$

is the probability that the state trajectory $\{x_{k+1}, \ldots, x_{N_t}\}$, starting from $x_k$, will reach the target set $K$ within the time horizon $[k, N_t]$, while staying within the safe set $A$. This function allows expressing the reach-avoid probability backward recursively, as the following result elucidates.

**Proposition 3** For a given policy $\mu = (\mu_0, \mu_1, \ldots, \mu_{N_t-1})$, define function $W^\mu_k : \mathcal{X} \rightarrow [0, 1]$ by the backward recursions

$$W^\mu_k(x) = \mathbf{E} \left[ 1_K(x_{k+1}) + 1_{A \setminus K}(x_{k+1})W^\mu_{k+1}(x_{k+1}) \big| x_{k+1} \sim T_x (\cdot | x, \mu_k(x)) \right],$$

for $k = N_t - 1, N_t - 2, \ldots, 0$, and initialized with $W^\mu_{N_t}(x) = 0$. Then for any initial state $x_0 \in \mathcal{X}$, the probabilistic reach-avoid property $r^\mu_{x_0}(K, A)$ can be expressed as

$$r^\mu_{x_0}(K, A) = 1_K(x_0) + 1_{A \setminus K}(x_0)W^\mu_0(x_0).$$
Proof The proof follows (Summers and Lygeros, 2010, Lemma 4), which is based on a scheme in (Abate et al., 2008a) and where the above statement is proven for a value function $V_k(x) = 1_K(x) + 1_{A\setminus K}(x)W_k^*(x)$. □

Notice that, while the probabilistic reach-avoid problem has been formulated above via DP recursions, it hinges on a sum-multiplicative characterization which is non-standard: much of the analytical and computational results in DP are formulated for additive (possibly discounted) cost functions (Bertsekas and Shreve, 1996).

Rather than selecting and fixing a policy $\mu$ as done above, we now focus on the controller synthesis problem, which seeks the Markov policy $\mu^*$ that maximizes the probabilistic reach-avoid property, and which is such that

$$r^*_{x_0}(K, A) = \sup_{\mu \in \mathcal{M}} r^\mu_{x_0}(K, A) = \sup_{\mu \in \mathcal{M}} \mathbb{E}^\mu_{x_0} \left[ \sum_{j \in [0, N_t]} \prod_{i=0}^{j-1} 1_{A\setminus K}(x_i) \mathbb{1}_K(x_j) \right].$$

Let us emphasize that the optimization is over finite-action policies, which are however functions of the continuous state space. The optimal policy can be characterized as follows.

**Proposition 4** Define functions $W_k^* : \mathcal{X} \to [0, 1]$, by the backward recursions

$$W_k^*(x) = \max_{a \in A} \mathbb{E} \left[ 1_K(x_{k+1}) + 1_{A\setminus K}(x_{k+1}) W_{k+1}^*(x_{k+1}) \mid x_{k+1} \sim T_x(\cdot \mid x, a) \right],$$

for $k = N_t - 1, N_t - 2, \ldots, 0$, and initialized by $W_N^*(x) = 0$. Then for any initial state $x_0 \in \mathcal{X}$ the optimal probabilistic reach-avoid property $r^*_{x_0}(K, A)$ can be expressed as

$$r^*_{x_0}(K, A) = 1_K(x_0) + 1_{A\setminus K}(x_0) W_0^*(x_0).$$

Furthermore, $\mu_k^* : \mathcal{X} \to A$ for $k = N_t - 1, N_t - 2, \ldots, 0$, is such that $\forall x \in \mathcal{X}$:

$$\mu_k^*(x) = \arg \max_{a \in A} \left\{ \mathbb{E} \left[ 1_K(x_{k+1}) + 1_{A\setminus K}(x_{k+1}) W_{k+1}^*(x_{k+1}) \mid x_{k+1} \sim T_x(\cdot \mid x, a) \right] \right\},$$

and $\mu^* = (\mu_0^*, \mu_1^*, \ldots, \mu_{N_t-1}^*)$ is the optimal probabilistic reach-avoid Markov policy.

**Proof** See again (Summers and Lygeros, 2010, Theorem 6) and (Abate et al., 2008a). □

For a given time horizon $N_t$, the computation of $r^*_{x_0}(K, A)$, as in Proposition 4, can be seen as the application of $N_t$ mappings. More precisely, let us define a dynamic programming operator $T$ as

$$W_{k+1}^* = TW_k^*,$$

such that for all states $x \in \mathcal{X}$, the function $W_k^* : \mathcal{X} \to [0, 1]$ is defined as

$$W_k^*(x) = (TW_k^*) (x) = \max_{a' \in A} \mathbb{E} \left[ 1_K(x_{k+1}) + 1_{A\setminus K}(x_{k+1}) W_{k+1}^*(x_{k+1}) \mid x_{k+1} \sim T_x(\cdot \mid x, a') \right].$$

The exact value of the optimal probabilistic reach-avoid property can thus be written as the composition of $N_t$ mappings,

$$r^*_{x_0}(K, A) = 1_K(x_0) + 1_{A\setminus K}(x_0) (T^{N_t} W_{N_t}^*) (x_0),$$

with $W_{N_t}^*(x) = 0$. 

6
Corollary 5 (Initial probability distribution) It is often relevant to compute the probability of the reach-avoid problem with reference to an initial distribution, rather than a single state. Given an initial state \( x_0 \sim P_0 \in M(X) \), the optimal probability that traces of the Markov model reach the target set \( K \) within \( N_t \) time steps, while staying in \( A \), is obtained as
\[
r^*_p(K, A) = \int_X 1_K(x_0) + 1_{A \setminus K}(x_0) \left( T^{N_t} W_{N_t} \right)(x_0) P_0(dx_0).
\]

2.3 Computation of the Reach-Avoid Probability

Notice that in general it is not possible to solve the above recursion exactly: in order to determine the backwards iteration at a single point \( x_i \in X \), namely \( W^*_k(x_i) = (TW^*_{k+1})(x_i) \), one should exactly solve (2). The exact solution of (2) however is seldom analytical and can possibly result in computationally expensive procedures. The absence of an analytical representation for \( W^*_k : X \rightarrow [0, 1] \) leads to the use of approximation techniques, which can be categorized in two main families:

1. Numerical approximation techniques, which provide an approximation of the optimal probabilistic reach-avoid problem with actual error bounds. More precisely, suppose that \( r^*_{x_0}(K, A) \) is the quantity that in the end we want to approximately compute. For a given error bound \( \Delta > 0 \) we seek a numerical scheme that obtains an approximation \( \hat{r}^*_{x_0}(K, A) \), which is such that \( |\hat{r}^*_{x_0}(K, A) - r^*_{x_0}(K, A)| \leq \Delta \). This approach is taken in (Abate et al., 2010, 2007; Esmaeil Zadeh Soudjani and Abate, 2013), and the scheme is prone to suffer from the curse of dimensionality, since it approximates a general stochastic system by a Markov chain by partitioning state and action spaces.

2. Probabilistic approximation techniques, which approximate the original problem with probabilistic guarantees. The obtained approximation scheme \( \hat{r}^*_{x_0}(K, A) \) for \( r^*_{x_0}(K, A) \) depends on a finite number of samples or sampled paths of the underlying model. Notice that there exist no such schemes in the literature, while there exist many results with bounds that are known to converge asymptotically as the number of sample grows (Kushner and Dupuis, 2001). For a given error bound \( \Delta > 0 \) and confidence \( 1 - \delta \), the probability that the approximation is not close to the the optimal value can be bounded probabilistically as
\[
P\left\{ |\hat{r}^*_{x_0}(K, A) - r^*_{x_0}(K, A)| > \Delta \right\} \leq \delta \Delta.
\]

In this work, we pursue the approach sketched in the second point by focusing on results from the area of learning, and in particular on algorithms for functional approximations. A learning approach is in particular considered suitable for complex systems, such as general Markov processes, since it replaces model-based evaluations by model-free, sample-based evaluations (Busoniu et al., 2010). In this work we adopt the Fitted Value Iteration scheme (FVI) (Munos andSzepesvari, 2008), which is a learning algorithm that can in particular be used within a finite horizon setting. In practice, probabilistic bounds on approximation techniques can be divided into

1. general model-free and sample-free bounds, which give a-priori guarantees on the achievable accuracy for a finite sample set. Though they can show convergence in probability up to a bias term, their generality can render them conservative when used as a tool to assign an accuracy guarantee.
2. model-based and sample-based bounds: these bounds verify the accuracy of a dynamic programming scheme by using information on the sample distribution, on the Markov model, and on the specific reach-avoid property under study.

In the analysis of the algorithm these bounds can be perceived as complementary. A-priori and sample-free bounds are derived in Section 3 based on model-free/distribution-free notions, whereas model-based and sample-based bounds are given in Section 5.

3. Fitted Value Iteration

In this section we consider a learning algorithm that has been developed to solve (possibly discounted) additive-cost optimal control problems, and adapt it to the reach-avoid optimal control setting. Known as fitted value iteration (FVI) (Munos and Szepesvari, 2008), the algorithm is based on extracting a finite number of samples from the underlying model to numerically approximate the value recursions as in (2), which we argue seldom admits analytical solutions. More precisely, the scheme generalizes the information gathered from the samples to approximate the “exact” optimal value function $W_k^*$ as $\hat{W}_k^*$ in two steps: first by estimating $W_k^*$ over a finite number of states, thereafter fitting $\hat{W}_k^* \in \mathcal{W}$ to the estimation by an analytical function. We define $\mathcal{W}$ to be a strict subset of $B(\mathcal{X};1)$, the class of measurable functions defined over $\mathcal{X}$, lower bounded by 0 and upper bounded by 1.

The FVI algorithm employs samples that are generated by the underlying model. At each iteration ($k = N_t - 1, \ldots, 0$) the data set comprises:

1. $N$ base points $(x_{i,k})_{1 \leq i \leq N}$, each of which is drawn from the distribution $\eta$ over the set $A \setminus K$;
2. $M$ samples for each base point $x_{i,k} \in \mathcal{X}$ and action $a \in \mathcal{A}$, which are drawn from the transition kernel as $x_{i,a,j,k+1} \sim T_x(\cdot | x_{i,k}, a)$. This set of samples is denoted by $(x_{i,a,j,k+1})_{1 \leq j \leq M}$.

Table 1: Sample generation for the FVI algorithm

| Step | Description |
|------|-------------|
| 1.   | Draw $N$ base points $(x_{i,k})_{1 \leq i \leq N}$ from the distribution $\eta$ in $A \setminus K$; |
| 2.   | Draw $M$ samples at each base point $x_{i,k}$ and at each actions $a \in \mathcal{A}$ from the stochastic kernel $T_x$, and denote the set of samples as $(x_{i,a,j,k+1})_{1 \leq j \leq M}$. |

Table 1 summarizes the sample generation. Let us remark that at each iteration $k = N_t - 1, \ldots, 0$, a new set of samples is generated and used. We denote as “sample complexity” the cardinality of the sample generation, $N$ and $M$.

At each (backward) iteration $k = N_t - 1, \ldots, 0$, the algorithm executes two steps in order to approximate the exact recursion in (2):
1. The first step consists of estimating the value of the backward mapping \( \left( T\hat{W}_{k+1}^* \right) (x_k^i) \) at \( N \) base points \( x_k^i \). The recursion \( \left( T\hat{W}_{k+1}^* \right) (x_k^i) \) is estimated by an empirical operator \( \hat{T} \) as follows:

\[
\left( T\hat{W}_{k+1}^* \right) (x_k^i) = \max_{a \in A} \frac{1}{M} \sum_{j=1}^M 1_K(x_{k+1}^{i,a,j}) + 1_{A \setminus K}(x_{k+1}^{i,a,j})\hat{W}_{k+1}^*(x_{k+1}^{i,a,j}).
\]

Here \( x_{k+1}^{i,a,j} \) represent \( M \) independent and identically distributed realizations obtained from \( T_x (\cdot | x_k^i, a) \). Hence, \( j \in [1, M], i \in [1, N], \) and \( a \in A \). For an increasing number of samples \( x_{k+1}^{i,a,j} \), the estimate \( \left( T\hat{W}_{k+1}^* \right) (x_k^i) \) converges to \( \left( T\hat{W}_{k+1}^* \right) (x_k^i) \) with probability 1, by the law of large numbers.

2. In the second step, function \( \hat{W}_k^* \in \mathcal{W} \) is estimated as the solution of

\[
\hat{W}_k^* = \arg \min_{w \in \mathcal{W}} \sum_{i=1}^N |w(x_k^i) - \hat{T}\hat{W}_{k+1}^*(x_k^i)|^p,
\]

where the function class \( \mathcal{W} \) is a subset of the class of measurable functions that are bounded within the interval \([0,1]\). We assume that the argument of the minimum belongs to the function class (this fact will be further discussed below). The base points \( (x_k^i)_{1 \leq i \leq N} \) are independently drawn from a distribution \( \eta \) supported over the set \( A \setminus K \). The power factor \( p \geq 1 \) is a given positive number. Given a function \( w(x) \) and an increasing value of \( N \), the summation in (3) converges to \( \int_{A \setminus K} |w(x) - \hat{T}\hat{W}_{k+1}^*(x)|^p \eta(x)dx \), by the law of large numbers. This leads to the convergence of the argument \( \hat{W}_k^* \) to the optimal fit that minimizes the distance between \( \hat{T}\hat{W}_{k+1}^* \) and the functions \( w \in \mathcal{W} \) with respect to the \( p \)-norm, weighted by a distribution with density \( \eta \) and supported over the set \( A \setminus K \), namely

\[
||\hat{W}_k^* - \hat{T}\hat{W}_{k+1}^*||_{p,\eta} = \left( \int_{A \setminus K} |\hat{W}_k^*(x) - \hat{T}\hat{W}_{k+1}^*(x)|^p \eta(x)dx \right)^{1/p}.
\]

The overall FVI algorithm is summarized in Table 2. The iterations are initialized as \( \hat{W}_{N_0}^* (x) = 0, x \in \mathcal{X} \), and updated over the functions \( \hat{W}_{N_1}^*, \hat{W}_{N_2}^*, \ldots, \hat{W}_1^* \). Finally the value function at \( k = 0 \) is approximated at the initial condition \( x_0 \) with a sample-based integration, similar to (3), using \( M_0 \) independent and identically distributed realizations of \( T_x (\cdot | x_0, a) \) for all \( a \in A \).

**Remark 6 (Initial probability distribution)** In order to approximate the computation of the probability of the reach-avoid problem with reference to an initial distribution (rather than a single state, as discussed in Remark 5), a straightforward adaptation of the last iteration of step (3.) in the algorithm of Table 2 is sufficient.

1. In reality convergence can only be proved for certain function classes, cf. Section 4.
Given an initial condition \( x_0 \in \mathcal{X} \), a safe set \( A \), a target set \( K \), a time horizon \( N_t \), a set of \( N \) base points and of \( M \) samples at each base point (for each iteration \( k \)), a number \( p \) and a distribution \( \eta \), proceed with the following value iteration:

1. Initialize \( \hat{W}^{*}_{N_t}(x) = 0, \forall x \in \mathcal{X} \);
2. For \( k = N_t - 1 \) to 1 do
   (a) Collect samples (cfr. Table 1);
   (b) Estimate \( \hat{T}W^{*}_{k+1} \) as
   \[
   \left( \hat{T}W^{*}_{k+1} \right) (x^i_k) = \max_{a \in A} \frac{1}{M} \sum_{j=1}^{M} \left[ 1_K(x^i_{k+1}) + 1_{A \setminus K}(x^i_{k+1})\hat{W}^{*}_{k+1}(x^i_{k+1}) \right];
   \]
   (c) Find the function that minimizes the empirical \( p \)-norm as
   \[
   \hat{W}^{*}_k = \arg \min_{w \in \mathcal{W}} \sum_{i=1}^{N} \left| w(x^i_k) - \hat{T}W^{*}_{k+1}(x^i_k) \right|^p;
   \]
3. Collect \( M_0 \) samples for the single initial condition \( x_0 \) and for every action \( a \in A \), and estimate \( \hat{W}^{*}_0(x_0) \) as in 2.(b);
4. Return reach-avoid probability
   \[
   \hat{r}^{*}_{x_0}(K, A) = 1_K(x_0) + 1_{A \setminus K}(x_0)\hat{W}^{*}_0(x_0), \forall x_0 \in \mathcal{X}.
   \]

Table 2: Fitted Value Iteration algorithm

Remark 7 (Synthesis of an Approximately Optimal Policy) The FVI algorithm can be extended to include the synthesis of a policy \( a = \hat{\mu}_k(x) \). At every iteration in Table 2, first the policy is estimated at all the base points \( x^i_k \), as the argument of (2)(b). Secondly a classification algorithm is used, which gives an approximately optimal policy \( \hat{\mu}_k : \mathcal{X} \rightarrow A \) for each \( k \).

4. A-Priori Probabilistic Error Bounds

Let us define the accuracy of the FVI algorithm as follows: we say that the FVI algorithm has an accuracy \( \Delta \), with a confidence \( 1 - \delta_\Delta \), if the probability that the error made by the approximate solution \( \hat{r}^{*}_{x_0}(K, A) \) is larger than \( \Delta \), is upper-bounded by \( \delta_\Delta \). More formally, the accuracy is given as follows:

\[
\mathbb{P} \left\{ \left| \hat{r}^{*}_{x_0}(K, A) - r^{*}_{x_0}(K, A) \right| > \Delta \right\} \leq \delta_\Delta.
\]  

Next, we provide an explicit expression for the bound in (5), which is done in two steps:

1. first by computing a bound on the error of a single iteration (Sec. 4.1),
2. then by studying the propagation of the single-step error over the complete horizon of the problem (Sec. 4.2).

Notice that this section provides a-priori bounds which are model- and distribution-free, hence computable before applying the FVI algorithm. Alternatively, a-posteriori error bounds based on an additional sample set are proposed in Section 5.

4.1 Error Bounds on Single Iterations of the FVI Algorithm

(In this subsection the use of the time parameter \( k + 1 \) aligns with the notations previously introduced, cfr. Equation (3) and step 2.(b), as well as 1(i) and 2.(c) in Table 2.)

Let \( \hat{T}_W^{*}_{k+1} : \mathcal{X} \to [0, 1] \) be an unknown map, and consider a function class \( \mathcal{W} \subset B(\mathcal{X}; 1) \). Recall that at each iteration \( k = N_t - 1, N_t - 2, ..., 1 \), the objective of the learning algorithm is to find a function \( w \in \mathcal{W} \) that is close to \( \hat{T}_W^{*}_{k+1} \) with respect to the following weighted, \( p \)-norm:

\[
\|w - \hat{T}_W^{*}_{k+1}\|_{p,\eta} = \left( \int_{\mathcal{X}} |w - \hat{T}_W^{*}_{k+1}|^p \eta(x) dx \right)^{\frac{1}{p}}. \tag{6}
\]

Notice that if \( \hat{T}_W^{*}_{k+1} \not\in \mathcal{W} \), the optimal approximation

\[
\inf_{w \in \mathcal{W}} \|w - \hat{T}_W^{*}_{k+1}\|_{p,\eta}
\]

provides only a lower bound on this error: the presence of this non-zero bias error indicates that the FVI scheme is not asymptotically consistent, namely that the error does not converge to zero for an increasing sample size. Of interest to this work, a general upper bound on \( \inf_{w \in \mathcal{W}} \|w - \hat{T}_W^{*}_{k+1}\|_{p,\eta} \) is derived as

\[
d_{p,\eta}(T_W, \mathcal{W}) = \sup_{g \in \mathcal{W}} \inf_{f \in \mathcal{W}} \|f - T g\|_{p,\eta}. \tag{7}
\]

In (Munos and Szepesvari 2008) the bias \( d_{p,\eta}(T_W, \mathcal{W}) \) is referred to as the inherent Bellman error of the function space \( W \).

As discussed, the FVI algorithm employs empirical estimates, given in (3), of the quantity in (6). Therefore, the single step error hinges both on the inherent Bellman error, and on the deviations caused by using estimates of the recursion step \( \hat{T}_W^{*}_{k+1} \) over the base points (cfr. Section 4.1.1) and of the norm \( \| \cdot \|_{p,\eta} \) as the integral in (6) (cfr. Section 4.1.2). The error contributions depend on the number of samples used \( (N, M) \) and on the capacity of the function class \( \mathcal{W} \) (Section 4.1.2 and Appendix [B]), whereas they do not depend on the distribution \( \eta \), nor on the stochastic state transitions characterizing the model dynamics: as such the bounds are general and “distribution-free”.

In the following subsections, two lemmas are derived, which are necessary to obtain a general upper bound for the single step error. First (Section 4.1.1), the error introduced by using the estimation \( \hat{T}_W^{*}_{k+1} \) of the recursion step is bounded using Hoeffding’s inequality (Hoeffding, 1963). Then in Section 4.1.2 the maximal deviation of the empirical evaluation of the integral in (6) is bounded using methods from Statistical Learning Theory (Vapnik, 1998).
4.1.1 Accuracy of the Estimation of $\hat{T}\hat{W}^*_{k+1}$

Recall that the estimate $\left(\hat{T}\hat{W}^*_{k+1}\right)(x^i_k)$ of the exact recursion $\left(\hat{T}\hat{W}^*_{k+1}\right)(x^i_k)$ uses, for a given state-action pair $(x^i_k, a)$, the individual Monte-Carlo estimates

$$\frac{1}{M} \sum_{j=1}^{M} 1_{K}(x^i_{k+1})^j + 1_{A \setminus K}(x^i_{k+1})\hat{W}^*_k(x^i_{k+1})$$

of the quantity

$$E \left[ 1_{K}(x_{k+1}) + 1_{A \setminus K}(x_{k+1})\hat{W}^*_k(x_{k+1}) \middle| x_{k+1} \sim T_x(\cdot | x^i_k, a) \right].$$

Since the cardinality of the action space $A$ is finite, a bound on the error of the estimates for a given state-action pair can lead to a bound on the error over the states $x^i_k$: we elaborate on this idea next.

The $M$ random quantities for $1 \leq j \leq M$, $1_{K}(x^i_{k+1}) + 1_{A \setminus K}(x^i_{k+1})\hat{W}^*_k(x^i_{k+1})$, are obtained via independent and identically distributed realizations over the closed interval $[0, 1]$. Hoeffding’s inequality ([Hoeffding, 1963]) leads to an upper bound on the deviation of the estimate from the expected value as follows:

$$P \left\{ \left| E \left[ 1_{K}(x_{k+1}) + 1_{A \setminus K}(x_{k+1})\hat{W}^*_k(x_{k+1}) \middle| x_{k+1} \sim T_x(\cdot | x^i_k, a) \right] \right| - \frac{1}{M} \sum_{j=1}^{M} 1_{K}(x^i_{k+1}) + 1_{A \setminus K}(x^i_{k+1})\hat{W}^*_k(x^i_{k+1}) \leq \epsilon_1 \right\} \geq 1 - 2e^{-2M(\epsilon_1)^2},$$

where $\epsilon_1$ is the bound on the error.

We can then provide a lower bound on the probability that the deviation incurred by the quantity $\left(\hat{T}\hat{W}^*_{k+1}\right)(x^i_k)$ is bounded by $\epsilon_1$ via the joint probability of $|A|$ independent events, as follows:

$$P \left\{ \left| \hat{T}\hat{W}^*_{k+1}(x^i_k) - \hat{T}\hat{W}^*_{k+1}(x^i_k) \right| \leq \epsilon_1 \right\}$$

$$\geq \prod_{a \in A} P \left\{ \left| E \left[ 1_{K}(x_{k+1}) + 1_{A \setminus K}(x_{k+1})\hat{W}^*_k(x_{k+1}) \middle| x_{k+1} \sim T_x(\cdot | x^i_k, a) \right] \right| - \frac{1}{M} \sum_{j=1}^{M} 1_{K}(x^i_{k+1}) + 1_{A \setminus K}(x^i_{k+1})\hat{W}^*_k(x^i_{k+1}) \leq \epsilon_1 \right\}.$$  

Let us now extend the above probabilistic bound for the error computed at a single point $x^i_k$ to a bound for the error over all the $N$ base points: we can express this bound via an empirical $p$-norm defined over the base points, as follows:

$$\|\hat{T}\hat{W}^*_{k+1} - \hat{T}\hat{W}^*_{k+1}\|_{p, \bar{\eta}} = \left( \frac{1}{N} \sum_{i=1}^{N} \left| \hat{T}\hat{W}^*_{k+1}(x^i_k) - \hat{T}\hat{W}^*_{k+1}(x^i_k) \right|^p \right)^{1/p}. \quad (8)$$

This leads to the following result.
Lemma 8 For a given error bound $\epsilon_1$ and sample complexity $N$ and $M$, the estimation error can be probabilistically bounded as follows:

$$P\left\{ \| \hat{T}W_{k+1}^* - \hat{T}W_{k+1}^* \|_{p, \hat{\eta}} \leq \epsilon_1 \right\} \geq 1 - \delta_1,$$

where $\delta_1 = 1 - (1 - 2e^{-2M(\epsilon_1)^2})|A|^N$, as long as $0 < 2e^{-2M(\epsilon_1)^2} \leq 1$.

Proof See Appendix A.

Notice that for an increasing number of samples $M$, by Lemma 8 the empirical norm of the error as in (8) is less than any given accuracy $\epsilon_1$ with a probability that increases to 1.

### 4.1.2 Accuracy of the Empirical Norm

Let $\hat{T}W_{k+1}^* : \mathcal{X} \rightarrow [0,1]$ be an unknown function and $\eta$ a probability measure on $\mathcal{X}$, $\eta \in M(\mathcal{X})$. The objective is to find a function $w \in \mathcal{W}$ that is close to $\hat{T}W_{k+1}^*$ with respect to the following expected loss:

$$\inf_{w \in \mathcal{W}} \| w - \hat{T}W_{k+1}^* \|^p_{p, \eta} = \inf_{w \in \mathcal{W}} \int_{\mathcal{X}} |w(x) - \hat{T}W_{k+1}^*(x)|^p \eta(x) dx.$$

This section provides a bound for the error originating from the use of a finite number of samples to evaluate the loss: this empirical loss is defined as

$$\| w - \hat{T}W_{k+1}^* \|^p_{p, \hat{\eta}} = \frac{1}{N} \sum_{i=1}^{N} |w(x_k^i) - \hat{T}W_{k+1}^*(x_k^i)|^p,$$

for a given set of $N$ random variables drawn independently over $A \setminus K$ according to $x_k^i \sim \eta$.

Let us express a probabilistic bound on this error that holds uniformly over all functions $w \in \mathcal{W}$ as follows:

$$P\left\{ \sup_{w \in \mathcal{W}} \left[ \| w - \hat{T}W_{k+1}^* \|^p_{p, \hat{\eta}} - \| w - \hat{T}W_{k+1}^* \|^p_{p, \hat{\eta}} \right] \geq \epsilon_2^p \right\} \leq \delta_2.$$

Observe that since the expected and the empirical losses can be reformulated respectively as the mean and the empirical mean of a loss function, defined informally as $f(x) = |w(x) - \hat{T}W_{k+1}^*(x)|^p$, the above problem can be framed as the uniform convergence of a standard learning problem (Pollard, 1984; Haussler, 1992, 1995; Vapnik, 1998). Resorting to the related literature, results on bounds for the error probability of the regression of real-valued functions employ capacity concepts (including Rademacher averages, covering numbers, and pseudo dimensions) of a function class (Pollard, 1990; Hoeffding, 1963; Bartlett et al., 2005). We focus on pseudo dimensions to deal with the capacity (or the complexity) of the function class. By the theory developed in (Pollard, 1984, 1990) and (Haussler, 1992), we obtain the following uniform convergence bound.

Lemma 9 Let $\mathcal{W}$ be a set of finitely parameterized, measurable functions on $\mathcal{X}$ taking values in the interval $[0,1]$ with pseudo dimension $\dim_p(\mathcal{W}) = d < \infty$. Let $(x_k^i)_{1 \leq i \leq N}$ be generated...
by $N$ independent draws according to any distribution $\eta$ on $A \setminus K$ and let $p \geq 1$. Then for any $\epsilon_2 > 0$ we have that

$$P \left\{ \sup_{w \in W} \left| \left| w - \hat{T}\hat{W}_{k+1}^* \right| \right|_{p,\eta}^p - \left| \left| w - TW_{k+1}^* \right| \right|_{p,\eta}^p \geq \epsilon_2^p \right\} \leq 4e(d + 1) \left( \frac{32e}{\epsilon_2} \right)^d e^{-N\epsilon_2^2}. \quad (10)$$

**Proof** See Appendix C, which also elaborates on the characterization and computability of the pseudo dimension of a function class. $lacksquare$

### 4.1.3 Global Accuracy of Single Iterations

The error bounds introduced in Lemmas 8 and 9, together with the inherent Bellman error, yield an upper bound on the error introduced at each iteration, which is recapitulated in the following statement. Since the bounds are independent of the distributions $\eta$ and $T_\pi(\cdot | x, a)$, they are in fact “distribution-free” bounds (Bartlett et al., 2005).

**Theorem 10** Consider a reach-avoid property defined over a general stochastic system with continuous state space $X$ and finite action space $A$. Let $A \subset X$ and $K \subset X$ be Borel measurable sets and fix $p \geq 1$, the distribution $\eta \in M(A \setminus K)$ and $W \subset B(X; 1)$. Pick any $\hat{W}_{k+1}^* \in B(X^*; 1)$ and let $\hat{T}\hat{W}_{k+1}^*$ and $W_k^*$ be calculated using (3) and (4). For given upper bounds on

$$P \left\{ \left| \left| \hat{T}\hat{W}_{k+1}^* - \hat{T}\hat{W}_{k+1}^* \right| \right|_{p,\eta}^p > \epsilon_1 \right\} \leq \delta_1, \quad (11)$$

and

$$P \left\{ \sup_{w \in W} \left| \left| w - \hat{T}\hat{W}_{k+1}^* \right| \right|_{p,\eta}^p - \left| \left| w - TW_{k+1}^* \right| \right|_{p,\eta}^p \geq \epsilon_2^p \right\} \leq \delta_2, \quad (12)$$

the bound on a single step update is as follows:

$$P \left\{ \left| \left| \hat{W}_k^* - TW_{k+1}^* \right| \right|_{p,\eta}^p > d_{p,\eta}(\hat{T}\hat{W}_{k+1}^*, W) + \epsilon \right\} \leq \delta_1 + \delta_2, \quad (13)$$

where the optimal approximation is given as a biasing term defined as $d_{p,\eta}(\hat{T}\hat{W}_{k+1}^*, W) = \inf_{w \in W} \left| \left| w - TW_{k+1}^* \right| \right|_{p,\eta}$, and the error $\epsilon$ is given as $\epsilon = 2\epsilon_1 + 2\epsilon_2$.

**Proof** See Appendix C. $lacksquare$

Note that the statement assumes that $W_k^*$ is the unique solution to the optimization problem in (4). This strict assumption can be weakened by adding a tolerance term to the theorem. The quantity $d_{p,\eta}(\hat{T}\hat{W}_{k+1}^*, W)$ admits the inherent Bellman error $d_{p,\eta}(TW, W)$ in (7) as a general upper bound.

### 4.2 Error Propagation over the Whole Horizon and Global Error Bounds of the FVI Algorithm

We express a global bound on the accuracy of the FVI algorithm $|\hat{r}^*_x(K, A) - r^*_x(K, A)|$ by using the probabilistic bounds derived in Section 4.1 on the errors introduced by the approximate one-step mappings $|\left( \hat{T}\hat{W}_1^* \right)(x_0) - \left( \hat{T}\hat{W}_1^* \right)(x_0)|$, as well as $||\hat{W}_1^* - TW_2^*||_{p,\eta}$.
\[ \| \hat{W}_k^* - T \hat{W}_k^* \|_{p,\eta}, \ldots, \| \hat{W}_{N_t-1}^* - T \hat{W}_{N_t}^* \|_{p,\eta}, \text{and by propagating the error introduced by each single iteration to the successive value iterations, as done in the next statement. More precisely, in the following lemma we show that the deviation of the approximate value function } \hat{W}_k^* \text{ from the optimal value function } T^{N_t-k} W_{N_t}^* \text{ can be expressed as a function of this deviation at step } k+1, \text{ plus the approximation error introduced at the } (N_t - k) \text{th iteration (which has been bounded above). Recall that the optimal value functions can be written as } W_k^* = T^{N_t-k} W_{N_t}^* \text{ and that by definition } W_{N_t}^* = \hat{W}_{N_t}^*. \]

**Lemma 11** Let \( \eta \) be the density of a probability distribution with support on \( A \setminus K \) and \( t_x \) be the density function of the stochastic kernel \( T_x \). Then
\[
\| \hat{W}_k^* - T^{N_t-k} \hat{W}_{N_t}^* \|_{p,\eta} \leq \| \hat{W}_k^* - T \hat{W}_{k+1}^* \|_{p,\eta} + B^{\frac{1}{p}} \| \hat{W}_{k+1}^* - T^{N_t-(k+1)} \hat{W}_{N_t}^* \|_{p,\eta},
\]
for \( k = 0, 1, \ldots, N_t - 1 \), and where \( B \) is defined as
\[
B = \sup_{x_k+1 \in A \setminus K} \int_{A \setminus K} \max_{a \in A} t_x(x_{k+1} \mid x_k, a) \frac{\eta(x_k)}{\eta(x_{k+1})} dx_k.
\]  

**Proof** See Appendix D.

Putting all the pieces together, the following theorem provides an expression for the global FVI error bound as the accumulation of the errors from the single iterations over the whole time horizon.

**Theorem 12** Consider a reach-avoid problem defined on a Markov process with a continuous state space \( X \) and a finite action space \( A \). The optimal reach-avoid probability \( r^*(x_0, K, A) \) for a given target set \( K \), safe set \( A \), initial state \( x_0 \), and time horizon \( N_t \), is approximated by the quantity \( \hat{r}^*(x_0, K, A) \) obtained with the FVI Algorithm in Table 2, which has an accuracy of \( \Delta \) and a confidence \( \delta \), as stated in (5), if the following holds:
\[
P \left\{ \sum_{k=1}^{N_t-1} B_0^{k-1} \| \hat{W}_k^* - T \hat{W}_{k+1}^* \|_{p,\eta} + | \hat{W}_0^*(x_0) - T \hat{W}_1^*(x_0) | > \Delta \right\} \leq \delta, \tag{15}
\]
where \( B_0 \) is given in (14), \( B_0 \) is defined as
\[
B_0 = \sup_{x_1 \in A \setminus K} \max_{a \in A} \frac{t_x(x_1 \mid x_0, a)}{\eta(x_1)},
\]  

and where \( \eta \) is the density of a probability distribution supported on \( A \setminus K \) and \( t_x \) is the density of the stochastic kernel \( T_x \) of the given Markov process.

**Proof** See Appendix D.

**Corollary 13** (Initial state distribution) **Theorem 12** can be extended to the case of an initial condition taken from a given distribution. Consider the density \( p_0 \) of a given initial state distribution \( P_0 \in M(X) \). Then, it is possible to show that:
1. Equation (15) can be extended to

\[
P \left\{ B_0 \sum_{k=1}^{N_t-1} B^{k-1}_\rho \| \hat{W}_k^* - \hat{T}W_{k+1}^* \|_{p,\eta} + \int_{A\setminus K} |\hat{W}_0^*(x_0) - \hat{T}W_1^*(x_0)| p_0(x_0) dx_0 > \Delta \right\} \leq \delta_\Delta.
\]

2. and the expression for \( B_0 \) can be generalized as

\[
B_0 = \sup_{x_1 \in A \setminus K} \int_{A \setminus K} \frac{\max_{a \in A} t_x(x_1 | x_0, a) p_0(x_0)}{\eta(x_1)} dx_0.
\]

**Remark 14** Notice that the scaling factor \( B \) has a significant influence on the error propagation. It is related to the notion of concentrability of the future-state distribution (Manos and Szepesvari, 2008; Farahmand et al., 2010). If \( B > 1 \) the error of the algorithm will increase exponentially with the time horizon, whereas if \( B < 1 \) the accuracy of the algorithm will depend mostly on the errors in the last few iterations (backwards in time).

The scaling factor \( B \) expresses the maximal concentration of the dynamics (relative to \( \eta \)) over the relevant set \( A \setminus K \) after one transition starting from distribution \( \eta \). It is known that for an autonomous, finite-state Markov process, \( B \leq 1 \) if \( \eta \) differs from the stationary distribution of the Markov process over \( X \) only by a constant factor for all \( A \setminus K \).

The case study discussed in Section 6 displays a choice of a density function \( \eta \) leading to bounded values for \( B \) and \( B_0 \), respectively. The relation between the scaling factor \( B \) and the model dynamics for the considered Markov process is also analyzed.

**4.2.1 Discussion on the Global Error Bounds**

Suppose that we require that the error in the estimation of \( r_{x_0}^*(K, A) \), with a confidence at least equal to \( \alpha \), is less than \( \Delta \), as per (5). Let us assume that there exist positive values \( \epsilon_0, \epsilon_1, \epsilon_2 \) such that the approximation error \( \Delta \) can be split up into individual bounds based on Theorem 10 and 12 as follows:

\[
\Delta = B_0 \sum_{k=1}^{N_t-1} B^{k-1}_\rho \frac{d p_\eta (T W, W)}{\eta (x_0)} - \quad \text{inherent Bellman error (Section 4.1.1)}
+ 2B_0 \sum_{k=1}^{N_t-1} B^{k-1}_\rho (\epsilon_1) - \quad \text{error on the estimation of } \hat{T}W_{k+1}^* \text{ (Section 4.1.1)}
+ 2B_0 \sum_{k=1}^{N_t-1} B^{k-1}_\rho (\epsilon_2) - \quad \text{error on the empirical norm (Section 4.1.2)}
+ \epsilon_0 - \quad \text{error related to } |\hat{W}_0^*(x_0) - \hat{T}W_1^*(x_0)|
\]

We compute a lower bound on the confidence that the approximation error is bounded by \( \Delta \) by upper bounding the complementary event. Using a union bounding argument on the probability of invalidating any of the bounding terms in the above equation, we obtain that the probability that the overall error is larger than \( \Delta \) is upper bounded with \( \delta_\Delta \), as

\[
\delta_\Delta = 0 + (N_t - 1) \left( 1 - (1 - 2e^{-2M(\epsilon_1)^2} |A| N) \right) - \quad \text{inherent Bellman error (Section 4.1.1)}
+ (N_t - 1) \left( 4e(d + 1) \left( \frac{32e}{\epsilon_2} \right)^d e^{\frac{N_t^2}{168}} \right) - \quad \text{Lemma 8 with sample complexity } M \text{ and } N
+ 1 - (1 - 2e^{-2M_0(\epsilon_0)^2} |A|) - \quad \text{Lemma 8 with } \dim_p (W) = d < \infty
\]

\[
(18)
\]

16
as long as $0 < 2e^{-2M(\epsilon_1)^2} \leq 1$ and $0 < 2e^{-2M_0(\epsilon_0)^2} \leq 1$. We can observe that it is possible to find finite values for $N, M, M_0$ such that $1 - \alpha > \delta_2$. Moreover for each choice of positive $\epsilon_0, \epsilon_1, \epsilon_2 > 0$ and confidence $0 \leq \alpha < 1$, the necessary number of samples can be upper bounded by polynomials in $\frac{1}{\epsilon_0}, \frac{1}{\epsilon_1}, \frac{1}{\epsilon_2}$ and $\frac{1}{1-\alpha}$, as follows:

$$
\begin{align*}
N &= \left\lfloor 128 \left( \ln(4e(d + 1)) + d \ln(32e) \right) \left( \frac{1}{\epsilon_2} \right)^{2p} \\
&\quad + 128dp \left( \frac{1}{\epsilon_2} \right)^{2p} \ln \left( \frac{1}{\epsilon_2} \right) + 128 \left( \frac{1}{\epsilon_2} \right)^{2p} \ln \left( \frac{1}{\epsilon_2} \right) \right\rfloor,
M &= \left\lfloor \frac{1}{2} \left( \frac{1}{\epsilon_1} \right)^2 \left( \ln(2|A|) + \ln(\frac{1}{\delta_1}) + \ln(N) \right) \right\rfloor,
M_0 &= \left\lfloor \frac{1}{2} \left( \frac{1}{\epsilon_0} \right)^2 \left( \ln(2|A|) + \ln(\frac{1}{\delta_0}) \right) \right\rfloor,
\end{align*}
$$

with positive parameters $\delta_0, \delta_1, \delta_2 > 0$ such that $1 - \alpha = \delta_0 + (N_t - 1)\delta_1 + (N_t - 1)\delta_2$ (see Appendix F for this derivation).

Note that the above accuracy does not depend on the dimensionality $n$ of the state space. Therefore the accuracy for models with higher state space dimension will directly depend on the complexity of the function class employed to approximate given value functions. This is unlike standard grid-based numerical approximation techniques, such as that proposed in (Esmaeil Zadeh Soudjani and Abate, 2013), which are known to break down over models with large state-space dimensionality. In grid-based approximation techniques, the accuracy obtained is proportional to the largest diameter of the grid, and the number of cells in the grid grows exponentially with the state space dimension $n$. Furthermore, in order to decrease the error by a factor $a$, the grid diameters has to be decreased proportionally, and the number of cells (and the related memory usage) increases as $a^n$.

Let us add a few comments on the dependency of the accuracy from several design variables of the FVI algorithm. Firstly, the choice of function class affects both the inherent Bellman error and the pseudo dimension: while the former gives a measure of how well the function class $W$ can represent the value functions $W_k^*$, the latter is directly related to the complexity of the function class. The objective is to obtain a low complexity function class that is capable to accurately fit the given value functions. A good accuracy can be hard to attain when a bad choice of the function class leads to both a large bias (due to the inherent Bellman error) and to a large number of samples. Secondly, the sample distribution $\eta$ defines, together with the state transitions, the scaling factors $B$ and $B_0$. In order to minimize the error propagation caused by $B$, the distribution $\eta$ should be “aligned” with the model dynamics characterized by the density of the transition kernel. Finally, the parameters $N, M, M_0$ follow from the required accuracy demands, which are reformulated as polynomial functions depending on $\epsilon_0, \epsilon_1, \epsilon_2$ and $\delta_0, \delta_1, \delta_2$ as in (19).

With regards to the single-step errors, Lemmas 8 and 9 determine a bound uniformly over the whole function class and for any possible probability distribution. The used distribution-free notions lead to conservative bounds (Bartlett et al., 2005), which then result in a large set of required samples. Notice however that the construction allows to compute the bounds a-priori, before any sample is drawn from the system.

The formal probabilistic bounds on the error made by the approximation algorithm show that the algorithm converges in probability to the best approximation for an increasing cardinality of the samples.
5. Sample-Based Error Bounds

In this section, a probabilistic bound on the error of the approximated reach-avoid probability is developed according to a model- and sample-based philosophy. This bound can be computed after the reach-avoid probability has been obtained via dynamic programming as time-dependent, approximate value functions \( \hat{\hat{W}}^*_k \) for \( 0 \leq k \leq N_t - 1 \). The obtained bounds are not only sample dependent but also distribution dependent, since knowledge of the transition kernel is necessary to compute scaling factors such as (14). Throughout the section it is assumed that the used samples are not correlated with \( \hat{\hat{W}}^*_k \), in other words if the estimated value functions \( \hat{\hat{W}}^*_k \) are a result of a sampled-based optimization, then the samples used for the bounds in this section are drawn anew and independently.

The probabilistic bound on the accuracy \( P \{ | \hat{\hat{r}}^*_x(K,A) - \hat{r}^*_x(K,A) | > \Delta \} \leq \delta_\Delta \), as given in (5), and computed now in a sample-based manner, includes an empirical estimate of the quantity \( | \hat{\hat{r}}^*_x(K,A) - \hat{r}^*_x(K,A) | \). This sample-based estimate is computed as follows:

a. collect samples \( (x_i)_{1 \leq i \leq \tilde{N}} \) according to (1) in Table 1 (with \( N = \tilde{N} \)), and subsequently use (2) with \( M = \tilde{M} \) to draw both \((y^i_{1,a,j})_{1 \leq i \leq \tilde{M}}\) and \((y^i_{2,a,j})_{1 \leq j \leq \tilde{M}}\);

b. estimate the single step error (20) for each \( k \);

c. estimate the bias (22) for each \( k \);

d. compute the multi-step error as a propagation and a composition of the estimates in (b.) and (c.).

The single step error is estimated as

\[
\| \hat{\hat{W}}^*_k - \hat{T}_k \hat{\hat{W}}^*_{k+1} \|_{1,\tilde{\eta}} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \left| \hat{\hat{W}}^*_k(x^i) - \max_{a \in A} \hat{T}^a_1 \hat{\hat{W}}^*_{k+1}(x^i) \right| \quad (20)
\]

for all \( 1 \leq k \leq N_t - 1 \). The term on the right is the empirical 1-norm and can be written as a 1-norm with weighting \( \tilde{\eta} \), which is the empirical distribution of \( \eta \) resulting from \( (x_i)_{1 \leq i \leq \tilde{N}} \).

Let the operator \( \hat{T}^a_\alpha \) be, for \( \alpha = 1, 2 \),

\[
\hat{T}^a_\alpha \hat{\hat{W}}^*_{k+1}(x^i) = \frac{1}{\tilde{M}} \sum_{j=1}^{\tilde{M}} 1_{K}(y^i_{\alpha,j}) + 1_{A \setminus K}(y^i_{\alpha,j}) \hat{\hat{W}}^*_{k+1}(y^i_{\alpha,j}). \quad (21)
\]

The estimate in (20) is biased due to the maximization over the action space, therefore as a second step we estimate a bound on this bias. The combination of the two sample sets \( \{(y^i_{1,a,j})_{1 \leq j \leq \tilde{M}}\} \) and \( \{(y^i_{2,a,j})_{1 \leq j \leq \tilde{M}}\} \), for each \( x^i \) and \( a \), allows us to estimate this bias as

\[
\left\| \max_{\alpha \in A} \left| \hat{T}^a_1 \hat{\hat{W}}^*_{k+1} - \hat{T}^a_2 \hat{\hat{W}}^*_{k+1} \right| \right\|_{1,\tilde{\eta}}, \quad (22)
\]

for \( 1 \leq k \leq N_t - 1 \).

In the following theorem, an expression for the bound on \( P \{ | \hat{\hat{r}}^*_x(K,A) - \hat{r}^*_x(K,A) | > \Delta \} \leq \delta_\Delta \) is derived, employing the error propagation technique first used in Section 4.2, the estimates of the single step error above, and the bias (22) in combination with Hoeffding’s inequality \( \text{Hoeffding} [1963] \).
Theorem 15 Consider a reach-avoid problem defined on a Markov process with a continuous state space $\mathcal{X}$ and a finite action space $A$. The optimal reach-avoid probability $r^*_{x_0}(K, A)$ for a given target set $K$, safe set $A$, initial state $x_0$, and time horizon $N_t$, is approximated by the quantity $\tilde{r}^*_{x_0}(K, A)$ obtained with the FVI Algorithm in Table 2, which has an accuracy of $\Delta$ and a confidence $\delta_\Delta$, as stated in (25), if the following holds:

$$\Delta = B_0 \sum_{k=1}^{N_1-1} B^{k-1} \left( \left\| \tilde{W}^*_k - \tilde{\bar{W}}^*_k \right\|_{1, \bar{\eta}} + \left\| \max_{a \in A} \tilde{T}^*_1 \tilde{W}^*_{k+1} - \tilde{T}^*_2 \tilde{W}^*_{k+1} \right\|_{1, \bar{\eta}} \right) + B_0 \epsilon + \epsilon_0,$$

$$\delta_\Delta = e^{-2 \frac{\delta_\Delta^2}{\epsilon^2}} - \delta_0,$$

with $L = 2 \sum_{k=1}^{N_1-1} B^{k-1}$, and sample sizes $\tilde{M}$ and $\tilde{N}$ according to the sample sets drawn according to the distribution $\eta$. Equation (23a) includes the estimated error as a combination of (20) and (22). The scaling factors $B$ and $B_0$ are computed as in (14) and (16) for the same sampling distribution $\eta$. The factors $\delta_0$ and $\epsilon_0$ are computed as in Lemma 8 with $M = M_0$ and $N = 1$.

Proof See Appendix G.

The accuracy $\Delta$ depends on two terms, the propagation of the estimated single-step and bias errors over the time horizon up to $k = 1$, and the estimation errors $B_0 \epsilon + \epsilon_0$ for $k = 0$, related to the confidence $\delta_\Delta$.

Suppose a close-to-optimal policy is given, for example a policy as detailed in Remark 7 and computed from the series of estimated value functions $\tilde{W}^*_k$. Then we know that $r^*_{x_0}(K, A) \geq \tilde{r}^*_{x_0}(K, A)$, therefore a lower bound on the value of $r^*_{x_0}(K, A)$ is also a lower bound on $r^*_{x_0}(K, A)$. Note that for a policy $\mu$, the closed-loop Markov process is time dependent. This allows us to estimate $r^\mu_{x_0}(K, A)$ directly from traces of this autonomous Markov process. The deviation of this empirical mean can be bounded probabilistically using Hoeffding’s inequality. Additionally upper bound on the deviation $|\tilde{r}^*_{x_0}(K, A) - r^\mu_{x_0}(K, A)|$ can be computed. The combination of the bound in Theorem 15 and of the bound on $|\tilde{r}^*_{x_0}(K, A) - r^\mu_{x_0}(K, A)|$ provide a bound on the performance deviation of $r^\mu_{x_0}(K, A)$: the triangle inequality leads to $|r^*_{x_0}(K, A) - \tilde{r}^*_{x_0}(K, A)| \leq |r^*_{x_0}(K, A) - r^\mu_{x_0}(K, A)| + |\tilde{r}^*_{x_0}(K, A) - r^\mu_{x_0}(K, A)|$.

In comparison to the a-priori bound derived in Section 4, the sample-based bounds do not depend on the inherent Bellman error and can be shown to be less conservative in general. Moreover, they provide insight into the accuracy of the iterations steps. However, they give no information about the expected convergence of the algorithm, and they can only be computed after a run of the algorithm. Similarly to the a-priori bounds, they do not depend on the dimensionality of the state space and can be expected to scale better than the bounds used for grid-based approaches such as (Esmaeil Zadeh Soudjani and Abate, 2013). The sample based bounds are elucidated further with the case study in the following section.
6. Case Study and Numerical Experiments

We consider a case study from the literature (Fehnker and Ivančić, 2004), where the goal is to maximize the probability that the temperature of two interconnected rooms, while staying within a comfortable range, reaches a target temperature range within a given finite time horizon. The temperature can be affected by using local heating systems. In this study the outcome is attained by applying the Fitted Value Iteration scheme. A reach-avoid problem is set up by selecting as the safe set \( A = [17.5 \ 22]^2 \), as the target set \( K = [19.25 \ 20.25]^2 \), and a fixed time horizon \( N_t = 10 \). The case study was implemented in Matlab R2013b on a notebook with 2.6 GHz Intel Core i5 and 16 GB of RAM.

6.1 Model

The dynamics of the temperature in the two rooms is described by a Markov model, with the temperature of the rooms making up the state space \( X = \mathbb{R}^2 \), and where the possible configurations of the two heaters \( \{\text{OFF, ON}\} = \{0, 1\} \) form the finite action space \( A \). Hence \( A = \{0, 1\} \times \{0, 1\} \), and as an example the action related to the first heater in the ON mode and the second in the OFF one is given as \( a = [1 \ 0]^T \in A \). The dynamics at discrete time \( k \), driven by a sequence \( n_k \) of i.i.d Gaussian random variables, is characterized by the following stochastic difference equation:

\[
x_{k+1} = Ax_k + Ba + C + n_k,
\]

where

\[
A = \begin{bmatrix} 1 - b_1 - a_{1,2} & a_{1,2} \\ a_{2,1} & 1 - b_2 - a_{2,1} \end{bmatrix}, \quad C = \begin{bmatrix} b_1 x_a \\ b_2 x_a \end{bmatrix}, \quad \text{and} \quad B = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix},
\]

and with the following parameters (instantiated in Table 3):

- \( x_a \) is the ambient temperature (assumed to be constant),
- \( b_i \geq 0 \) is a constant for the average heat loss rate of room \( i \) to the environment,
- \( a_{ij} \geq 0 \) is a constant for the average heat exchange rate of room \( i \) to room \( j \neq i \),
- \( c_i \geq 0 \) is a constant for the rate of heat supplied by the heater in room \( i \).

The noise process \( n_k \) is a realization of zero-mean Gaussian random variables with covariance \( \nu^2 I_{2 \times 2} \) (2-dimensional identity matrix \( I_{2 \times 2} \)) and \( \nu = 0.5 \). Let \( \mathcal{N}(\cdot | \mu, \Sigma) \) be a 2-dimensional multivariate normal distribution over \( (X, B(X)) \) with mean \( \mu \) and covariance matrix \( \Sigma \), then the stochastic kernel \( T_x \) is given as

\[
T_x (\cdot | x, a) = \mathcal{N} (\cdot | Ax + Ba + C, \nu^2 I_{2 \times 2})
\]

| \( b_1 \) | \( c_1 \) | \( x_a \) | \( a_{ij} \) |
|---|---|---|---|
| 0.0375 | 0.65 | 6 | 0.0625 |

Table 3: Parameters of the 2-room heating case study, taken from Abate et al. (2008b).
and gives the probability distribution of the stochastic transitions in (24). The stochastic kernel (25) admits the probability density

$$
t_x(y \mid x, a) = \frac{1}{\sqrt{|\Sigma|(2\pi)^2}}e^{(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu))}, \quad (26)
$$

where $|\cdot|$ denotes the determinant of a matrix, and as before the covariance matrix equals $\Sigma = \nu^2 I_{2 \times 2}$ and the mean value is $\mu = Ax + Ba + C$.

### 6.2 Application of the Fitted Value Iteration Algorithm

The FVI scheme is implemented as in Algorithm 2 and solves the reach-avoid problem approximatively. We obtain an approximation of $\tilde{r}_{\eta}(K, A) = T^{10}\tilde{W}_0^*(x_0)$ by $\tilde{W}_0^*(x_0)$, while using the auxiliary functions $\tilde{W}_1^*, \ldots, \tilde{W}_9^*$ to approximate $T\tilde{W}_1^*, \ldots, T\tilde{W}_9^*$ in the FVI scheme – equivalently, function $\tilde{W}_k^*$ approximates $T^{N_t-k}\tilde{W}_9^*$. For a given temperature $x_k$ at time instant $k$, the function $\tilde{W}_k^*(x_k)$ gives the approximate probability that the consecutive temperature values $x_{k+1}, \ldots, x_{N_t}$ will reach the temperature range $[19.25, 20.25]^2$ within $N_t - k$ time steps, while staying inside the safe set $[17.5, 22]^2$.

In order to apply the FVI algorithm, we first select a distribution $\eta$ to sample from, then select a function class $W$ and a value for $p \geq 1$ to solve (4). A uniform distribution $\eta$ over $A \backslash K$ is selected. This has a scaling factor $B$ less than $\frac{1}{|A| |A|}$ (see Appendix H for the derivation).

To approximate the value functions, we use a function class $W$ composed of Gaussian radial basis function (RBF) neural networks with 50 RBFs with a uniform width of 0.7. The neural network toolbox of Matlab is used to solve the regression problem in (4) as a least-square problem (with $p = 2$). The neural network with a single layer of hidden units of Gaussian type radial basis functions is proved to be a universal approximator for real-valued functions (Hartman et al., 1990). Furthermore the pseudo dimension of an artificial neural network with $W$ free parameters and $k$ hidden nodes has been upper bounded by $O(W^2k^2)$ (Karpinski and Macintyre, 1997; Anthony and Bartlett, 1999). This means that for any desired precision the required number of samples is bounded by a polynomial in the number of hidden nodes.

The following quantities are obtained for the sample complexities: $N = 600$, $M = 10^3$, $M_0 = 10^3$. The approximate value functions for $\tilde{W}_9^*$, $\tilde{W}_5^*$, and $\tilde{W}_1^*$ are displayed in Figure 1. On the top plots, a point on the state space is associated with a probability for the reach-avoid property over the given time horizon. At the bottom, the contour plots (level sets) characterize the set of points that verify the reach-avoid property with a probability at least equal to the given level.

Table 4 provides, for several initial conditions $x_0$, the optimal action at the initial time and the approximated solution $(\tilde{r}_{x_0}^*(K, A))$, cf. Figure 1 of the reach-avoid problem. Figure 2 displays a suboptimal policy $\tilde{\mu}^*$ that is obtained via the FVI algorithm as discussed in Remark 7, by employing the tree classification method ClassificationTree.fit of Matlab. The selected policy corresponds to heating actions for the lower temperature regions of the rooms. Observe that policy $\tilde{\mu}_9^*$ for $k = 9$ is not accurate over the flat regions of $\tilde{W}_9^*$ (corresponding to the blue spots in the left side of Figure 2 - left plot), which are far away from the reach set $K$. Since the average heat loss rate of room 1 is the highest, we
expect that the heating should be turned on relatively longer. Figure 2 confirms this, i.e. the red (ON,ON) region is not square-shaped as the heaters stay ON for higher temperatures in room 1 than in room 2.

6.3 Performance of the Value Iteration

We are interested in the performance of the FVI algorithm and in analyzing how the computed accuracy deteriorates over the iterations from $N_t - 1$ to 1. Note that the last iteration is of little interest, since it does not include the fitting step. The accuracy is computed using the model-based and sample-based bounds of Section 5. The sample-based esti-
Table 4: Approximate solutions $\tilde{r}^*_{x_0}(K,A)$ are given in the table for several initial conditions $x_0$, together with the related sub-optimal action at the initial time.

![Policy plots](image)

Figure 2: The policy $\hat{\mu}^*$ for $k = 9, 5, 1$ obtained from the same computations as in Figure 1. The action (ON,ON) is labeled in red, (OFF, ON) is orange, (ON,OFF) is yellow, and (OFF,OFF) is blue.

The approximate solutions $\tilde{r}^*_{x_0}(K,A)$ are given in the table for several initial conditions $x_0$, together with the related sub-optimal action at the initial time.

| $x_0$ | $\tilde{r}^*_{x_0}(K,A)$ | $\tilde{r}^*_{x_0}(K,A)$ | $\tilde{r}^*_{x_0}(K,A)$ | $\tilde{r}^*_{x_0}(K,A)$ |
|-------|--------------------------|--------------------------|--------------------------|--------------------------|
| $x_0$ | $[19, 19]^T$             | $[20.5, 19]^T$           | $[19, 20.5]^T$           | $[20.5, 20.5]^T$         |
| $a$   | (ON,ON)                  | (OFF,ON)                 | (ON,OFF)                 | (OFF,OFF)                |
|       | 0.8808                   | 0.9454                   | 0.9206                   | 0.9557                   |
| $x_0$ | $[18, 18]^T$             | $[21.5, 18]^T$           | $[18, 21.5]^T$           | $[21.5, 21.5]^T$         |
| $a$   | (ON,ON)                  | (ON,OFF)                 | (OFF,ON)                 | (OFF,OFF)                |
|       | 0.5204                   | 0.7635                   | 0.8596                   | 0.8312                   |

The estimates of the single step error (20), namely $\parallel \hat{W}^*_{k} - \hat{T}\hat{W}^*_{k+1} \parallel_{1,\hat{\eta}}$; and of the bias (22), namely $\max_{a \in A} \parallel \hat{T}\hat{W}^*_{k+1} - \hat{T}_{a\mu} \hat{W}^*_{k+1} \parallel_{1,\hat{\eta}}$, are provided in Figure 3. Observe that the values of both (20) and (22) fall in the interval between $3 \times 10^{-3}$ and $5 \times 10^{-3}$. The bias estimate (22) appears distributed all over this interval, whereas there is a noticeable trend in the plot of (20), which suggests that the first iterations can be fitted more easily than the later ones. In Figure 4, the accuracy of the FVI algorithm propagated over the iterations is given, starting from the first iteration $\parallel \hat{W}^*_{0} - T \hat{W}^*_{N_t} \parallel_{1,\hat{\eta}}$ until the last iteration $\parallel \hat{W}^*_{1} - T^{N_t} W^*_{N_t} \parallel_{1,\hat{\eta}}$. This accuracy is computed using Theorem 15. The estimates in Figure 3 are used to compute the estimate of the accuracy $\parallel \hat{W}^*_{k} - T W^*_{N_t} \parallel_{1,\hat{\eta}}$ and the accuracy $\Delta$ for a given $\delta_{\Delta}$. For each iteration step, it can be observed in Figure 4 that the error caused by estimating the dynamic programming operator $\hat{T}$ and by fitting a function is relatively small ($< 10^{-2}$). However, when considering the error propagated over the whole horizon, the error grows exponentially. This is caused by the scaling factor $B$, which has been computed numerically and amounts to 3.07. As expected, the accuracy of the algorithm depends strongly on $B$. Therefore good accuracy of the algorithm can only be achieved if the number of iterations $N_t$ is limited or if the scaling factor is reasonably close to or smaller than 1.
An independent set of samples of size $\tilde{N} = 4 \cdot 10^3$ and $\tilde{M} = 10^4$ has been used.

7. Conclusions and Future Work

This article has investigated the performance of a sample-based approximation scheme for the synthesis of optimal controllers maximizing the probability associated to a property of interest, known as the “reach-avoid” specification. The reach-avoid problem focuses on the maximization of the likelihood that any finite-horizon trajectory of the model enters a given goal set, while avoiding a fixed set of undesired states. The approximate computational scheme is based on the Fitted Value Iteration algorithm, which hinges on random sample extractions. Formal and explicit probabilistic bounds on the error made by the approximation algorithm show that the algorithm converges in probability to a neighborhood of the optimal DP solution, bounded by the inherent Bellman error. Additionally, the work provides novel sample-based probabilistic error bounds for general dynamic programming solutions of the reach-avoid problem.

The authors are interested in the non-trivial extension to continuous control spaces, as well as in the assessment of the performance of synthesized approximate policies over the concrete model. For this extended goal, future work will focus on approximating the policy and the value function using two separate function classes.

Finally, the development of better sampling distributions that minimize the error propagation can lead to tighter errors, which can be more relevant in practice. To this end, the optimal sampling distribution should be used to optimize the scaling factors by resembling more closely the local stochastic kernels.
Figure 4: The plot shows the accuracy based on the error propagation over the iterations estimated with an hold out set of size $\hat{N} = 4 \cdot 10^3$ and $\hat{M} = 10^4$. The accuracy $\Delta$, marked as ($\times$), is given for $1 - \delta\Delta = 0.9$. In the graph, the estimate of the accuracy, $\|W^*_k - T^{N_1-k}W^*_N\|$, is given as ($\circ$).

Acknowledgments

We thank the reviewers of an earlier version of this work for the feedback. We would like to acknowledge the support of the Netherlands Organisation for Scientific Research and of the Dutch Institute of Systems and Control. This work has also been supported by the European Commission STREP project MoVeS 257005, by the European Commission Marie Curie grant MANTRAS 249295, by the European Commission IAPP project AMBI 324432, by the European Commission NoE Hycon2 257462, and by the NWO VENI grant 016.103.020.

References

A. Abate, S. Amin, M. Prandini, J. Lygeros, and S. Sastry. Computational approaches to reachability analysis of stochastic hybrid systems. In A. Bemporad, A. Bicchi, and G. Buttazzo, editors, *Hybrid Systems: Computation and Control*, volume 4416 of *Lecture Notes in Computer Science*, pages 4–17. Springer Berlin / Heidelberg, 2007.

A. Abate, M. Prandini, J. Lygeros, and S. Sastry. Probabilistic reachability and safety for controlled discrete time stochastic hybrid systems. *Automatica*, 44(11):2724–2734, November 2008a.

A. Abate, M. Prandini, J. Lygeros, and S. Sastry. An approximate dynamic programming approach to probabilistic reachability for stochastic hybrid systems. In *Proceedings of the*
A. Abate, J.P. Katoen, J. Lygeros, and M. Prandini. Approximate model checking of stochastic hybrid systems. *European Journal of Control*, 16(6):624–641, 2010.

A. Abate, J.-P. Katoen, and A. Mereacre. Quantitative automata model checking of autonomous stochastic hybrid systems. In *Proceedings of the 14th ACM international conference on Hybrid Systems: computation and control*, pages 83–92, 2011.

M. Anthony. Uniform Glivenko-Cantelli theorems and concentration of measure in the mathematical modelling of learning. Research report, Department of Mathematics London School of Economics, May 2002.

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

C. Baier and J.-P. Katoen. *Principles of Model Checking*. The MIT Press, 2008.

P.L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher Complexities. *Annals of Statistics*, 33:1497–1537, 2005.

D.P. Bertsekas and S.E. Shreve. *Stochastic Optimal Control : The discrete time case*. Athena Scientific, 1996.

H.A.P. Blom and J. Lygeros. *Stochastic Hybrid Systems: Theory and Safety Critical Applications*. Number 337 in Lecture Notes in Control and Information Sciences. Springer Verlag, Berlin Heidelberg, 2006.

L. Busoniu, R. Babuska, B.D. Schutter, and D. Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Automation and Control Engineering. CRC Press, 2010.

C.G. Cassandras and J. Lygeros (Eds.). *Stochastic Hybrid Systems*. Number 24 in Control Engineering. CRC Press, Boca Raton, 2006.

E. M. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, 1999.

S. Esmaeil Zadeh Soudjani and A. Abate. Adaptive and sequential gridding procedures for the abstraction and verification of stochastic processes. *SIAM Journal on Applied Dynamical Systems*, 12(2):921–956, 2013.

A.M. Farahmand, R. Munos, and C. Szepesvari. Error Propagation for Approximate Policy and Value Iteration. In *Neural Information Processing Systems*, pages 568–576, 2010.

A. Fehnker and F. Ivančić. Benchmarks for hybrid systems verification. In R. Alur and G.J. Pappas, editors, *In Hybrid Systems: Computation and Control (HSCC 2004)*, pages 326–341. Springer, 2004.

E.J. Hartman, J.D. Keeler, and J.M. Kowalski. Layered neural networks with gaussian hidden units as universal approximations. *Neural computation*, 2(2):210–215, 1990.
D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation/information and Control*, 100:78–150, 1992.

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.

A. Hinton, M. Kwiatkowska, G. Norman, and D. Parker. PRISM: A tool for automatic verification of probabilistic systems. In H. Hermanns and J. Palsberg, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, volume 3920 of *Lecture Notes in Computer Science*, pages 441–444. Springer Verlag, Berlin Heidelberg, 2006.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

M. Karpinski and A. Macintyre. Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and System Sciences*, 54(1):169–176, 1997.

J.-P. Katoen, M. Khattri, and I. S. Zapreev. A Markov reward model checker. In *IEEE Proceedings of the International Conference on Quantitative Evaluation of Systems*, pages 243–244, Los Alamos, CA, USA, 2005.

M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

K. Koutsoukos and D. Riley. Computational methods for reachability analysis of stochastic hybrid systems. In J. Hespanha and A. Tiwari, editors, *Hybrid Systems: Computation and Control*, number 3927 in *Lecture Notes in Computer Sciences*, pages 377–391. Springer-Verlag, Berlin, 2006.

H. J. Kushner and P.G. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer-Verlag, New York, 2001.

H.J. Kushner. *Stochastic stability and control*, volume 33. Academic Press, 1967.

S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Springer Verlag, 1993.

R. Munos and C. Szepesvari. Finite time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

L. Peshkin and S. Mukherjee. Bounds on sample size for policy evaluation in markov environments. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pages 616–630, London, UK, 2001. Springer-Verlag.

D. Pollard. *Convergence of stochastic processes*. Springer Verlag, 1984.

D. Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages 1–86, 1990.
M. Prandini and J. Hu. Stochastic reachability: Theory and numerical approximation. In C.G. Cassandras and J. Lygeros, editors, *Stochastic hybrid systems*, Automation and Control Engineering Series 24, pages 107–138. Taylor & Francis Group/CRC Press, 2006.

S. Summers and J. Lygeros. Verification of discrete time stochastic hybrid systems: A stochastic reach-avoid decision problem. *Automatica*, 46:1951–1961, 2010.

I. Tkachev and A. Abate. Characterization and computation of infinite-horizon specifications over markov processes. *Theoretical Computer Science*, 515:1 – 18, 2014.

V. Vapnik. *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. John Wiley & Sons, Inc., 1998.
Appendix A. Proof of Lemma 8

We employ results on the concentration of random variables (Anthony, 2002), which in general raise conditions on a random variable ensuring its realizations to be concentrated around its expectation, in the sense that the probability of a deviation from the expectation is exponentially small (as a function of the deviation). Of interest to this work is a known bound holding for sums of bounded and independent random variables (Hoeffding, 1963).

**Proposition 16 (Hoeffding’s inequality, Hoeffding (1963))** Suppose that \( X_i \), for \( i = 1, 2, \ldots, N \), are independent random variables supported on \([0, 1]\). Then

\[
P\left\{ \left| \sum_{i=1}^{N} x_i - \mathbb{E} \sum_{i=1}^{N} X_i \right| \geq N \epsilon \right\} \leq 2e^{-2N\epsilon^2},
\]

where \( \mathbb{E} \sum_{i=1}^{N} X_i \) is the mean of the random variable \( \sum_i X_i \), whereas the empirical mean is defined as \( \sum_{i=1}^{N} x_i \), where \( x_i \) is a realization of \( X_i \).

Using Proposition 16 the proof of Lemma 8 is provided as follows.

**Proof** Let us express a probabilistic error bound on the accuracy of the estimate \( \hat{\mathcal{T}} \mathcal{W}^*_{k+1} \) at each base point \( x_i \) and given any \( a \in \mathcal{A} \) as

\[
P\left\{ \left\| \mathcal{T}\mathcal{W}^*_{k+1} - \hat{\mathcal{T}} \mathcal{W}^*_{k+1} \right\|_{p, \hat{\eta}} > \epsilon_1 \right\} \leq \delta_1,
\]

where we have used the empirical norm based on \( \hat{\eta} \). We obtain

\[
P\left\{ \left\| \mathcal{T}\mathcal{W}^*_{k+1} - \hat{\mathcal{T}} \mathcal{W}^*_{k+1} \right\|_{p, \hat{\eta}} \leq \epsilon_1 \right\} = P\left\{ \left\| \mathcal{T}\mathcal{W}^*_{k+1} - \hat{\mathcal{T}} \mathcal{W}^*_{k+1} \right\|_p \leq \epsilon_1 \right\}
= \text{[definition of the empirical norm in (8)]}
= P\left\{ \left( \frac{1}{N} \sum_{i=1}^{N} \left| \mathcal{T}\mathcal{W}^*_{k+1}(x_i^j) - \hat{\mathcal{T}} \mathcal{W}^*_{k+1}(x_i^j) \right|^p \right) \leq \epsilon_1^p \right\}
= \text{[mutual independence of the sample set \( \bigcup_{a \in \mathcal{A}} \left( x_{i,a,j} \right)_{1 \leq j \leq M} \) at each base point \( x_i \)}
\geq P\left\{ \bigcap_{i=1}^{N} \left\{ \left| \mathcal{T}\mathcal{W}^*_{k+1}(x_i^j) - \hat{\mathcal{T}} \mathcal{W}^*_{k+1}(x_i^j) \right|^p \leq \epsilon_1^p \right\} \right\} = \prod_{i=1}^{N} P\left\{ \left| \mathcal{T}\mathcal{W}^*_{k+1}(x_i^j) - \hat{\mathcal{T}} \mathcal{W}^*_{k+1}(x_i^j) \right| \leq \epsilon_1 \right\}.
\]

Let us now express the argument of the probability operator as follows

\[
\left| \mathcal{T}\mathcal{W}^*_{k+1}(x_i^j) - \hat{\mathcal{T}} \mathcal{W}^*_{k+1}(x_i^j) \right| = \max_{a \in \mathcal{A}} \mathbb{E} \left[ 1_{K(x_{k+1})} + 1_{A \setminus K(x_{k+1})} \hat{\mathcal{W}}^*_{k+1}(x_{k+1}) \mid x_{k+1} \sim T_x (\cdot \mid x_i^j, a) \right]
- \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \left[ 1_{K(x_{k+1})} + 1_{A \setminus K(x_{k+1})} \hat{\mathcal{W}}^*_{k+1}(x_{k+1}) \right]
\leq \max_{a \in \mathcal{A}} \mathbb{E} \left[ 1_{K(x_{k+1})} + 1_{A \setminus K(x_{k+1})} \hat{\mathcal{W}}^*_{k+1}(x_{k+1}) \mid x_{k+1} \sim T_x (\cdot \mid x_i^j, a) \right]
- \frac{1}{M} \sum_{j=1}^{M} \left[ 1_{K(x_{k+1})} + 1_{A \setminus K(x_{k+1})} \hat{\mathcal{W}}^*_{k+1}(x_{k+1}) \right].
\]
Therefore the probability of the last event above can be lower bounded by the probability associated to several independent events over the finite action space, as follows:

\[
\mathbb{P}\left\{ \left| \hat{\mathcal{W}}_{k+1}^* (x_k^i) - \hat{\mathcal{W}}_{k+1}^* (x_k^j) \right| \leq \epsilon_1 \right\} \\
\geq \prod_{a \in \mathcal{A}} \mathbb{P}\left\{ \mathbb{E}\left[ 1_K(x_{k+1}) + 1_{\mathcal{A}\setminus K}(x_{k+1})\hat{\mathcal{W}}_{k+1}^*(x_{k+1}) \right] \mid x_{k+1} \sim T_x (\cdot \mid x_k^i, a) \right\} \\
- \frac{1}{M} \sum_{j=1}^{M} \left[ 1_K(x_{k+1}^{i,a,j}) + 1_{\mathcal{A}\setminus K}(x_{k+1}^{i,a,j})\hat{\mathcal{W}}_{k+1}^*(x_{k+1}^{i,a,j}) \right] \leq \epsilon_1 \right\}.
\]

For a given base point \( x_k^i \in \mathcal{X} \), action \( a \in \mathcal{A} \), and function \( \hat{\mathcal{W}}_{k+1}^* \in \mathcal{W} \), define random variables \( Z_j \) via their realizations \( 1_K(x_{k+1}^{i,a,j}) + 1_{\mathcal{A}\setminus K}(x_{k+1}^{i,a,j})\hat{\mathcal{W}}_{k+1}^*(x_{k+1}^{i,a,j}) \), with \( j = 1, \ldots, M \). Since each \( x_{k+1}^{i,a,j} \) is independently drawn from \( T_x (\cdot \mid x_k^i, a) \), the random variables \( Z_j \) are independent, identically distributed, and take values within the closed interval \([0, 1]\). By application of Hoeffding’s inequality (as in Proposition 16), the concentration of the \( M \) samples around the expected value of \( Z_j \) can be expressed as

\[
\mathbb{P}\left\{ \mathbb{E}\left[ 1_K(x_{k+1}) + 1_{\mathcal{A}\setminus K}(x_{k+1})\hat{\mathcal{W}}_{k+1}^*(x_{k+1}) \right] \mid x_{k+1} \sim T_x (\cdot \mid x_k^i, a) \right\} \\
- \frac{1}{M} \sum_{j=1}^{M} \left[ 1_K(x_{k+1}^{i,a,j}) + 1_{\mathcal{A}\setminus K}(x_{k+1}^{i,a,j})\hat{\mathcal{W}}_{k+1}^*(x_{k+1}^{i,a,j}) \right] \geq \epsilon_1 \right\} \leq 2e^{-2M(\epsilon_1)^2}.
\]

Therefore as long as \( 0 \leq 2e^{-2M(\epsilon_1)^2} \leq 1 \), it follows that

\[
\mathbb{P}\left\{ \| \hat{\mathcal{W}}_{k+1}^* - \hat{\mathcal{W}}_{k+1}^* \|_{p, \eta} \leq \epsilon_1 \right\} \geq \left( 1 - 2e^{-2M(\epsilon_1)^2} \right)^{|\mathcal{A}|}.
\]

**Remark 17** As long as we only know that the random variables \( Z_j \) are bounded, the use of Hoeffding’s inequality is sufficient. If we further have information on the variance of \( Z_j \), one can leverage the inequalities of Chebyshev and of Bienaymè-Chebyshev (Hoeffding, 1963), or alternatively Bernstein’s inequality (Peshkin and Mukherjee, 2001): the former bounds are only function of the variance, whereas the latter inequality depends not only on variance of \( Z_j \) but also on its bounded domain. Upper bounds on either the variance of \( Z_j \) or on its range can be derived exploiting prior knowledge on properties of the function space \( \mathcal{W} \) and of the distribution \( T_x (\cdot \mid x, a) \).

**Appendix B. Proof of Lemma 9**

We derive a general, analytical bound on the error of a single backward recursion using notions from statistical learning theory (Haussler, 1992, 1995; Pollard, 1984). The error bound takes into account that, for any \( \hat{\mathcal{W}}_{k+1}^* \), the optimal fit can be anywhere in the function class. Furthermore the bound will be distribution-free, namely holding for any
Markov process (with dynamics characterized by $T_x$) and any sample distribution $\eta$ over the set $A \setminus K$.

We exclusively consider function classes $W \subset B(\mathcal{X}; 1)$ endowed with a finite pseudo-dimension: this includes all finitely-parameterized function classes (Munos and Szepesvari, 2008). The notion of pseudo dimension (Pollard, 1984; Anthony, 2002; Haussler, 1992) expresses the capability of a function class $W$ to fit a set of samples.

**Proof** In order to prove Lemma 9, we show that the inequality in (10) holds for any $W_{k+1}^* \in W$ at any time instant $k = 0, \ldots, N_t - 1$. For the sake of notation in the following, we substitute $\hat{W}_{k+1}^*$ by $W$, and instead of considering the set of base points $(x_k^i)_{1 \leq i \leq N}$ drawn at the time instant $k$ we simply introduce $\bar{x} = (x^1, \ldots, x^N)$ as a sequence of $N$ independent realizations drawn from a distribution over $A \setminus K$ with density $\eta$.

For any given function $W \in \mathcal{W}$, induce a new function class $l_W = \{ |w - TW|^p : w \in W \}$ with elements $l_w \in l_W : l_w = |w - TW|^p$. The inequality in (10) can be rewritten over the function class $l_W$ as follows

$$
P \left\{ \sup_{w \in W} |\|w - TW\|_p^\eta - \|w - TW\|_p^\eta| \geq \epsilon_2 \right\} = P \left\{ \sup_{l_w \in l_W} |E_{\eta} [l_w] - \frac{1}{N} \sum_{i=1}^{N} l_w(x^i)| \geq \epsilon_2 \right\},$$

where $E_{\eta}$ denotes the expected value with respect to $\eta$. This allows us to use a result in (Pollard, 1984), which provides an upper bound on the probability of the above event as a function of the covering number of the metric space $((l_W)_{l_1}, \| \cdot \|_1)$.

**Proposition 18 (Pollard (1984))** Let $\mathcal{F}$ be a permissible set of functions on $\mathcal{X}$ with $0 \leq f(x) \leq K$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Let $\bar{x} = (x_1, \ldots, x_N)$ be a sequence of $N$ samples drawn independently from $\mathcal{X}$ according to any distribution on $\mathcal{X}$. Then for all $\epsilon > 0$

$$
P \left\{ \forall f \in \mathcal{F} : \left| E f - \frac{1}{N} \sum_{i=1}^{N} f(x^i) \right| \geq \epsilon \right\} \leq 4E \mathcal{N}(\epsilon/16, F_{l_{l_1}}, \| \cdot \|_1) e^{-\frac{N\epsilon^2}{128K^2}},$$

(27)

where the quantity $\mathcal{N}$ will be introduced shortly and where the definition of a permissible set of functions (Pollard, 1984) includes all finitely parameterized functions.

Let us introduce the concept of covering number of a metric space (Haussler, 1992). Given a (pseudo-)metric space $(A, \rho)$ and a subset $S$ of $A$, we say that the set $T \subseteq A$ is an $\epsilon$-cover for $S$ (where $\epsilon > 0$) if, for every $s \in S$ there is a $t \in T$ such that $\rho(s, t) < \epsilon$. For a given $\epsilon > 0$ we denote the covering number $\mathcal{N}(\epsilon, S, \rho)$ (Haussler, 1992) as the cardinality of the smallest $\epsilon$-cover of $S$.

For a given set of samples $x^i$ with $i = 1, \ldots, N$, the evaluation of a function $l_w \in l_W$ over each of these samples is given as the $N$ dimensional vector in $[0, 1]^N$: $(l_w)_{l_1} = (l_w(x^1), l_w(x^2), \ldots, l_w(x^N))$. The induced set of vectors is

$$(l_W)_{l_1} = \{(l_w)_{l_1} = (l_w(x^1), l_w(x^2), \ldots, l_w(x^N)) : l_w \in l_W \} \subseteq [0, 1]^N.$$

The minimal $\epsilon$-cover of $((l_W)_{l_1}, \| \cdot \|_1)$ is denoted as $\mathcal{N}(\epsilon, (l_W)_{l_1}, \| \cdot \|_1)$.

The deviation of the expected value from the empirical mean can be bounded using Pollard’s proposition (Pollard, 1984)

$$
P \left\{ \sup_{l_w \in l_W} \left| E_{\bar{x}} [l_w(x)] - \frac{1}{N} \sum_{i=1}^{N} l_w(x^i) \right| \geq \epsilon_2 \right\} \leq 4E \mathcal{N}(\epsilon_2/16, (l_W)_{l_1}, \| \cdot \|_1) e^{-\frac{N\epsilon_2^2}{128}}.$$
The expected value of $\mathcal{N}(e^p_2/16, (l_W)_{\|\cdot\|_1})$ is computed over the samples $x^i$ of $\bar{X}$, drawn independently from a probability distribution with density $\eta$. Since there is a trivial isometry (Haussler, 1992) between $(l_W)_{\|\cdot\|_1}$ and $(l_W, \|\cdot\|_{1,\hat{\eta}})$, both spaces have equal covering numbers

$$\mathcal{N}(e^p_2/16, (l_W)_{\|\cdot\|_1}) = \mathcal{N}(e^p_2/16, l_W, \|\cdot\|_{1,\hat{\eta}}).$$

In practice a value for $E[\mathcal{N}(e^p_2/16, (l_W)_{\|\cdot\|_1})]$ can be obtained by upper bounding $\mathcal{N}(e^p_2/16, l_W, \|\cdot\|_{1,\hat{\eta}})$ independently of the sample distribution. For this we introduce the pseudo dimension of a function class, formally defined as follows (see Pollard, 1984; Anthony, 2002; Haussler, 1992). Suppose $F$ is a class of functions, $f \in F$, $f : \mathcal{X} \rightarrow [0,1]$. Then $S \subseteq \mathcal{X}$ is shattered by $F$ if there are numbers $r_x \in [0,1]$ for $x \in S$ such that for every $T \subseteq S$ there is some $f_T \in F$ with the property that $f_T \geq r_x$ if $x \in T$ and $f_T < r_x$ if $x \in S \setminus T$. We say that $F$ has a finite pseudo dimension $\dim_p(F) = d$ if $d$ is the maximum cardinality of a shattered set.

For any distribution $P \in M(\mathcal{X})$, the packing number (Haussler, 1992) and therefore also the covering number of the metric space $(l_W, \|\cdot\|_{1,\hat{\eta}})$ can be upper bounded as a function of the pseudo-dimension and the base of the natural logarithm $e$: for any $\epsilon > 0$,

$$\mathcal{N}(\epsilon, l_W, \|\cdot\|_{1,\hat{\eta}}) \leq e(d + 1) \left( \frac{2e}{\epsilon} \right)^d, \text{ with } \dim_p(l_W) = d.$$

We have proved that a sufficient upper bound is given as

$$P \left\{ \sup_{w \in W} \left\| w - TW \right\|_{p,\hat{\eta}} - \left\| w - TW \right\|_{p,\hat{\eta}} \geq \epsilon_p \right\} \leq 4e(d + 1) \left( \frac{32e}{\epsilon^p} \right)^d e^{-N(\epsilon_p)^2/128}.$$  

The proof can be concluded by showing that the pseudo dimension $d$ of the induced class $l_W$ is the same as the pseudo dimension of $W$. Let $\{w - TW : w \in W\}$ be a new function class induced from $W$. The invariance properties of the pseudo dimension $\dim_p(W)$ shown in (Haussler, 1992) allow to conclude that $\dim_p(\{w - TW : w \in W\}) = \dim_p(W)$. The induced function class $l_W$ can then be defined as follows: $l_W = \{|k|^p \mid k \in \{w - TW : w \in W\}\}$. Since it was shown in (Kearns and Schapire, 1994) that the pseudo dimension is invariant over function composition ($|\cdot|^p$), we conclude that the pseudo dimension is $\dim_p(l_W) = \dim_p(\{w - TW : w \in W\}) = \dim_p(W) = d$.  

**Remark 19 (Computation of the pseudo-dimension)** When the function class $W$ is a vector space of real-valued functions, the pseudo dimension is equal to the dimensionality of the function class (Anthony and Bartlett, 1999, Theorem 11.4). (Anthony and Bartlett, 1999) elaborates the details of the computation of pseudo dimensions of parameterized function classes, especially for function classes defined over neural networks.

Since it is possible to bound the pseudo dimension of $l_W$ (as introduced in the proof) by the pseudo dimension of $W$, this capacity concept has been used to bound the error caused by using an empirical estimate of the weighted $p$-norm. Notice that for non-parametric function classes, concepts such as covering number or Rademacher average of the function class $l_W$ can be used instead (Bartlett et al., 2005).
Let us shortly discuss how the derived bounds can be tightened. A first option is to circumvent the notion of pseudo dimension and work with the covering numbers in Pollard inequality (Proposition 18), however the increase in assumptions on the function class and in overall computations make the gain in accuracy undeserving. A second option is to explore alternatives over Pollard inequality in (18) with better constants (Bartlett et al., 2005). An alternative concentration inequality based on Bernstein’s inequality is used in (Peshkin and Mukherjee, 2001). Hoeffding inequality gives a concentration inequality on the sum of bounded random variables, whereas Bernstein inequality gives a tighter bound based on knowledge of both the boundness and the variance of the random variables. Even with improved constants or alternative inequalities, the error bounds can still result to be conservative for reasonable sample complexities.

Appendix C. Proof of Theorem 10

The proof of Theorem 10 is adapted from the proof of the single-step error bound for Fitted Value Iteration with multiple sample batches in (Munos and Szepesvari, 2008).

**Proof** Let us introduce a simplified notation for $W_{k+1}$ by replacing it with a general function $W'$ that minimizes the empirical norm as $W' = \arg \min_{W \in W} \|W - \hat{T}W\|_p,\eta$. Let us further define a space $\Omega$ for the batch of samples drawn at any of the iterations, such that at any instant $k$ the realized sample batch $\omega := \bigcup_{i \in \{1, \ldots, N\}} \left( x_k^i \cup \left( \bigcup_{a \in A} \left( x_k^{i,a,j} \right)_{1 \leq j \leq M} \right) \right)$ is an element of the sample space, $\omega \in \Omega$.

For any given $\epsilon' > 0$, consider a function $W^* \in W$ such that $\|W^* - TW\|_{p,\eta} \leq \inf_{W \in W} \|W - TW\|_{p,\eta} + \epsilon'$ (this in particular holds since $W$ has been assumed to be close and bounded).

The error bound in (13) holds for a sample realization $\omega$ if the following sequence of inequalities holds simultaneously:

\[
\|W' - TW\|_{p,\eta} \leq \|W' - \hat{T}W\|_{p,\eta} + \epsilon_1 + \epsilon_2 \tag{28a}
\]
\[
\leq \|W' - \hat{T}W\|_{p,\eta} + \epsilon_1 + \epsilon_2 + \epsilon' \tag{28b}
\]
\[
\leq \|w^* - \hat{T}W\|_{p,\eta} + \epsilon_1 + \epsilon_2 \tag{28c}
\]
\[
\leq \|w^* - TW\|_{p,\eta} + 2\epsilon_1 + \epsilon_2 \tag{28d}
\]
\[
\leq \|w^* - TW\|_{p,\eta} + 2\epsilon_1 + 2\epsilon_2 \tag{28e}
\]

As long as the previous sequence of inequalities is true, the following one also holds:

\[
\|W' - TW\|_{p,\eta} \leq \inf_{w \in W} \|w - TW\|_{p,\eta} + 2\epsilon_1 + 2\epsilon_2 + \epsilon'.
\]

We claim that the sequence of inequalities holds with a probability at least $1 - (\delta_1 + \delta_2)$. Since there exists a function $w^*$ for any $\epsilon' > 0$ it follows with a probability at least $1 - (\delta_1 + \delta_2)$ that

\[
\|W' - TW\|_{p,\eta} \leq d_{p,\eta}(TW, W) + 2\epsilon_1 + 2\epsilon_2.
\]

By the union bound argument (Anthony, 2002), the probability of the union of events can be bounded by the sum of the probabilities of the single events. Using this argument it is possible to define a lower bound on the probability associated with the simultaneous
Observe that the event sets \( W' \) and \( W \) are defined as subsets of the more general event \( B \) defined as

\[
B : \quad e_2^p < \sup_{w \in W} \left| w - TW \right|_{p, \eta} - \left| w - TW \right|_{p, \eta}.
\]

Given two functions \( w^* \) and \( W' \) define events \( A_1 \) and \( A_2 \)

\[
A_1 : \quad e_2^p < \left| w^* - TW \right|_{p, \eta} - \left| w^* - TW \right|_{p, \eta}
\]

\[
A_2 : \quad e_2^p < \left| W' - TW \right|_{p, \eta} - \left| W' - TW \right|_{p, \eta}.
\]

Observe that the event sets \( A_1 \) and \( A_2 \) are subsets of the more general event \( B \) defined as

\[
B : \quad e_2^p < \sup_{w \in W} \left| w - TW \right|_{p, \eta} - \left| w - TW \right|_{p, \eta}.
\]
Thus it follows that for any $\epsilon_2 > 0$: $P\{A_1 \cup A_2\} \leq P\{B\}$ and, based on (12), we have

$$P\left\{\left|\|W' - TW\|_{p,\eta} - \|W' - TW\|_{p,\hat{\eta}}\right| > \epsilon_2\right\} \cup \left\{\left|\|w^* - TW\|_{p,\eta} - \|w^* - TW\|_{p,\hat{\eta}}\right| > \epsilon_2\right\} \leq \delta_2.$$

Thus the probability that the inequalities (28a) and (28e) do not hold is less than $\delta_2$.

The second and fourth inequalities (28b), (28d) depend the accuracy of the estimation of the backward recursion at each base point $x_k^i$. Employing the inequality $\|w - g\|_{p,\eta} < \|w - h\|_{p,\hat{\eta}}$, we can see that

$$\left|\|W' - TW\|_{p,\eta} - \|W' - \hat{TW}\|_{p,\eta}\right| \leq \|TW - \hat{TW}\|_{p,\hat{\eta}},$$

and

$$\left|\|w^* - \hat{TW}\|_{p,\eta} - \|w^* - TW\|_{p,\eta}\right| \leq \|TW - \hat{TW}\|_{p,\hat{\eta}}.$$ 

For every sample set $\omega$ the inequalities (28b), (28d) apply if $\|TW - \hat{TW}\|_{p,\hat{\eta}} \leq \epsilon_1$. Thus

$$P\left\{\left|\|W' - TW\|_{p,\eta} - \|W' - \hat{TW}\|_{p,\eta}\right| > \epsilon_1\right\} \cup \left\{\left|\|w^* - \hat{TW}\|_{p,\eta} - \|w^* - TW\|_{p,\eta}\right| > \epsilon_1\right\} \leq \delta_1.$$ 

The probability that at least one of the inequalities in (28) does not hold can be expressed using the union bound as $\delta_1 + \delta_2$. Thus the sequence of inequalities holds with at least a probability of $1 - \delta_1 - \delta_2$.

**Appendix D. Proof of Lemma 11**

**Proof** Let us set up the following chain of inequalities:

$$\|T^{N_t - k}\hat{W}^*_{N_t - k} - \hat{W}^*_k\|_{p,\eta}$$

$\text{[Add and subtract function $T\hat{W}^*_{k+1}$]}$

$=$ $\|T\left(T^{N_t - k - 1}\hat{W}^*_{N_t - k} - \hat{W}^*_k + \hat{W}^*_k\right) - \hat{W}^*_k\|_{p,\eta}$

$\text{[Definition of $T$ in (2), where we have considered a single $x_k \sim \eta$]}$

$=$ $\max_{a \in A} E\left[1_{K}(x_{k+1}) + 1_{A \setminus K}(x_{k+1}) \left(T^{N_t - k - 1}\hat{W}^*_{N_t - k} - \hat{W}^*_k\right)(x_{k+1}) | x_{k+1} \sim T_x(\cdot | x_k, a)\right]$

$\text{[Max $E[\xi_1 + \xi_2] \leq \max E[\xi_1] + \max E[\xi_2]$]}$

$\leq \max_{a \in A} E\left[1_{K}(x_{k+1}) \left(T^{N_t - k - 1}\hat{W}^*_{N_t - k} - \hat{W}^*_k\right)(x_{k+1}) | x_{k+1} \sim T_x(\cdot | x_k, a)\right]$

$+$ $\max_{a \in A} E\left[1_{K}(x_{k+1}) + 1_{A \setminus K}(x_{k+1})\hat{W}^*_k(x_{k+1}) | x_{k+1} \sim T_x(\cdot | x_k, a)\right] - \hat{W}^*_k$
Let us now show that the first term is bounded by $B^\frac{1}{p} \left\| \mathbf{T}^{N_i - k - 1} \hat{W}_{N_i}^* - \hat{W}_{k+1}^* \right\|_{p,\eta}$:

$$
\left\| \max_{a \in A} \int_{\mathcal{X}} \mathbf{1}_{A \setminus K} (x_{k+1}) \left( \mathbf{T}^{N_i - k - 1} \hat{W}_{N_i}^* - \hat{W}_{k+1}^* \right) t_x (x_{k+1}|x_k, a) dx_{k+1} \right\|_{p,\eta} \\
\leq \left\| \max_{a \in A} \left[ \mathbf{1}_{A \setminus K} (x_{k+1}) \left( \mathbf{T}^{N_i - k - 1} \hat{W}_{N_i}^* - \hat{W}_{k+1}^* \right) \right] \right\|_{p,\eta} \\
+ \left\| \mathbf{T} \hat{W}_{k+1}^* - \hat{W}_{k+1}^* \right\|_{p,\eta}
$$

= [ Definition of $\mathbf{T}$ in (2) ]

$$
\left\| \max_{a \in A} \left[ \mathbf{1}_{A \setminus K} (x_{k+1}) \left( \mathbf{T}^{N_i - k - 1} \hat{W}_{N_i}^* - \hat{W}_{k+1}^* \right) \right] \right\|_{p,\eta} \\
+ \left\| \mathbf{T} \hat{W}_{k+1}^* - \hat{W}_{k+1}^* \right\|_{p,\eta}
$$

= [ Introduce density function $t_x (x_{k+1}|x_k, a)$ for kernel $T_x$ ]

$$
\left\| \max_{a \in A} \int_{\mathcal{X}} \mathbf{1}_{A \setminus K} (x_{k+1}) \left( \mathbf{T}^{N_i - k - 1} \hat{W}_{N_i}^* - \hat{W}_{k+1}^* \right) t_x (x_{k+1}|x_k, a) dx_{k+1} \right\|_{p,\eta} \\
+ \left\| \mathbf{T} \hat{W}_{k+1}^* - \hat{W}_{k+1}^* \right\|_{p,\eta}
$$

= [ Express the $\eta$-weighted $p$-norm over $A \setminus K$ ]

$$
\left( \int_{A \setminus K} \right) \left\| \max_{a \in A} \int_{A \setminus K} \left( \mathbf{T}^{N_i - k - 1} \hat{W}_{N_i}^* - \hat{W}_{k+1}^* \right) t_x (x_{k+1}|x_k, a) dx_{k+1} \right\|_{p,\eta} \\
\leq \left( \int_{A \setminus K} \max_{a \in A} \int_{A \setminus K} \left( \mathbf{T}^{N_i - k - 1} \hat{W}_{N_i}^* - \hat{W}_{k+1}^* \right) \right)^{\frac{1}{p}} \eta(x_k) dx_k
$$

= [ Introduce dummy term $\frac{\eta(x_{k+1})}{\eta(x_k)}$, which is defined over $x_{k+1} \in A \setminus K$ ]

$$
\left( \int_{A \setminus K} \int_{A \setminus K} \max_{a \in A} \left( \mathbf{T}^{N_i - k - 1} \hat{W}_{N_i}^* - \hat{W}_{k+1}^* \right) t_x (x_{k+1}|x_k, a) \right)^{\frac{1}{p}} \eta(x_k) dx_k \eta(x_{k+1}) dx_{k+1}
$$
Introduce now the upper bound on $\mathbb{E}_\eta(\hat{x}_{k+1})$ through the following chain of inequality:

$$
\mathbb{E}_\eta(\hat{x}_{k+1}) \leq \hat{Z}_{k+1}^\ast - \hat{Z}_{k+1}^\ast \leq \hat{Z}_{k+1}^\ast - \hat{Z}_{k+1}^\ast \leq \hat{Z}_{k+1}^\ast - \hat{Z}_{k+1}^\ast.
$$

Introduce now the upper bound on $\mathbb{E}_\eta(\hat{x}_{k+1})$ through the following chain of inequality:

$$
\mathbb{E}_\eta(\hat{x}_{k+1}) \leq \hat{Z}_{k+1}^\ast - \hat{Z}_{k+1}^\ast \leq \hat{Z}_{k+1}^\ast - \hat{Z}_{k+1}^\ast \leq \hat{Z}_{k+1}^\ast - \hat{Z}_{k+1}^\ast.
$$

We have finally shown that

$$
\left\| T^{N_{t-k}} \hat{W}_{k+1} - \hat{W}_k \right\|_{p, \eta} \leq \left\| T \hat{W}_{k+1} - \hat{W}_k \right\|_{p, \eta} + B_{\eta} \left\| T^{N_{t-k}} \hat{W}_{k+1} - \hat{W}_k \right\|_{p, \eta}.
$$

**Appendix E. Proof of Theorem 12**

**Proof** If we estimate the quantity $r_{x_0}(K, A) = (T^{N_{t}} \hat{W}_{N_{t}}^\ast) (x_0) = (T^{N_{t}} \hat{W}_{N_{t}}^\ast) (x_0)$ by $\left( \hat{T} \hat{W}_{1}^\ast \right) (x_0)$, then we have that $r_{x_0}(K, A) = 1_K(x_0) + A_{A \setminus K}(x_0) \left( \hat{T} \hat{W}_{1}^\ast \right) (x_0)$. The absolute deviation of the approximated $r_{x_0}(K, A)$ from the exact $r_{x_0}(K, A)$ is given as

$$
\left| r_{x_0}(K, A) - r_{x_0}(K, A) \right| = \left| (T^{N_{t}} \hat{W}_{N_{t}}^\ast) (x_0) - \left( \hat{T} \hat{W}_{1}^\ast \right) (x_0) \right|.
$$

The objective is to present this error as a function of the errors introduced by the approximate mappings $\left( \left( T^{N_{t}} \hat{W}_{N_{t}}^\ast \right) (x_0) - \left( \hat{T} \hat{W}_{1}^\ast \right) (x_0) \right)$, as well as of the quantities $\left\| \hat{W}_{1}^\ast - \hat{W}_{2}^\ast \right\|_{p, \eta}, \left\| \hat{W}_{2}^\ast - \hat{W}_{3}^\ast \right\|_{p, \eta}, \ldots, \left\| \hat{W}_{N_{t}-1}^\ast - \hat{W}_{N_{t}}^\ast \right\|_{p, \eta}$.

To this end, we first express a bound on $\left| (T^{N_{t}} \hat{W}_{N_{t}}^\ast) (x_0) - \left( \hat{T} \hat{W}_{1}^\ast \right) (x_0) \right|$ as a function of $\left| \left( \hat{T} \hat{W}_{1}^\ast \right) (x_0) - (\hat{W}_{1}^\ast (x_0)) \right|$ and of $\left\| T^{N_{t}-1} \hat{W}_{N_{t}}^\ast - \hat{W}_{1}^\ast \right\|_{p, \eta}$. Then Lemma 11 is used to express $\left\| T^{N_{t}-1} \hat{W}_{N_{t}}^\ast - \hat{W}_{1}^\ast \right\|_{p, \eta}$ as a function of the errors introduced by the approximate mappings. Similar to the first chain of inequality in the proof of Lemma 11 applied at step $k = 0$ and point $x_0$, we obtain that

$$
\left| (T^{N_{t}} \hat{W}_{N_{t}}^\ast) (x_0) - \left( \hat{T} \hat{W}_{1}^\ast \right) (x_0) \right| \leq \max_{a \in A} \int_{A \setminus K} T^{N_{t}-1} \hat{W}_{N_{t}}^\ast (x_1) | x_1 (x_0, a) d x_1 + \left| \left( \hat{T} \hat{W}_{1}^\ast \right) (x_0) - \left( \hat{T} \hat{W}_{1}^\ast \right) (x_0) \right|.
$$

Let us now introduce a measure for the maximum concentration of the density function $t_x (x_1 | x_0, a)$ over $x_1 \in A \setminus K$, for any $a \in A$, defined relative to the density of the distribution $\eta$ in $\{16\}$, as $B_0 = \sup_{x_1 \in A \setminus K} \max_{a \in A} t_x (x_1 | x_0, a) \eta (x_1)$. Since $B_0 \eta (x_1) \geq t_x (x_1 | x_0, a)$, it
follows that
\[
\max_{a \in A} \int_{A \setminus K} \left| T^{N_1-1} \hat{W}_N^*(x_1) - \hat{W}_1^*(x_1) \right| t_x(x_1 \mid x_0, a) \, dx_1 \\
\leq B_0 \int_{A \setminus K} \left| T^{N_1-1} \hat{W}_N^*(x_1) - \hat{W}_1^*(x_1) \right| \eta(x_1) \, dx_1.
\]

The last expression corresponds to a 1-norm with respect to a probability measure \( \eta \) over \( A \setminus K \). Exploiting the monotonicity of the \( p \)-norm with respect to a probability measure, a more general expression for the approximation error is obtained as
\[
\left| \left( T^{N_1} \hat{W}_N^* \right)(x_0) - \hat{W}_1^*(x_0) \right| \leq \left| \left( \hat{T} \hat{W}_1^* \right)(x_0) - \left( T \hat{W}_1^* \right)(x_0) \right| + B_0 \left\| \left( T^{N_1-1} \hat{W}_N^* \right) - \hat{W}_1^* \right\|_{p, \eta}.
\]

The second term can be expressed as a function of the weighted \( p \)-norm of the approximations by applying Lemma 11. This leads to the expression for an upper bound on the approximation error as
\[
\left| \hat{r}_{x_0}^*(K, A) - r_{x_0}^*(K, A) \right| \leq \left| \left( \hat{T} \hat{W}_1^* \right)(x_0) - \left( T \hat{W}_1^* \right)(x_0) \right| + B_0 \sum_{k=1}^{N_1-1} B^{k-1} \left\| \hat{W}_k^* - T \hat{W}_{k+1}^* \right\|_{p, \eta}.
\]

From the above expression, a sufficient condition the accuracy in (5) to hold is
\[
P \left\{ \left| \left( \hat{T} \hat{W}_1^* \right)(x_0) - \left( T \hat{W}_1^* \right)(x_0) \right| + B_0 \sum_{k=1}^{N_1-1} B^{k-1} \left\| \hat{W}_k^* - T \hat{W}_{k+1}^* \right\|_{p, \eta} > \Delta \right\} \leq \delta_1.
\]

### Appendix F. Sample Complexities

Given \( \epsilon_{0,1,2} \) and \( \alpha \), select \( \delta_{0,1,2} > 0 \) such that \( 1 - \alpha = \delta_0 + (N_1 - 1) \delta_1 + (N_1 - 1) \delta_2 \), and let us pick values for \( N, M, M_0 \) such that
\[
\delta_0 \leq 2|A| \epsilon^{-2N_0(\epsilon_0)^2}, \quad \delta_1 \leq 2|A| N \epsilon^{-2M(\epsilon_1)^2}, \quad \delta_2 \leq 4e(d + 1) \left( \frac{32e}{\epsilon_0^2} \right)^d e^{-N_0^2 \epsilon^{-2M(\epsilon_1)^2}}.
\]

Note that the first two inequalities are approximated with first order approximation for which we know that \( 1 - (1 - 2e^{-2M_0(\epsilon_0)^2}) |A| \leq 2|A| \epsilon^{-2M_0(\epsilon_0)^2} \) and \( 1 - (1 - 2e^{-2M(\epsilon_1)^2}) |A| N \leq 2|A| N \epsilon^{-2M(\epsilon_1)^2} \). The obtained integer values for \( N, M, M_0 \) are given as
\[
\begin{align*}
N &= \left[ 128 \ln(4e(d + 1)) + d \ln(32e) \left( \frac{1}{\epsilon_0^2} \right)^{2p} + 128dp \left( \frac{1}{\epsilon_0^2} \right)^2 \ln \left( \frac{1}{\epsilon_0^2} \right) + 128 \left( \frac{1}{\epsilon_0^2} \right)^2 \ln \left( \frac{1}{\delta_2} \right) \right], \\
M &= \left[ \frac{d}{2} \left( \frac{1}{\epsilon_1} \right)^2 \ln(2|A|) + \ln(\frac{1}{\epsilon_1}) + \ln(N) \right], \\
M_0 &= \left[ \frac{1}{2} \left( \frac{1}{\epsilon_0} \right)^2 \ln(2|A|) + \ln(\frac{1}{\epsilon_0}) \right].
\end{align*}
\]

The use of the obtained \( M, M_0, N \) in (18) leads to a confidence of at least \( \alpha \).
Appendix G. Proof of Theorem 15

Proof The proof of Theorem 15 is built observing that (a.) the single step error \( \| \hat{W}_k^* - T\hat{W}_{k+1}^* \|_{1,\eta} \) is bounded by the sum of the expectations of (20) and (22); that (b.) the propagation of the single step errors gives a bound on the overall approximation error, see Theorem 12 - hence the expected value of the estimates, propagated over the time horizon, also gives a bound on the approximation error; and that (c.) the one-sided application of the Hoeffding’s inequality provides a probabilistic upper bound on the deviation of the estimate from its mean, and therefore also bounds the approximation error probabilistically.

Part (a.)

\[ \| \hat{W}_k^* - T\hat{W}_{k+1}^* \|_{1,\eta} = E_x \left[ \| \hat{W}_k^* (x) - T\hat{W}_{k+1}^* (x) \| \right] \text{ with } E_x [f(x)] \text{ the mean of } f(x) \text{ for } x \sim \eta. \]

Define a set of i.i.d. random variables \( \tilde{y}_1 = [y_1^{a,1}, y_1^{a,2}, \ldots, y_1^{a,M}] \) drawn from the distribution \( y_1^{a,j} \sim T_x (\cdot | x, a) \). Introduce \( E_{\tilde{y}_1} [ \max_{a \in A} T_k \hat{W}_{k+1}^* (x) | x] \) as an auxiliary variable with \( \hat{T}_k^* \) the estimated operator as defined in (21) and computed over the \( \tilde{y}_1 \).

\[
E_x \left[ \hat{W}_k^* (x) - E_{\tilde{y}_1} \left[ \max_{a \in A} \hat{T}_k^* \hat{W}_{k+1}^* (x) \right] \right] + E_{\tilde{y}_1} \left[ \max_{a \in A} \hat{T}_k^* \hat{W}_{k+1}^* (x) - T\hat{W}_{k+1}^* (x) \right]
\]

\[
\leq E_x \left[ \hat{W}_k^* (x) - E_{\tilde{y}_1} \left[ \max_{a \in A} \hat{T}_k^* \hat{W}_{k+1}^* (x) \right] \right] + E_{\tilde{y}_1} \left[ \max_{a \in A} \hat{T}_k^* \hat{W}_{k+1}^* (x) - T\hat{W}_{k+1}^* (x) \right]
\]

Observe that the single step error, \( E_{x,\tilde{y}_1} \left[ \hat{W}_k^* (x) - \max_{a \in A} \hat{T}_k^* \hat{W}_{k+1}^* (x) \right] \) is equal to \( \| \hat{W}_k^* - T\hat{W}_{k+1}^* \|_{1,\eta} \) and \( E\| \hat{W}_k^* - T\hat{W}_{k+1}^* \|_{1,\tilde{y}_1} \). The bias term gives the bias introduced by using an estimate of the operator and it can be rewritten as the expected value of (22). Note that \( \max_{a \in A} E_{\tilde{y}_1} [V_{k+1} (y) | x, a] \) is a function of \( x, a \) and \( E_{\tilde{y}_1} [f(\tilde{y})] \leq E_{\tilde{y}_1} [f(\tilde{y})] \), thus it follows that

\[
\| \hat{W}_k^* - T\hat{W}_{k+1}^* \|_{1,\eta} \leq E_{x,\tilde{y}_1} \left[ \max_{a \in A} \hat{T}_k^* \hat{W}_{k+1}^* (x) - \max_{a \in A} T_k \hat{W}_{k+1}^* (x) \right] \| x \|.
\]

The second inequality follows from introducing a secondary set of random variables \( \tilde{y}_2 = [y_2^{a,1}, y_2^{a,2}, \ldots, y_2^{a,1M}] \) for which the elements are i.i.d. as \( y_2^{a,j} \sim T_x (\cdot | x, a) \) and which are independent of \( \tilde{y}_1 \). Substituting \( T\hat{W}_{k+1}^* (x) \) with \( \max_{a \in A} E_{\tilde{y}_2} \left[ \hat{T}_2^* \hat{W}_{k+1}^* (x) | x, a \right] \) we have

\[
\| \hat{W}_k^* - T\hat{W}_{k+1}^* \|_{1,\tilde{y}_2} \leq E_{x,\tilde{y}_1,\tilde{y}_2} \left[ \max_{a \in A} \hat{T}_2^* \hat{W}_{k+1}^* (x) - \max_{a \in A} T_2 \hat{W}_{k+1}^* (x) \right] \| x, a, \tilde{y}_1 \|.
\]

The last equality is equal to the expected value of the estimated bias \( E \| \max_{a \in A} \hat{T}_2^* \hat{W}_{k+1}^* - T_2 \hat{W}_{k+1}^* \|_{1,\tilde{y}_2} \).

This proves statement (a.).

Part (b.) & (c.) Based on Theorem 12 \( \hat{\nu}_{x_0}^*(K, A) \) has accuracy \( \Delta \) with probability \( \delta_\Delta \) if

\[
P \left\{ \| W_0^* (x_0) - T\hat{W}_1^* (x_0) \| + B_0 \sum_{k=1}^{N_1} B_k \| \hat{W}_k^* (x) - \max_{a \in A} T_k \hat{W}_{k+1}^* (x) \| + \max_{a \in A} \| T_k \hat{W}_{k+1}^* (x) - T_2 \hat{W}_{k+1}^* (x) \| \geq \Delta \right\} < \delta_\Delta.
\]
Which holds under a union bounding argument if

\[
P \left \{ |\hat{W}_n(x_0) - \mathfrak{T}\hat{W}_n(x_0)| \geq \varepsilon_0 \right \} < \delta_0
\]

(30)

\[
P \left \{ B_0 \sum_{k=1}^{N_1} B^{k-1} \mathbb{E} \left [ |\hat{W}_n(x) - \max_{a \in A} \hat{T}_1 \hat{W}_{k+1}^*(x)| + \max_{a \in A} |\hat{T}_1 \hat{W}_{k+1}^*(x) - \hat{T}_2 \hat{W}_{k+1}^*(x)| \right ] \geq B_0 \varepsilon \right \}
\]

(31)

\[
< e^{-\frac{N_1 \varepsilon^2}{2}}
\]

and \( \Delta \) and \( \delta_\Delta \) are given as (23a) and (23b).

The probabilistic bound (30) follows from Lemma 8 for the estimation error of an empirical norm with accuracy \( \varepsilon_0, \delta_0 \) obtained for \( p = 1, M = M_0 \) and \( N = 1 \) as long as \( 0 < 2e^{-2M_0 \varepsilon_0^2} < 1 \). The probabilistic bound (31) follows from a one-sided Hoeffding’s inequality (Hoeffding, 1963) with random variable

\[
\sum_{k=1}^{N_1} B^{k-1} \left ( |\hat{W}_n^*(x) - \max_{a \in A} \hat{T}_1 \hat{W}_{k+1}^*(x)| + \max_{a \in A} |\hat{T}_1 \hat{W}_{k+1}^*(x) - \hat{T}_2 \hat{W}_{k+1}^*(x)| \right ),
\]

obtained from the combination of random variable \( x \sim \eta \) and conditional random variables \( \tilde{y}_1 \) and \( \tilde{y}_2 \) and taking values in the range \( [0, 2 \sum_{k=1}^{N_1} B^{k-1}] \). Note that its estimated of interest over \( \tilde{N} \) samples can be rewritten in the form of (23a),

\[
\sum_{k=1}^{N_1} B^{k-1} \left ( |\hat{W}_n^* - \hat{T}_1 \hat{W}_{k+1}^*| + \left \| \max_{a \in A} \hat{T}_1 \hat{W}_{k+1}^* - \hat{T}_2 \hat{W}_{k+1}^* \right \| \right )
\]

This concludes the proof of Theorem 15.

**Appendix H. Scaling factor for case study**

Compute \( B \) as in (14) using the given density distribution of the transitions (26), as

\[
B = \sup_{y \in A \setminus K} \int_{A \setminus K} \frac{1}{\sqrt{\Sigma(2\pi)^2}} \max_{a \in A} \left ( \exp \left ( -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right ) \eta(y) \right ) dx
\]

\[
= \left [ \mu \text{ is a function of } a \text{ and } x, \text{ and } \eta(\cdot) \text{ is constant over } A \setminus K \right ]
\]

\[
= \sup_{y \in A \setminus K} \int_{A \setminus K} \frac{1}{\sqrt{\Sigma(2\pi)^2}} \max_{a \in A} \exp \left ( -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right ) dx
\]

\[
= [ \text{Suppose } A \text{ is invertible, and define } \bar{\mu}(y, a) = A^{-1} y - A^{-1} B a - A^{-1} C, \Sigma = A^{-1} \Sigma A^{-T} ]
\]

\[
= \sup_{y \in A \setminus K} \frac{1}{|A|} \int_{A \setminus K} \frac{1}{\sqrt{\Sigma(2\pi)^2}} \max_{a \in A} \exp \left ( -\frac{1}{2} (x - \bar{\mu}(y, a))^T \Sigma^{-1} (x - \bar{\mu}(y, a)) \right ) dx
\]

\[
\leq \sup_{y \in A \setminus K} \frac{1}{|A|} \int_{A \setminus K} \frac{1}{\sqrt{\Sigma(2\pi)^2}} \sum_{a \in A} \exp \left ( -\frac{1}{2} (x - \bar{\mu}(y, a))^T \Sigma^{-1} (x - \bar{\mu}(y, a)) \right ) dx
\]

40
\[
\begin{align*}
&= \sup_{y \in A \setminus K} \frac{1}{|A|} \left( \sum_{a \in A} \int_{A \setminus K} \frac{1}{\sqrt{|\Sigma|}} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( x - \bar{\mu}(y,a) \right)^T \Sigma^{-1} \left( x - \bar{\mu}(y,a) \right) \right) dx \right). \\
\text{The integral is rewritten as one over a scaled 2-dimension multivariate Gaussian density distribution with mean } \bar{\mu} \text{ and covariance } \Sigma. \text{ With this result, it can be deduced that } B \text{ is smaller than } \frac{1}{|A||A|}\text{ as}
\end{align*}
\]

\[
B \leq \sup_{y \in A \setminus K} \frac{1}{|A|} \left( \sum_{a \in A} \int_X \frac{1}{\sqrt{|\Sigma|}} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( x - \bar{\mu}(y,a) \right)^T \Sigma^{-1} \left( x - \bar{\mu}(y,a) \right) \right) dx \right), \tag{32}
\]

\[
\leq \sup_{y \in A \setminus K} \frac{1}{|A|} \left( \sum_{a \in A} 1 \right) = \frac{1}{|A|}|A|. 
\]