Sampling-based approximations with quantitative performance for the probabilistic reach-avoid problem over general Markov processes

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.
Link to publication

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

Sofie Haesaert   s.haesaert@tue.nl

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

Robert Babuska   r.babuska@tudelft.nl

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

Alessandro Abate   alessandro.abate@cs.ox.ac.uk

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 computation of maximal reach-avoid specifications, together with the synthesis of the corresponding optimal controllers. 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. 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 gives a-priori computable 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. Furthermore, we provide tighter probabilistic error bounds that are sample-based. 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 few 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 (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 (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 a 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. We are further interested in a class of such models known as stochastic hybrid systems (SHS) (Abate et al., 2008), which are endowed with a "hybrid" (that is, both continuous and discrete) state space, which are relevant for a number of applications in Engineering and the Life Sciences (Blom and Lygeros, 2006; Cassandras and J. Lygeros, 2006). This work investigates the problem of controller synthesis over these models, namely the selection of sequences of control inputs (which in particular can be functions of the states of the model) over a finite time horizon, in order to optimise a given figure of merit.

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, we focus on the reach-avoid specification, a property that is well known and central within the Formal Verification field (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 computational 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 finite time horizon, a trajectory of the model enters a goal set, while avoiding a given set of undesired states (both sets are arbitrary measurable 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 studied 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 Time Logic (Baier and Katoen, 2008). From a controller synthesis perspective, the goal becomes that of either maximizing or minimizing the above specification over the given time horizon.

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., 2008; Summers and Lygeros, 2010). These results have recently led to the study of other properties, either richer (Abate et al., 2011) or defined over unbounded time horizons (Tkachev and Abate, 2014). These 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 (Koutsoukos and Riley, 2006; Kushner and Dupuis, 2001; Prandini and Hu, 2006), which provide results with properties that are only known asymptotically.

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. 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 guarantees. Additionally, novel and tighter probabilistic error bounds for dynamic programming solutions of the reach-avoid problem based on samples extraction are presented. As a comparison to the alternative techniques in the literature (Abate et al., 2010; Esmaeil Zadeh Soudjani and Abate, 2013), we show the related techniques provide bounds that are valid deterministically, whereas the proposed result yields tighter results in general that are valid with a certain (tunable) confidence. The outcomes lead to an approach providing controller synthesis with a certified performance, which is relevant for safety-critical applications (Blom and Lygeros, 2006). The proofs of the statements are included in the Appendix.

2. Probabilistic Reach-Avoid Problem over General Markov Processes

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 $T_x (\cdot \mid x, a)$ 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 σ-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 $\int_A T_x (dy \mid x, a) = \int_A t_x (y \mid x, a) \, dy$.

Definition 2 (Markov policy) *A Markov policy μ over 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 μ, 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) σ-algebra and allows computing probability associated to events over trajectories – we denote this probability by $\mathbf{P}$, and further define the probabilities $\mathbf{P}_{x_0}, \mathbf{P}_{x_0}^{\mu}$ as $\mathbf{P}$ 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 $T_x(\cdot \mid x_k, \mu_k(x_k))$. The model can be initialized according to an initial probability measure $P_0 \in M(\mathcal{X})$, where $M(\cdot)$ denotes the collection of probability distributions over a given set.

2.1 Probabilistic Reach-Avoid Problem: Definition

Let us define the probabilistic reach-avoid problem, also known as constrained reachability (Baier and Katoen, 2008), and provide its characterization. Consider a safe set $A \in \mathcal{B}(\mathcal{X})$, a target set $K \in \mathcal{B}(\mathcal{X})$, and a finite 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 it reaches the target set K within the time horizon, while staying inside the safe set A. This property can be expressed as

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

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 $\mathcal{X} \setminus A$. Formally,

$$r_{x_0}^{\mu}(K, A) = \mathbf{P}_{x_0}^{\mu} \{\exists j \in [0, N_t] : x_j \in K \wedge \ \forall i \in [0, j-1] : x_i \in A \setminus K\}, \quad (1)$$

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

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

where $\mathbf{1}_B(x) = 1$ 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 over set $\mathcal{X} \setminus B$.

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

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

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 follows.

Proposition 3 *Given a policy $\mu = (\mu_0, \mu_1, \ldots, \mu_{N_t-1})$, define function $W_k^\mu : \mathcal{X} \to [0,1]$ by backward recursion*

$$W_k^\mu(x) = \mathbf{E}_x^{\mu_k} \left[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A \backslash K}(x_{k+1}) W_{k+1}^\mu(x_{k+1}) \right],$$

and initialized with $W_{N_t}^\mu(x) = 0$. Then for any initial state $x_0 \in \mathcal{X}$, the probabilistic reach-avoid property $r_{x_0}^\mu(K, A)$ can be expressed as

$$r_{x_0}^\mu(K, A) = \mathbf{1}_K(x_0) + \mathbf{1}_{A \backslash K}(x_0) W_0^\mu(x_0) .$$

Proof The proof follows (Summers and Lygeros, 2010, Lemma 4), where the above statement is proven for a value function $V_k^\mu(x) = \mathbf{1}_K(x) + \mathbf{1}_{A \backslash K}(x) W_k^\mu(x)$. ∎

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 μ as done above, we now focus on the controller synthesis problem, which seeks the Markov policy μ^* that maximizes the probabilistic reach-avoid property, and which is such that $r_{x_0}^*(K, A) = \sup_{\mu \in \mathcal{M}} r_{x_0}^\mu(K, A)$. 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*

$$W_k^*(x) = \max_{a \in \mathcal{A}} \mathbf{E}_x^a \left[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A \backslash K}(x_{k+1}) W_{k+1}^*(x_{k+1}) \right],$$

with $x_{k+1} \sim T_x(\cdot \mid x, a)$ for $k = N_t - 1, N_t - 2, \ldots, 0$, and initialized by $W_{N_t}^*(x) = 0$. 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

$$r_{x_0}^*(K, A) = \mathbf{1}_K(x_0) + \mathbf{1}_{A \backslash K}(x_0) W_0^*(x_0).$$

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

$$\mu_k^*(x) = \arg \max_{a \in \mathcal{A}} \mathbf{E}_x^a \left[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A \backslash K}(x_{k+1}) W_{k+1}^*(x_{k+1}) \right]$$

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., 2008). ∎

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

$$
\begin{aligned}
W_k^*(x) &= \left(\mathsf{T}W_{k+1}^*\right)(x) \\
&= \max_{a \in \mathcal{A}} \mathbf{E}_x^a\big[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\setminus K}(x_{k+1})W_{k+1}^*(x_{k+1})\big] .
\end{aligned}
\tag{2}
$$

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

$$
r_{x_0}^*(K, A) = \mathbf{1}_K(x_0) + \mathbf{1}_{A\setminus K}(x_0) \left(\mathsf{T}^{N_t}W_{N_t}^*\right)(x_0).
$$

2.2 Computation of the Reach-Avoid Probability

Notice that generally it is not possible to solve the above recursions exactly: in order to determine the backwards iteration at a single point $x_i \in \mathcal{X}$, namely $W_k^*(x_i) = \left(\mathsf{T}W_{k+1}^*\right)(x_i)$, 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 $W_k^* : \mathcal{X} \to [0, 1]$ leads to the use of approximation techniques, which can be categorized in two families:

1. *Numerical approximation techniques*, which provide an approximation of the optimal probabilistic reach-avoid problem with actual error bounds. More precisely, for a given error bound $\boldsymbol{\Delta} > 0$ we seek a numerical scheme that obtains an approximation $\hat{r}_{x_0}^*(K, A)$, which is such that $|\hat{r}_{x_0}^*(K, A) - r_{x_0}^*(K, A)| \leq \boldsymbol{\Delta}$. This approach is taken in (Abate et al., 2010; 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 with a Markov chain by partitioning the state space.

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. For a given error bound $\boldsymbol{\Delta} > 0$ and confidence $1 - \delta_{\boldsymbol{\Delta}}$, the probability that the approximation is not close to the the optimal value can be bounded probabilistically as

$$
\mathbf{P}\left\{\left|\hat{r}_{x_0}^*(K, A) - r_{x_0}^*(K, A)\right| > \boldsymbol{\Delta}\right\} \leq \delta_{\boldsymbol{\Delta}}.
\tag{3}
$$

In this work, we newly pursue the second approach by focusing on results from the area of learning, and in particular on algorithms for functional approximations. A learning approach is 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). We adopt the Fitted Value Iteration scheme (FVI) (Munos and Szepesvari, 2008), a learning algorithm fit in particular for finite horizon settings.

In practice, bounds for probabilistic approximation methods (3) can be divided into two groups. Firstly, general model-free and sample-free bounds, which provide *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. Alternatively, model-based and sample-based bounds: these bounds verify the accuracy of a dynamic programming scheme by drawing samples of the model and using available information from the model, and from 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 4 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 additive-cost optimal control problems, and adapt it to the reach-avoid optimal control setting. Known as FVI Munos and Szepesvari (2008), the algorithm extracts a finite number of samples from the underlying model to numerically approximate the value recursions in (2). 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 the analytical function $\hat{W}_k^* \in \mathcal{W}$ to the estimate. 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.

Algorithm 1: Sample generation for the FVI algorithm

Given a safe set A, a target set K, a time horizon N_t, and distribution η, generate N base points, and M samples at each base point as follows:

1. *Draw N base points $\left(x_k^i\right)_{1 \le i \le N}$ from the distribution η in $A \setminus K$;*

2. *Draw M samples at each base point x_k^i and at each actions $a \in \mathcal{A}$ from the stochastic kernel T_x, and denote the set of samples as $\left(x_{k+1}^{i,a,j}\right)_{1 \le j \le M}$.*

The FVI algorithm employs samples that are generated by the underlying model. Samples referred to as "base points" are taken from a chosen distribution η, and additional samples are drawn at the base points from the transition kernel. Algorithm 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, namely 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 $(\mathsf{T}\hat{W}_{k+1}^*)(x_k^i)$ at N base points x_k^i. The recursion $(\mathsf{T}\hat{W}_{k+1}^*)(x_k^i)$ is estimated by an empirical operator $\hat{\mathsf{T}}$ as follows:

$$\left(\hat{\mathsf{T}}\hat{W}_{k+1}^*\right)(x_k^i) = \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_K(x_{k+1}^{i,a,j}) + \mathbf{1}_{A \setminus K}(x_{k+1}^{i,a,j})\hat{W}_{k+1}^*(x_{k+1}^{i,a,j}). \qquad (4)$$

Here $x_{k+1}^{i,a,j}$ represent M independent and identically distributed realizations obtained from $T_x\left(\cdot \mid x_k^i, a\right)$. Hence, $j \in [1, M], i \in [1, N]$, and $a \in \mathcal{A}$. For an increasing number

of samples $x_{k+1}^{i,a,j}$, the estimate $(\hat{\mathsf{T}}\hat{W}_{k+1}^*)(x_k^i)$ converges to $(\mathsf{T}\hat{W}_{k+1}^*)(x_k^i)$ 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

$$\hat{W}_k^* = \arg\min_{w \in \mathcal{W}} \sum_{i=1}^{N} \left| w(x_k^i) - \hat{\mathsf{T}}\hat{W}_{k+1}^*(x_k^i) \right|^p. \tag{5}$$

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 η 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 (5) converges to $\int_{A \setminus K} |w(x) - \hat{\mathsf{T}}\hat{W}_{k+1}^*(x)|^p \eta(x)dx$, by the law of large numbers. This leads (cf. Section 4) to the convergence of the argument $\hat{W}_k^*$ to the optimal fit that minimizes the distance between $\hat{\mathsf{T}}\hat{W}_{k+1}^*$ and the functions $w \in \mathcal{W}$ with respect to the p-norm, weighted by a distribution with density η and supported over the set $A \setminus K$, namely

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

The overall FVI algorithm is summarized in Algorithm 2. The iterations are initialized as $\hat{W}_{N_t}^*(x) = 0, x \in \mathcal{X}$, and updated over the functions $\hat{W}_{N_t-1}^*, \hat{W}_{N_t-2}^*, \ldots, \hat{W}_1^*$. Finally the value function at $k = 0$ is approximated at the initial condition x_0 with a sample-based integration, similar to (4), using M_0 independent and identically distributed realizations of $T_x(\cdot \mid x_0, a)$ for all $a \in \mathcal{A}$.

Algorithm 2: Fitted Value Iteration algorithm

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 η, perform:

  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. Algorithm 1);*
     (b) *Estimate $\mathsf{T}\hat{W}_{k+1}^*$ as $(\hat{\mathsf{T}}\hat{W}_{k+1}^*)(x_k^i)$ (cfr. (4))*
     (c) *Find the function that minimizes the empirical p-norm as*

$$\hat{W}_k^* = \arg\min_{w \in \mathcal{W}} \sum_{i=1}^{N} \left| w(x_k^i) - \hat{\mathsf{T}}\hat{W}_{k+1}^*(x_k^i) \right|^p;$$

  3. *Collect M_0 samples for the single initial condition x_0 and for every action $a \in \mathcal{A}$, and estimate $\hat{W}_0^*(x_0)$ as in step (2)(b);*
  4. *Return reach-avoid probability*

$$\hat{r}_{x_0}^*(K, A) = \mathbf{1}_K(x_0) + \mathbf{1}_{A \setminus K}(x_0)\hat{W}_0^*(x_0), \forall x_0 \in \mathcal{X}.$$

Remark 5 (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 Algorithm 2, first the policy is estimated at all the base points x_k^i, as the argument of (2)(b). Secondly a classification algorithm is used, providing an approximately optimal policy $\hat{\mu}_k : \mathcal{X} \to \mathcal{A}$ for each k.*

4. A-Priori Probabilistic Error Bounds

Let us recall the accuracy of the FVI algorithm as in (3): we say that the FVI algorithm has an accuracy $\mathbf{\Delta}$, with a confidence $1 - \delta_{\mathbf{\Delta}}$, if the probability that the error made by the approximate solution $\hat{r}_{x_0}^*(K, A)$ is larger than $\mathbf{\Delta}$, is upper-bounded by $\delta_{\mathbf{\Delta}}$. We explicitly quantify the accuracy in (3) and analyse it in two steps: first by computing a bound on the error of a single iteration (Sec. 4.1); then by studying the propagation of the single-step error over multiple iterations (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 a Single Iteration of the FVI

Let $\mathsf{T}\hat{W}_{k+1}^* : \mathcal{X} \to [0, 1]$ 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 $\mathsf{T}\hat{W}_{k+1}^*$ with respect to the following weighted, p-norm:

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

Notice that if $\mathsf{T}\hat{W}_{k+1}^* \notin \mathcal{W}$, the optimal approximation $\inf_{w \in \mathcal{W}} \|w - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\eta}$ 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 $\inf_{w \in \mathcal{W}} \|w - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\eta}$ is derived as

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

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

As discussed, the FVI algorithm employs empirical estimates, given in (4)-(5), 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 $\mathsf{T}\hat{W}_{k+1}^*$ 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 η, 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 $\hat{\mathsf{T}}\hat{W}^*_{k+1}$ 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 $\mathsf{T}\hat{W}^*_{k+1}$

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

$$\frac{1}{M}\sum_{j=1}^M \mathbf{1}_K(x^{i,a,j}_{k+1}) + \mathbf{1}_{A\backslash K}(x^{i,a,j}_{k+1})\hat{W}^*_{k+1}(x^{i,a,j}_{k+1})$$

of $\mathbf{E}^a_{x^i_k}\left[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\backslash K}(x_{k+1})\hat{W}^*_{k+1}(x_{k+1})\right]$. Since the cardinality of the action space $\mathcal{A}$ 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 $1 \leq j \leq M$, $\mathbf{1}_K(x^{i,a,j}_{k+1}) + \mathbf{1}_{A\backslash K}(x^{i,a,j}_{k+1})\hat{W}^*_{k+1}(x^{i,a,j}_{k+1})$, 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:

$$\mathbf{P}\Big\{\Big|\mathbf{E}^a_{x^i_k}\Big[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\backslash K}(x_{k+1})\hat{W}^*_{k+1}(x_{k+1})\Big]$$
$$-\frac{1}{M}\sum_{j=1}^M \Big[\mathbf{1}_K(x^{i,a,j}_{k+1}) + \mathbf{1}_{A\backslash K}(x^{i,a,j}_{k+1})\hat{W}^*_{k+1}(x^{i,a,j}_{k+1})\Big]\Big| \leq \epsilon_1\Big\} \geq 1 - 2e^{-2M(\epsilon_1)^2},$$

where ϵ_1 is the bound on the error.
We can then provide a lower bound on the probability that the deviation incurred by the quantity $(\hat{\mathsf{T}}\hat{W}^*_{k+1})(x^i_k)$ is bounded by ϵ_1 via the joint probability of $|\mathcal{A}|$ independent events, as follows:

$$\mathbf{P}\Big\{\Big|\mathsf{T}\hat{W}^*_{k+1}(x^i_k) - \hat{\mathsf{T}}\hat{W}^*_{k+1}(x^i_k)\Big| \leq \epsilon_1\Big\}$$
$$\geq \prod_{a\in\mathcal{A}}\mathbf{P}\Big\{\Big|\mathbf{E}^a_{x^i_k}\Big[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\backslash K}(x_{k+1})\hat{W}^*_{k+1}(x_{k+1})\Big]$$
$$-\frac{1}{M}\sum_{j=1}^M\Big[\mathbf{1}_K(x^{i,a,j}_{k+1}) + \mathbf{1}_{A\backslash K}(x^{i,a,j}_{k+1})\hat{W}^*_{k+1}(x^{i,a,j}_{k+1})\Big]\Big| \leq \epsilon_1\Big\}.$$

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:

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

This leads to the following result.

Lemma 6 *For a given error bound ϵ_1 and sample complexity N and M, the estimation error can be probabilistically bounded as follows:*

$$\mathbf{P}\left\{\|\mathsf{T}\hat{W}_{k+1}^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\|_{p,\hat{\eta}} \le \epsilon_1\right\} \ge 1 - \delta_1, \tag{9}$$

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

Notice that for an increasing number of samples M, by Lemma 6 the empirical norm of the error as in (8) is less than ϵ_1 with a probability that increases to 1.

4.1.2 ACCURACY OF THE EMPIRICAL NORM

Let $\mathsf{T}\hat{W}_{k+1}^* : \mathcal{X} \to [0,1]$ be an unknown function and η 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 $\mathsf{T}\hat{W}_{k+1}^*$ with respect to the following *expected loss*:

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

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

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

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:

$$\mathbf{P}\big\{\sup_{w\in\mathcal{W}}\big|\|w - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\eta}^p - \|w-\mathsf{T}\hat{W}_{k+1}^*\|_{p,\hat{\eta}}^p\big| \ge \epsilon_2^p\big\} \le \delta_2.$$

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 $f(x) = |w(x) - \mathsf{T}\hat{W}_{k+1}^*(x)|^p$, the above problem can be framed as the uniform convergence of a standard learning problem (Haussler, 1992; Pollard, 1984; Vapnik, 1998). Resorting to 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 (Bartlett et al., 2005; Hoeffding, 1963). We focus on pseudo dimensions to deal with the capacity (or the complexity) of the function class. By results in (Haussler, 1992) and (Pollard, 1984), we obtain the following uniform convergence bound.

Lemma 7 *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 $(x_k^i)_{1\le i\le N}$ be generated by N independent draws according to any distribution η on $A \setminus K$ and let $p \ge 1$. Then for any $\epsilon_2 > 0$ we have that*

$$\mathbf{P}\left\{\sup_{w\in\mathcal{W}}\big|\|w - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\eta}^p - \|w - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\hat{\eta}}^p\big| \ge \epsilon_2^p\right\} \le 4e(d+1)\left(\frac{32e}{\epsilon_2^p}\right)^d e^{-\frac{N\epsilon_2^{2p}}{128}}. \tag{10}$$

The characterization and computability of the pseudo dimension of a function class is given in Appendix B.

4.1.3 GLOBAL ACCURACY OF SINGLE ITERATIONS

The error bounds introduced in Lemmas 6 and 7, 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 η and $T_x(\cdot \mid x, a)$, they are in fact *distribution-free* bounds (Bartlett et al., 2005).

Theorem 8 *Consider a reach-avoid property defined over a general stochastic system with continuous state space $\mathcal{X}$ and finite action space $\mathcal{A}$. Let $A \subset \mathcal{X}$ and $K \subset \mathcal{X}$ be Borel measurable sets and fix $p \geq 1$, the distribution $\eta \in M(A \setminus K)$ and $\mathcal{W} \subset B(\mathcal{X}; 1)$. Pick any $\hat{W}_{k+1}^* \in B(\mathcal{X}; 1)$ and let $\hat{\mathsf{T}} \hat{W}_{k+1}^*$ and $\hat{W}_k^*$ be calculated using (4) and (5). For given upper bounds on*

$$\mathbf{P}\big\{\|\mathsf{T}\hat{W}_{k+1}^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\|_{p,\hat{\eta}} > \epsilon_1\big\} \leq \delta_1, \ \ and \tag{11}$$

$$\mathbf{P}\big\{\sup_{w \in \mathcal{W}} \big|\|w - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\eta}^p - \|w - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\hat{\eta}}^p\big| > \epsilon_2^p\big\} \leq \delta_2, \tag{12}$$

the bound on a single step update is as follows:

$$\mathbf{P}\Big\{\|\hat{W}_k^* - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\eta} > d_{p,\eta}(\mathsf{T}\hat{W}_{k+1}^*, \mathcal{W}) + \epsilon\Big\} \leq \delta_1 + \delta_2, \tag{13}$$

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

Note that the statement assumes that $\hat{W}_k^*$ is the unique solution to the optimization problem in (5). This strict assumption can be weakened by adding a tolerance term to the theorem. The quantity $d_{p,\eta}(\mathsf{T}\hat{W}_{k+1}^*, \mathcal{W})$ admits the inherent Bellman error $d_{p,\eta}(\mathsf{T}\mathcal{W}, \mathcal{W})$ in (7) as a general upper bound.

4.2 Error Propagation and Global Error Bounds

We express a global bound on the accuracy of the FVI algorithm $\big|\hat{r}_{x_0}^*(K, A) - r_{x_0}^*(K, A)\big|$ by using the probabilistic bounds (derived in Section 4.1) on the errors introduced by the approximate one-step mappings $|(\hat{\mathsf{T}}\hat{W}_1^*)(x_0) - (\mathsf{T}\hat{W}_1^*)(x_0)|$, as well as $\|\hat{W}_1^* - \mathsf{T}\hat{W}_2^*\|_{p,\eta}$, $\|\hat{W}_2^* - \mathsf{T}\hat{W}_3^*\|_{p,\eta}, \dots, \|\hat{W}_{N_t-1}^* - \mathsf{T}\hat{W}_{N_t}^*\|_{p,\eta}$, 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 $\hat{W}_k^*$ from the optimal value function $\mathsf{T}^{N_t-k}W_{N_t}^*$ can be expressed as a function of this deviation at step $k + 1$, plus the approximation error introduced at the $(N_t - k)$th iteration (which has been bounded above). Recall that the optimal value functions can be written as $W_k^* = \mathsf{T}^{N_t-k}W_{N_t}^*$ and that by definition $W_{N_t}^* = \hat{W}_{N_t}^*$.

Lemma 9 *Let η 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*

$$\|\hat{W}_k^* - \mathsf{T}^{N_t-k}\hat{W}_{N_t}^*\|_{p,\eta} \leq \|\hat{W}_k^* - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\eta} + B^{\frac{1}{p}} \Big\|\hat{W}_{k+1}^* - \mathsf{T}^{N_t-(k+1)}\hat{W}_{N_t}^*\Big\|_{p,\eta},$$

for $k = 0, 1, \ldots, N_t - 1$, and where B is defined as

$$B = \sup_{x_{k+1} \in A \setminus K} \int_{A \setminus K} \frac{\max_{a \in \mathcal{A}} t_x\left(x_{k+1} \mid x_k, a\right) \eta(x_k)}{\eta(x_{k+1})} dx_k.$$

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 10 *Consider a reach-avoid problem defined on a Markov process with a continuous state space $\mathcal{X}$ and a finite action space $\mathcal{A}$. 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 Algorithm 2, which has an accuracy of $\boldsymbol{\Delta}$ and a confidence $\delta_{\boldsymbol{\Delta}}$, as stated in (3), if the following holds:*

$$\mathbf{P}\left\{ B_0 \sum_{k=1}^{N_t - 1} B^{\frac{k-1}{p}} \|\hat{W}_k^* - \mathsf{T}\hat{W}_{k+1}^*\|_{p,\eta} \right. \tag{14}$$

$$\left. + |\hat{W}_0^*(x_0) - \mathsf{T}\hat{W}_1^*(x_0)| > \boldsymbol{\Delta} \right\} \leq \delta_{\boldsymbol{\Delta}}, \tag{15}$$

where B is given in (14), B_0 is defined as

$$B_0 = \sup_{x_1 \in A \setminus K} \max_{a \in \mathcal{A}} \frac{t_x\left(x_1 \mid x_0, a\right)}{\eta(x_1)}, \tag{16}$$

and where η 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.

Remark 11 *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 (Munos 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). It B expresses the maximal concentration of the dynamics (relative to η) over the relevant set $A \setminus K$ after one transition starting from distribution η.*

The case study discussed in Section 6 displays a choice of a density function η 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 α, is less than $\boldsymbol{\Delta}$, as per (3). Let us assume that there exist positive values $\epsilon_0, \epsilon_1, \epsilon_2$ such that the approximation error $\boldsymbol{\Delta}$ can be split up into individual bounds based on Theorem 8 and 10 as follows:

$$\boldsymbol{\Delta} = B_0 \sum_{k=1}^{N_t - 1} B^{\frac{k-1}{p}} d_{p,\eta}\left(\mathsf{T}\mathcal{W}, \mathcal{W}\right) + 2B_0 \sum_{k=1}^{N_t - 1} B^{\frac{k-1}{p}}\left(\epsilon_1\right)$$

$$+ 2B_0 \sum_{k=1}^{N_t-1} B^{\frac{k-1}{p}} (\epsilon_2) + \epsilon_0,$$

accounting for, respectively, the inherent Bellman error (Section 4.1); the error on the estimation of $\mathsf{T}\hat{W}_{k+1}^*$ (Section 4.1.1); the error on the empirical norm (Section 4.1.2); and the error related to $|\hat{W}_0^*(x_0) - \mathsf{T}\hat{W}_1^*(x_0)|$. We compute a lower bound on the confidence that the approximation error is bounded by $\mathbf{\Delta}$ 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 Δ is upper bounded with $\delta_{\mathbf{\Delta}}$, as

$$\begin{aligned} \delta_{\mathbf{\Delta}} =& 0 + (N_t - 1)\big(1 - (1 - 2e^{-2M(\epsilon_1)^2})^{|\mathcal{A}|N}\big) \\ &+ (N_t - 1)\Big(4e(d+1)\Big(\tfrac{32e}{\epsilon_2^p}\Big)^d e^{-\frac{N\epsilon_2^{2p}}{128}}\Big) \\ &+ 1 - (1 - 2e^{-2M_0(\epsilon_0)^2})^{|\mathcal{A}|}, \end{aligned} \tag{17}$$

with respectively the inherent Bellman error (Section 4.1), the confidence terms from Lemma 6 with sample complexity M and N, from Lemma 7 with $\mathsf{dim}_p(\mathcal{W}) = d < \infty$, and from Lemma 6 with $M = M_0$ and $N = 1$. Eqn. (17) holds 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 $1 - \alpha > \delta_{\mathbf{\Delta}}$. 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:

$$\begin{aligned} N =& \Big\lceil 128\big(\ln(4e(d+1)) + d\ln(32e)\big)\Big(\tfrac{1}{\epsilon_2}\Big)^{2p} \\ &+ 128dp\Big(\tfrac{1}{\epsilon_2}\Big)^{2p} \ln\Big(\tfrac{1}{\epsilon_2}\Big) + 128\Big(\tfrac{1}{\epsilon_2}\Big)^{2p} \ln\Big(\tfrac{1}{\delta_2}\Big)\Big\rceil, \end{aligned} \tag{18a}$$

$$M = \Big\lceil \tfrac{1}{2}\Big(\tfrac{1}{\epsilon_1}\Big)^2 \Big(\ln(2|\mathcal{A}|) + \ln(\tfrac{1}{\delta_1}) + \ln(N)\Big)\Big\rceil, \tag{18b}$$

$$M_0 = \Big\lceil \tfrac{1}{2}\Big(\tfrac{1}{\epsilon_0}\Big)^2 \Big(\ln\big(2|\mathcal{A}|\big) + \ln\big(\tfrac{1}{\delta_0}\big)\Big)\Big\rceil, \tag{18c}$$

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

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.

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 the function class $\mathcal{W}$ 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 η 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 η 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 $\delta_0, \delta_1, \delta_2$, e.g. as in (18a).

With regards to the single-step errors, Lemmas 6 and 7 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.

In conclusion, 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 $\hat{W}_k^*$ 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 $\hat{W}_k^*$, in other words if the estimated value functions $\hat{W}_k^*$ 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 $\mathbf{P}\left\{\ |\ \hat{r}_{x_0}^*(K, A) - r_{x_0}^*(K, A)\ |\ > \mathbf{\Delta}\ \right\} \leq \delta_{\mathbf{\Delta}}$, as given in (3), and computed now in a sample-based manner, includes an empirical estimate of the quantity $|\ \hat{r}_{x_0}^*(K, A) - r_{x_0}^*(K, A)\ |$. This sample-based estimate is computed as follows:

a. collect samples $(x_i)_{1 \leq i \leq \tilde{N}}$ according to (1) in Algorithm1 (with $N = \tilde{N}$), and subsequently use (2) with $M = \tilde{M}$ to draw both $(y_1^{i,a,j})_{1 \leq j \leq \tilde{M}}$ and $(y_2^{i,a,j})_{1 \leq j \leq \tilde{M}}$;

b. estimate the single step error (19) for each k;

c. estimate the bias (21) 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

$$\left\|\hat{W}_k^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\right\|_{1,\tilde{\eta}} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \left|\hat{W}_k^*(x^i) - \max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x^i)\right| \tag{19}$$

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 η resulting from $(x_i)_{1 \leq i \leq \tilde{N}}$.

Let the operator $\hat{\mathsf{T}}_\alpha^a$ be, for $\alpha = 1, 2$,

$$\hat{\mathsf{T}}_\alpha^a \hat{W}_{k+1}^*(x^i) = \frac{1}{\tilde{M}}\sum_{j=1}^{\tilde{M}} \mathbf{1}_K(y_\alpha^{i,a,j}) + \mathbf{1}_{A\setminus K}(y_\alpha^{i,a,j})\hat{W}_{k+1}^*(y_\alpha^{i,a,j}). \tag{20}$$

The estimate in (19) 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 $(y_1^{i,a,j})_{1\leq j\leq \tilde{M}}$ and $(y_2^{i,a,j})_{1\leq j\leq \tilde{M}}$, for each x^i and a, allows us to estimate this bias for $1 \leq k \leq N_t - 1$ as

$$\left\| \max_{a\in\mathcal{A}} \left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^* - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^* \right| \right\|_{1,\tilde{\eta}}. \tag{21}$$

In the following theorem, an expression for the bound on $\mathbf{P}\{\,|\,\hat{r}_{x_0}^*(K,A) - r_{x_0}^*(K,A)\,|\, > \Delta\,\} \leq \delta_\Delta$ is derived, employing the error propagation technique first used in Section 4.2, the estimates of the single step error above, and the bias (21) in combination with Hoeffding's inequality (Hoeffding, 1963).

Theorem 12 *Consider a reach-avoid problem defined on a Markov process with a continuous state space $\mathcal{X}$ and a finite action space $\mathcal{A}$. 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 Algorithm 2, which has an accuracy of Δ and a confidence δ_Δ, as stated in (3), if the following holds:*

$$\Delta = B_0 \sum_{k=1}^{N_t-1} B^{k-1}\left(\left\| \hat{W}_k^* - \hat{\mathsf{T}}\hat{W}_{k+1}^* \right\|_{1,\tilde{\eta}} + \left\| \max_{a\in\mathcal{A}} \left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^* - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^* \right| \right\|_{1,\tilde{\eta}} \right) + B_0\epsilon + \epsilon_0, \tag{22a}$$

$$\delta_\Delta = e^{-2\frac{\tilde{N}\epsilon^2}{L^2}} - \delta_0, \tag{22b}$$

with $L = 2\sum_{k=1}^{N_t-1} B^{k-1}$, and sample sizes $\tilde{M}$ and $\tilde{N}$ according to the sample sets drawn according to the distribution η. Equation (22a) includes the estimated error as a combination of (19) and (21). The scaling factors B and B_0 are computed as in (14) and (16) for the same sampling distribution η. The factors δ_0 and ϵ_0 are computed as in Lemma 6 with $M = M_0$ and $N = 1$.

The accuracy Δ 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 δ_Δ.

Suppose that a close-to-optimal policy is given, for example a policy as detailed in Remark 5 and computed from the series of estimated value functions $\hat{W}_k^*$. Then we know that $r_{x_0}^*(K,A) \geq r_{x_0}^{\hat{\mu}^*}(K,A)$, therefore a lower bound on the value of $r_{x_0}^{\hat{\mu}^*}(K,A)$ is also a lower bound on $r_{x_0}^*(K,A)$. Note that for a policy μ, the closed-loop Markov process is time dependent. This allows us to estimate $r_{x_0}^\mu(K,A)$ directly from traces of this autonomous Markov process. The deviation of this empirical mean can be bounded probabilistically using Hoeffding's inequality. Additionally an upper bound on the deviation $|\hat{r}_{x_0}^*(K,A) - r_{x_0}^\mu(K,A)|$ can be computed. The combination of the bound in Theorem 12 and of the bound on $|\hat{r}_{x_0}^*(K,A) - r_{x_0}^\mu(K,A)|$ provide a bound on the performance deviation of $r_{x_0}^\mu(K,A)$: the

triangle inequality leads to $|r_{x_0}^*(K, A) - r_{x_0}^\mu(K, A)| \leq |r_{x_0}^*(K, A) - \hat{r}_{x_0}^*(K, A)| + |\hat{r}_{x_0}^*(K, A) - r_{x_0}^\mu(K, A)|$.

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 are expected to scale better than those used for grid-based approaches such as (Esmaeil Zadeh Soudjani and Abate, 2013).

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 smaller target range within a given finite time horizon. The temperature can be affected using local heaters. 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 $\mathcal{X} = \mathbb{R}^2$, and where the possible configurations $\{OFF, ON\} = \{0, 1\}$ of the two heaters form the finite action space $\mathcal{A}$. Hence $\mathcal{A} = \{0, 1\} \times \{0, 1\}$, 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 $a = [1\ 0]^T \in \mathcal{A}$. The dynamics at discrete time k is characterized by the following stochastic difference equation:

$$x_{k+1} = \mathbf{A}x_k + \mathbf{B}a + \mathbf{C} + n_k, \text{ where} \tag{23}$$
$$\mathbf{A} = \begin{bmatrix} 1-b_1-a_{1,2} & a_{1,2} \\ a_{2,1} & 1-b_2-a_{2,1} \end{bmatrix}, \mathbf{C} = \begin{bmatrix} b_1 x_a \\ b_2 x_a \end{bmatrix}, \text{ and } \mathbf{B} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix},$$

and with the following parameters: 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 parameters re instantiated as $b_1 = 0.0375$, $c_1 = 0.65$, $x_a = 6$, $b_2 = 0.025$, $c_2 = 0.6$, and $a_{ij} = 0.0625$. 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 $\mathcal{N}(\cdot \mid \mu, \Sigma)$ be a 2-dimensional multivariate normal distribution over $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with mean μ and covariance matrix Σ, then the stochastic kernel T_x is given as

$$T_x(\cdot \mid x, a) = \mathcal{N}\left(\cdot \mid \mathbf{A}x + \mathbf{B}a + \mathbf{C}, \nu^2 I_{2\times 2}\right) \tag{24}$$

and characterises the probability distribution of the stochastic transitions in (23). The stochastic kernel (24) admits the probability density

$$t_x\left(y \mid x, a\right) = \frac{1}{\sqrt{|\Sigma|(2\pi)^2}} e^{\left(-\frac{1}{2}(y-\bar{\mu})^T \Sigma^{-1}(y-\bar{\mu})\right)}, \tag{25}$$

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 $\bar{\mu} = \mathbf{A}x + \mathbf{B}a + \mathbf{C}$.

6.2 Application of the Fitted Value Iteration Algorithm

The FVI scheme is implemented as in Algorithm 2, and approximates the solution of the reach-avoid problem. We obtain an approximation of $r_{x_0}^*(K, A) = \mathsf{T}^{10}\hat{W}_{10}^*(x_0)$ by $\hat{\mathsf{T}}\hat{W}_1^*(x_0)$, while using the auxiliary functions $\hat{W}_9^*, \dots, \hat{W}_1^*$ to approximate $\mathsf{T}\hat{W}_{10}^*, \dots, \mathsf{T}\hat{W}_2^*$ in the FVI scheme – equivalently, function $\hat{W}_k^*$ approximates $\mathsf{T}^{N_t-k}\hat{W}_{N_t}^*$. For a given temperature x_k at time instant k, the function $\hat{W}_k^*(x_k)$ gives the approximate probability that the consecutive temperature values $x_{k+1}, \dots, x_{N_t}$ 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 select a uniform distribution η over $A \setminus K$ to sample from, then select a function class $\mathcal{W}$ and a value for $p \geq 1$ to solve (5). We consider a function class $\mathcal{W}$ 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 (5) as a least-square problem (with $p = 2$). A 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 $\hat{W}_9^*, \hat{W}_5^*$, and $\hat{W}_1^*$ are displayed in Fig. 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.

Fig. 2 displays a suboptimal policy $\hat{\mu}^*$ that is obtained via the FVI algorithm as discussed in Remark 5, by employing the tree classification method `ClassificationTree.fit` of `Matlab`. Observe that policy $\hat{\mu}_9^*$ for $k = 9$ is not accurate over the flat regions of $\hat{W}_9^*$ (corresponding to the blue spots in the left side of Fig. 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. Fig. 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.
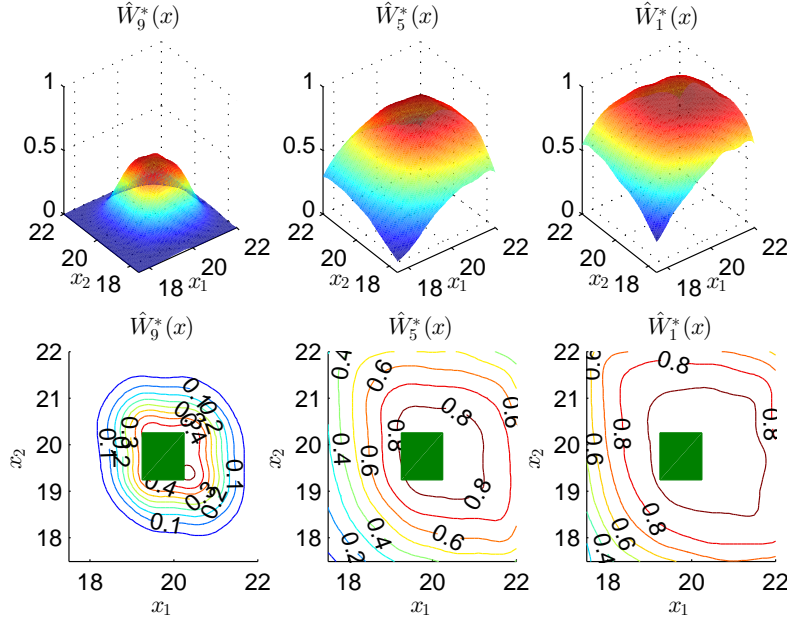
Figure 1: Function approximations $\hat{W}_9^*, \hat{W}_5^*$, and $\hat{W}_1^*$ – level sets (top) and contour plots (bottom). The function approximations of the value functions are obtained with the FVI algorithm as in Algorithm 2, using a radial basis function (RBF) neural network with 50 radial basis functions and a given width of 0.7, a uniform sampling distribution η over $A \setminus K$ and $M = 10^3$ and $N = 600$. The safe set A is $[17.5\ 22]^2$, and reach set $K = [19.25\ 20.25]^2$. The $\hat{W}_9^*, \hat{W}_5^*$, and $\hat{W}_1^*$ approximate $\mathsf{T}W_{10}^*, \mathsf{T}^5 W_{10}^*$, and $\mathsf{T}^9 W_{10}^*$ over the set $A \setminus K$. In the contour plots (bottom) the green squares denote set K.

Table 1: Approximate solutions $\hat{r}_{x_0}^*(K, A)$ are given in the table for several initial conditions x_0, together with the related sub-optimal action at the initial time.

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

6.3 Performance of the Fitted 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. Fig. 3 plots the sample-based estimates of the single step error (19), namely $\|\hat{W}_k^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\|_{1,\tilde{\eta}}$ and of the bias (21), namely
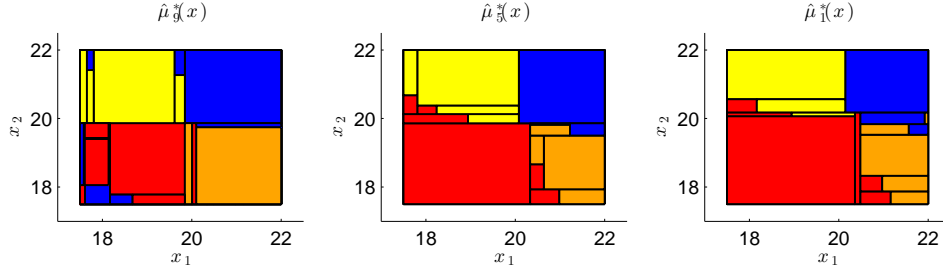
Figure 2: The policy $\hat{\mu}^*$ for $k = 9, 5, 1$ obtained form the same computations as in Fig. 1. The action (ON,ON) is labeled in red, (OFF, ON) is orange, (ON,OFF) is yellow, and (OFF,OFF) is blue.
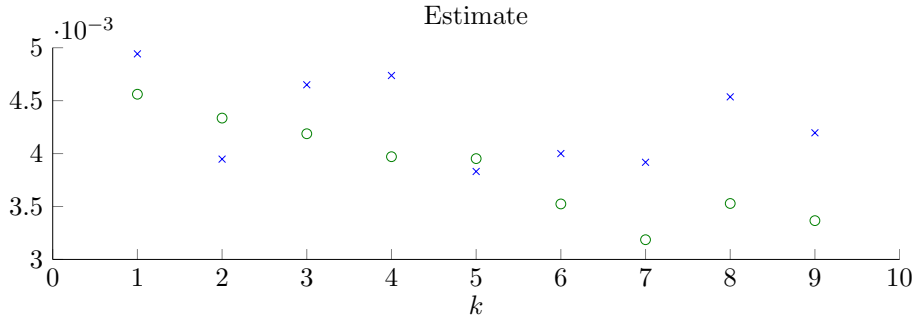

Figure 3: Sample-based estimation of the single step error $\|\hat{W}_k^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\|_{1,\tilde{\eta}}$ ($\times$) and $\left\|\max_{a \in \mathcal{A}}\left|\hat{\mathsf{T}}_1^a\hat{W}_{k+1}^* - \hat{\mathsf{T}}_2^a\hat{W}_{k+1}^*\right|\right\|_{1,\tilde{\eta}}$ ($\circ$) at each iteration for $k = 9$ until $k = 1$. An independent set of samples of size $\tilde{N} = 4 \cdot 10^3$ and $\tilde{M} = 10^4$ has been used.

$\left\|\max_{a \in \mathcal{A}}\left|\hat{\mathsf{T}}_1^a\hat{W}_{k+1}^* - \hat{\mathsf{T}}_2^a\hat{W}_{k+1}^*\right|\right\|_{1,\tilde{\eta}}$. Observe that the values of both (19) and (21) fall in the interval between 3×10^{-3} and 5×10^{-3}. The bias estimate (21) appears distributed all over this interval, whereas there is a noticeable trend in the plot of (19), which suggests that the first iterations can be fitted more easily. In Fig. 4, the accuracy of the FVI algorithm propagated over the iterations is given, starting from the first iteration $\|\hat{W}_9^* - \mathsf{T}W_{N_t}^*\|_\eta$ until the last iteration $\|\hat{W}_1^* - \mathsf{T}^{N_t}W_{N_t}^*\|_\eta$. This accuracy is computed using Theorem 12. The estimates in Fig. 3 are used to compute the estimate of the accuracy $\|\hat{W}_k^* - \mathsf{T}W_{N_t}^*\|_\eta$ and the accuracy $\boldsymbol{\Delta}$ for a given $\delta_{\boldsymbol{\Delta}}$. For each iteration step, it can be observed in Fig. 3 that the error caused by estimating the dynamic programming operator T and by fitting a function is relatively small ($< 10^{-2}$). However, the error grows exponentially over the whole horizon: as expected, the accuracy of the algorithm depends strongly on B, which has been computed numerically and amounts to 3.07.

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 of the known "reach-avoid" specification. The approximate computational scheme is based on the Fitted Value Iteration
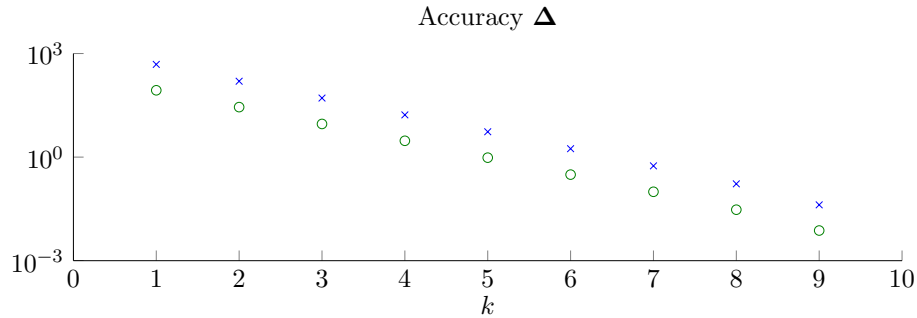
Figure 4: The plot shows the accuracy based on the error propagation over the iterations estimated with an hold out set of size $\tilde{N} = 4 \cdot 10^3$ and $\tilde{M} = 10^4$. The accuracy $\mathbf{\Delta}$, marked as ($\times$), is given for $1 - \delta_{\mathbf{\Delta}} = 0.9$. In the graph, the estimate of the accuracy, $\|\hat{W}_k^* - \mathsf{T}^{N_t - k} W_{N_t}^*\|$, is given as ($\circ$).

algorithm, which hinges on random sample extractions.

We 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. 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.

References

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 2008.

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. *Stochastic Hybrid Systems.* Number 24 in Control Engineering. CRC Press, Boca Raton, 2006.

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.

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.

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.

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.

Leonid Peshkin and Sayan Mukherjee. Bounds on sample size for policy evaluation in Markov environments. In *Computational Learning Theory*, pages 616–629. Springer, 2001.

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

Maria Prandini and Jianghai Hu. Stochastic reachability: Theory and numerical approximation. *Stochastic hybrid systems, Automation and Control Engineering Series*, 24:107–138, 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 6: Bound on Estimation Error

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

$$\mathbf{P}\left\{\left|\sum_{i=1}^{N} x_k^i - \mathbf{E}\sum_{i=1}^{N} X_i\right| \geq N\epsilon\right\} \leq 2e^{-2N\epsilon^2},$$

where $\mathbf{E}\sum_{i=1}^{N} X_i$ is the mean of the random variable $\sum_i X_i$, whereas the empirical mean is defined as $\sum_{i=1}^{N} x_k^i$, where x_k^i is a realization of X_i.

Using Proposition 13 the proof of Lemma 6 is provided as follows.

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

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

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

$$\mathbf{P}\left\{\|\mathsf{T}\hat{W}_{k+1}^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\|_{p,\hat{\eta}} \leq \epsilon_1\right\} = \mathbf{P}\left\{\|\mathsf{T}\hat{W}_{k+1}^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\|_{p,\hat{\eta}}^p \leq \epsilon_1^p\right\}$$

via definition of the empirical norm in (8)

$$= \mathbf{P}\left\{\frac{1}{N}\sum_{i=1}^{N}\left|\mathsf{T}\hat{W}_{k+1}^*(x_k^i) - \hat{\mathsf{T}}\hat{W}_{k+1}^*(x_k^i)\right|^p \leq \epsilon_1^p\right\}$$

Note the mutual independence between the sample sets at different base points x_k^i given as $\bigcup_{a\in\mathcal{A}}\left(x_{k+1}^{i,a,j}\right)_{1\leq j\leq M}$

$$\geq \mathbf{P}\left\{\bigcap_{i=1}^{N}\left\{\left|\mathsf{T}\hat{W}_{k+1}^*\left(x_k^i\right) - \hat{\mathsf{T}}\hat{W}_{k+1}^*(x_k^i)\right|^p \leq \epsilon_1^p\right\}\right\} = \prod_{i=1}^{N}\mathbf{P}\left\{\left|\mathsf{T}\hat{W}_{k+1}^*\left(x_k^i\right) - \hat{\mathsf{T}}\hat{W}_{k+1}^*(x_k^i)\right| \leq \epsilon_1\right\}.$$

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

$$\left|\mathsf{T}\hat{W}_{k+1}^*\left(x_k^i\right) - \hat{\mathsf{T}}\hat{W}_{k+1}^*(x_k^i)\right|$$

$$= \left|\max_{a\in\mathcal{A}}\mathbf{E}_{x_{k+1}}\left[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\backslash K}(x_{k+1})\hat{W}_{k+1}^*(x_{k+1})\right]\right.$$

$$\left. - \max_{a\in\mathcal{A}}\frac{1}{M}\sum_{j=1}^{M}[\mathbf{1}_K(x_{k+1}^{i,a,j}) + \mathbf{1}_{A\backslash K}(x_{k+1}^{i,a,j})\hat{W}_{k+1}^*(x_{k+1}^{i,a,j})]\right|$$

W.r.t $\mathbf{E}_{x_{k+1}}$ defined over random variable $x_{k+1} \sim T_x\left(\cdot \mid x_k^i, a\right)$

$$\leq \max_{a\in\mathcal{A}}\left|\mathbf{E}_{x_{k+1}}\left[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\backslash K}(x_{k+1})\hat{W}_{k+1}^*(x_{k+1})\right]\right.$$

$$\left. - \frac{1}{M}\sum_{j=1}^{M}[\mathbf{1}_K(x_{k+1}^{i,a,j}) + \mathbf{1}_{A\backslash K}(x_{k+1}^{i,a,j})\hat{W}_{k+1}^*(x_{k+1}^{i,a,j})]\right|.$$

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:

$$\mathbf{P}\left\{\left|\mathsf{T}\hat{W}^*_{k+1}\left(x^i_k\right) - \hat{\mathsf{T}}\hat{W}^*_{k+1}(x^i_k)\right| \leq \epsilon_1\right\}$$
$$\geq \prod_{a\in\mathcal{A}} \mathbf{P}\left\{\left|\mathbf{E}_{x_{k+1}}\left[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\setminus K}(x_{k+1})\hat{W}^*_{k+1}(x_{k+1})\right]\right.\right.$$
$$\left.\left. - \frac{1}{M}\sum_{j=1}^M \left[\mathbf{1}_K(x^{i,a,j}_{k+1}) + \mathbf{1}_{A\setminus K}(x^{i,a,j}_{k+1})\hat{W}^*_{k+1}(x^{i,a,j}_{k+1})\right]\right| \leq \epsilon_1\right\}.$$

For a given base point $x^i_k \in \mathcal{X}$, action $a \in \mathcal{A}$, and function $\hat{W}^*_{k+1} \in \mathcal{W}$, define random variables Z_j via their realizations $\mathbf{1}_K(x^{i,a,j}_{k+1}) + \mathbf{1}_{A\setminus K}(x^{i,a,j}_{k+1})\hat{W}^*_{k+1}(x^{i,a,j}_{k+1})$, with $j = 1, \ldots, M$. Since each $x^{i,a,j}_{k+1}$ is independently drawn from $T_x\left(\cdot \mid x^i_k, a\right)$, 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 13), the concentration of the M samples around the expected value of Z_j can be expressed as

$$\mathbf{P}\left\{\left|\mathbf{E}_{x_{k+1}}\left[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\setminus K}(x_{k+1})\hat{W}^*_{k+1}(x_{k+1})\right]\right.\right.$$
$$\left.\left. - \frac{1}{M}\sum_{j=1}^M \left[\mathbf{1}_K(x^{i,a,j}_{k+1}) + \mathbf{1}_{A\setminus K}(x^{i,a,j}_{k+1})\hat{W}^*_{k+1}(x^{i,a,j}_{k+1})\right]\right| \epsilon_1\right\} \leq 2e^{-2M(\epsilon_1)^2}.$$

Therefore as long as $0 \leq 2e^{-2M(\epsilon_1)^2} \leq 1$, it follows that $\mathbf{P}\left\{\|\mathsf{T}\hat{W}^*_{k+1} - \hat{\mathsf{T}}\hat{W}^*_{k+1}\|_{p,\hat{\eta}} \leq \epsilon_1\right\} \geq \left(1 - 2e^{-2M(\epsilon_1)^2}\right)^{N|\mathcal{A}|}$.

■

Remark 14 *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 Bienaym-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 $T_x\left(\cdot \mid x, a\right)$.*

Appendix B. Proof of Lemma 7

We derive a general, analytical bound on the error of a single backward recursion using notions from statistical learning theory (Haussler, 1992; Pollard, 1984). The error bound takes into account that, for any $\mathsf{T}\hat{W}^*_{k+1}$, 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 η over the set $A\setminus K$.

We exclusively consider function classes $\mathcal{W} \subset B(\mathcal{X}; 1)$ 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 $\mathcal{W}$ to the fit a set of samples.

Proof In order to prove Lemma 7, we show that the inequality in (10) holds for any $\hat{W}_{k+1}^* \in \mathcal{W}$ 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 $\vec{x} = (x^1, \ldots, x^N)$ as a sequence of N independent realizations drawn from a distribution over $A \setminus K$ with density η.

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

$$\mathbf{P}\Big\{ \sup_{w \in \mathcal{W}} \big| \|w - \mathsf{T}W\|_{p,\eta}^p - \|w - \mathsf{T}W\|_{p,\hat{\eta}}^p \big| \geq \epsilon_2^p \Big\} = \mathbf{P}\Big\{ \sup_{l_w \in l_\mathcal{W}} \big| \mathbf{E}_\eta [l_w] - \frac{1}{N} \sum_{i=1}^N l_w(x^i) \big| \geq \epsilon_2^p \Big\},$$

where $\mathbf{E}_\eta$ denotes the expected value with respect to η. 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 $((l_\mathcal{W})_{\vec{x}}, \|\cdot\|_1)$.

Proposition 15 (Pollard (1984)) *Let $\mathcal{F}$ be a permissable 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 $\vec{x} = (x_1, \ldots, x_N)$ be a sequence of N samples drawn independently from $\mathcal{X}$ according to any distribution on $\mathcal{X}$. Then for all $\epsilon > 0$*

$$\mathbf{P}\Big\{ \forall f \in \mathcal{F} : \big| \mathbf{E}f - \frac{1}{N} \sum_{\vec{x}} f(x_i) \big| \geq \epsilon \Big\} \leq 4 \mathbf{E}\mathcal{N}(\epsilon/16, F_{|\vec{x}}, \|\cdot\|_1) e^{-\frac{N\epsilon^2}{128K^2}}, \tag{26}$$

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, ρ) and a subset S of A, we say that the set $T \subseteq A$ is an ϵ-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 ϵ-cover of S.

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

$$(l_\mathcal{W})_{|\vec{x}} = \{(l_w)_{|\vec{x}} = (l_w(x^1), l_w(x^2), \ldots, l_w(x^N)), l_w \in l_\mathcal{W}\} \subseteq [0,1]^N.$$

The minimal ϵ-cover of $((l_\mathcal{W})_{|\vec{x}}, \|\cdot\|_1)$ is denoted as $\mathcal{N}(\epsilon, (l_\mathcal{W})_{|\vec{x}}, \|\cdot\|_1)$.

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

$$\mathbf{P}\Big\{ \sup_{l_w \in l_\mathcal{W}} \big| \mathbf{E}_x [l_w(x)] - \frac{1}{N} \sum_{i=1}^N l_w(x_i) \big| \geq \epsilon_2^p \Big\} \leq 4\mathbf{E} \Big[\mathcal{N}(\epsilon_2^p/16, (l_\mathcal{W})_{|\vec{x}}, \|\cdot\|_1) \Big] e^{-\frac{N(\epsilon_2)^{2p}}{128}}.$$

The expected value of $\mathcal{N}(\epsilon_2^p/16, (l_{\mathcal{W}})_{|\vec{x}}, \|\cdot\|_1)$ is computed over the samples x^i of $\vec{x}$, drawn independently from a probability distribution with density η. Since there is a trivial isometry (Haussler, 1992) between $(l_{\mathcal{W}|\vec{x}}, \|\cdot\|_1)$ and $(l_{\mathcal{W}}, \|\cdot\|_{1,\hat{\eta}})$, both spaces have equal covering numbers

$$\mathcal{N}(\epsilon_2^p/16, l_{\mathcal{W}|\vec{x}}, \|\cdot\|_1) = \mathcal{N}(\epsilon_2^p/16, l_{\mathcal{W}}, \|\cdot\|_{1,\hat{\eta}}).$$

In practice a value for $\mathbf{E}\left[\mathcal{N}(\epsilon_2^p/16, l_{\mathcal{W}}, \|\cdot\|_{1,\hat{\eta}})\right]$ can be obtained by upper bounding $\mathcal{N}(\epsilon_2^p/16, l_{\mathcal{W}}, \|\cdot\|_{1,\hat{\eta}})$ independently of the sample distribution. For this we introduce the pseudo dimension of a function class, formally defined as follows Pollard (1984); Anthony (2002); Haussler (1992). Suppose $\mathcal{F}$ is a class of functions, $f \in \mathcal{F}, f : \mathcal{X} \to [0,1]$. Then $S \subseteq \mathcal{X}$ is shattered by $\mathcal{F}$ if there are numbers $r_x \in [0,1]$ for $x \in \mathcal{S}$ such that for every $T \subseteq S$ there is some $f_T \in \mathcal{F}$ 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 $\mathcal{F}$ has a finite pseudo dimension $\mathsf{dim}_p(\mathcal{F}) = d$ 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 tho covering number of the metric space $(l_{\mathcal{W}}, \|\cdot\|_{1,P})$ can be upper bounded as a function of the pseudo-dimension and the base of the natural logarithm e: for any $\epsilon > 0$,

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

We have proved that a sufficient upper bound is given as

$$\mathbf{P}\left\{\sup_{w \in \mathcal{W}}\left|\|w - \mathsf{T}W\|_{p,\eta}^p - \|w - \mathsf{T}W\|_{p,\hat{\eta}}^p\right| \geq \epsilon_2^p\right\} \leq 4e(d+1)\left(\frac{32e}{\epsilon_2^p}\right)^d e^{-\frac{N(\epsilon_2)^{2p}}{128}}.$$

The proof can be concluded by showing that the pseudo dimension d of the induced class $l_{\mathcal{W}}$ is the same as the pseudo dimension of $\mathcal{W}$. Let $\{w - \mathsf{T}W : w \in \mathcal{W}\}$ be a new function class induced from $\mathcal{W}$. The invariance properties of the pseudo dimension $\dim_p(\mathcal{W})$ shown in (Haussler, 1992) allow to conclude that $\dim_p(\{w - \mathsf{T}W | w \in \mathcal{W}\}) = \dim_p(\mathcal{W})$. The induced function class $l_{\mathcal{W}}$ can then be defined as follows: $l_{\mathcal{W}} = \{|k|^p | k \in \{w - \mathsf{T}W : w \in \mathcal{W}\}\}$. 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 $\dim_p(l_{\mathcal{W}}) = \dim_p(\{w - \mathsf{T}W : w \in \mathcal{W}\}) = \dim_p(\mathcal{W}) = d$. ∎

Remark 16 (Computing the pseudo-dimension) *When the function class $\mathcal{W}$ 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 $l_{\mathcal{W}}$ (as introduced in the proof) by the pseudo dimension of $\mathcal{W}$, 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 $l_{\mathcal{W}}$ 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 15), 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 (15) 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 8

The proof of Theorem 8 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 $\hat{W}_{k+1}^*$ by replacing it with a general function $W' \in \mathcal{W}$ that minimizes the empirical norm as $W' = \arg\min_{w \in \mathcal{W}} \|w - \hat{\mathsf{T}}W\|_{p,\hat{\eta}}$. Let us further define a space Ω for the batch of samples drawn at any of the iterations, such that at any instant k the realized sample batch $\omega := \bigcup_{i \in \{1,\dots,N\}} \left(x_k^i \cup \left(\bigcup_{a \in \mathcal{A}} \left(x_{k+1}^{i,a,j} \right)_{1 \leq j \leq M} \right) \right)$ is an element of the sample space, $\omega \in \Omega$.

For any given $\epsilon' > 0$, consider a function $w^* \in \mathcal{W}$ such that $\|w^* - \mathsf{T}W\|_{p,\eta} \leq \inf_{w \in \mathcal{W}} \|w - \mathsf{T}W\|_{p,\eta} + \epsilon'$ (this in particular holds since $\mathcal{W}$ has been assumed to be close and bounded).

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

$$\|W' - \mathsf{T}W\|_{p,\eta} \leq \|W' - \mathsf{T}W\|_{p,\hat{\eta}} + \epsilon_2 \tag{27a}$$
$$\leq \|W' - \hat{\mathsf{T}}W\|_{p,\hat{\eta}} + \epsilon_1 + \epsilon_2 \tag{27b}$$
$$\leq \|w^* - \hat{\mathsf{T}}W\|_{p,\hat{\eta}} + \epsilon_1 + \epsilon_2 \tag{27c}$$
$$\leq \|w^* - \mathsf{T}W\|_{p,\hat{\eta}} + 2\epsilon_1 + \epsilon_2 \tag{27d}$$
$$\leq \|w^* - \mathsf{T}W\|_{p,\eta} + 2\epsilon_1 + 2\epsilon_2. \tag{27e}$$

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

$$\|W' - \mathsf{T}W\|_{p,\eta} \leq \inf_{w \in \mathcal{W}} \|w - \mathsf{T}W\|_{p,\eta} + 2\epsilon_1 + 2\epsilon_2 + \epsilon'.$$

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

$$\|W' - \mathsf{T}W\|_{p,\eta} \leq d_{p,\eta}(\mathsf{T}W, \mathcal{W}) + 2\epsilon_1 + 2\epsilon_2.$$

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

occurrence of the five inequalities in (27). We first show that the third inequality is always true. Then we give the probability associated to the first inequality (27a) and the fifth (27e) (this is based on (12)). Afterwards we provide an upper bound on the probability associated to the second and fourth inequalities (27b),(27d), based on the bound given in (11).

The third inequality (27c) is true for the whole sample space Ω due to the choice of W'. For all functions w in $\mathcal{W}$ it follows that $\|W' - \hat{\mathsf{T}}W\|_{p,\hat{\eta}} \leq \|w - \hat{\mathsf{T}}W\|_{p,\hat{\eta}}$ holds, because $W' = \arg\min_{w \in \mathcal{W}} \|w - \hat{\mathsf{T}}W\|_{p,\hat{\eta}}$.

The first and last inequalities (27a),(27e) bound the deviation between the empirical loss and the expected loss. This can be bounded with the worst case error. Firstly we observe that the inequality

$$\left| \|w - \mathsf{T}W\|_{p,\hat{\eta}} - \|w - \mathsf{T}W\|_{p,\eta} \right|^p \leq \left| \|w - \mathsf{T}W\|_{p,\eta}^p - \|w - \mathsf{T}W\|_{p,\hat{\eta}}^p \right|$$

is always true. In the case that $\|w - \mathsf{T}W\|_{p,\hat{\eta}} \leq \|w - \mathsf{T}W\|_{p,\eta}$ then

$$\left| \|w - \mathsf{T}W\|_{p,\hat{\eta}} - \|w - \mathsf{T}W\|_{p,\eta} \right| = \|w - \mathsf{T}W\|_{p,\hat{\eta}} - \|w - \mathsf{T}W\|_{p,\eta}$$

$$\|w - \mathsf{T}W\|_{p,\hat{\eta}} = (\|w - \mathsf{T}W\|_{p,\hat{\eta}} - \|w - \mathsf{T}W\|_{p,\eta}) + \|w - \mathsf{T}W\|_{p,\eta}$$

$$\|w - \mathsf{T}W\|_{p,\hat{\eta}}^p = ((\|w - \mathsf{T}W\|_{p,\hat{\eta}} - \|w - \mathsf{T}W\|_{p,\eta}) + \|w - \mathsf{T}W\|_{p,\eta})^p$$

$$\|w - \mathsf{T}W\|_{p,\hat{\eta}}^p \geq (\|w - \mathsf{T}W\|_{p,\hat{\eta}} - \|w - \mathsf{T}W\|_{p,\eta})^p + \|w - \mathsf{T}W\|_{p,\eta}^p$$

$$\|w - \mathsf{T}W\|_{p,\hat{\eta}}^p - \|w - \mathsf{T}W\|_{p,\eta}^p \geq (\|w - \mathsf{T}W\|_{p,\hat{\eta}} - \|w - \mathsf{T}W\|_{p,\eta})^p$$

$$\left| \|w - \mathsf{T}W\|_{p,\hat{\eta}}^p - \|w - \mathsf{T}W\|_{p,\eta}^p \right| \geq \left| \|w - \mathsf{T}W\|_{p,\hat{\eta}} - \|w - \mathsf{T}W\|_{p,\eta} \right|^p$$

On the other hand, for the case when $\|w - \mathsf{T}W\|_{p,\hat{\eta}} > \|w - \mathsf{T}W\|_{p,\eta}$ a similar argument can be used. We can then observe that

$$\left| \|w^* - \mathsf{T}W\|_{p,\hat{\eta}} - \|w^* - \mathsf{T}W\|_{p,\eta} \right|^p \leq \sup_{w \in \mathcal{W}} \left| \|w - \mathsf{T}W\|_{p,\eta}^p - \|w - \mathsf{T}W\|_{p,\hat{\eta}}^p \right|,$$

and that

$$\left| \|W' - \mathsf{T}W\|_{p,\eta} - \|W' - \mathsf{T}W\|_{p,\hat{\eta}} \right|^p \leq \sup_{w \in \mathcal{W}} \left| \|w - \mathsf{T}W\|_{p,\eta}^p - \|w - \mathsf{T}W\|_{p,\hat{\eta}}^p \right|.$$

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

$$A_1: \ \epsilon_2^p < \left| \|w^* - \mathsf{T}W\|_{p,\hat{\eta}} - \|w^* - \mathsf{T}W\|_{p,\eta} \right|^p, \quad A_2: \ \epsilon_2^p < \left| \|W' - \mathsf{T}W\|_{p,\eta} - \|W' - \mathsf{T}W\|_{p,\hat{\eta}} \right|^p.$$

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

$$B: \ \epsilon_2^p < \sup_{w \in \mathcal{W}} \left| \|w - \mathsf{T}W\|_{p,\eta}^p - \|w - \mathsf{T}W\|_{p,\hat{\eta}}^p \right|.$$

Thus it follows that for any $\epsilon_2 > 0$: $\mathbf{P}\{A_1 \cup A_2\} \leq \mathbf{P}\{B\}$ and, based on (12), we have

$$\mathbf{P}\left\{ \left\{ \left| \|W' - \mathsf{T}W\|_{p,\eta} - \|W' - \mathsf{T}W\|_{p,\hat{\eta}} \right| > \epsilon_2 \right\} \cup \left\{ \left| \|w^* - \mathsf{T}W\|_{p,\hat{\eta}} - \|w^* - \mathsf{T}W\|_{p,\eta} \right| > \epsilon_2 \right\} \right\}$$

$$\leq \mathbf{P}\left\{ \sup_{w \in \mathcal{W}} \left| \|w - \mathsf{T}W\|_{p,\eta}^p - \|w - \mathsf{T}W\|_{p,\hat{\eta}}^p \right| > \epsilon_2^p \right\} \leq \delta_2.$$

Thus the probability that the inequalities (27a) and (27e) do not hold is less then δ_2.

The second and fourth inequalities (27b),(27d) depend the accuracy of the estimation of the backward recursion at each base point x_k^i. Employing the inequality $\big| \|w - g\|_{p,\hat{\eta}} - \|w - h\|_{p,\hat{\eta}} \big| \leq \|g - h\|_{p,\hat{\eta}}$, we can see that

$$\left| \|W' - \mathsf{T}W\|_{p,\hat{\eta}} - \|W' - \hat{\mathsf{T}}W\|_{p,\hat{\eta}} \right| \leq \|\mathsf{T}W - \hat{\mathsf{T}}W\|_{p,\hat{\eta}},$$

and

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

For every sample set ω the inequalities (27b),(27d) apply if $\|\mathsf{T}W - \hat{\mathsf{T}}W\|_{p,\hat{\eta}} \leq \epsilon_1$. Thus

$$\mathbf{P}\left\{ \left\{ \|W' - \mathsf{T}W\|_{p,\hat{\eta}} - \|W' - \hat{\mathsf{T}}W\|_{p,\hat{\eta}} > \epsilon_1 \right\} \cup \left\{ \|w^* - \hat{\mathsf{T}}W\|_{p,\hat{\eta}} - \|w^* - \mathsf{T}W\|_{p,\hat{\eta}} > \epsilon_1 \right\} \right\}$$
$$\leq \mathbf{P}\left\{ \|\mathsf{T}W - \hat{\mathsf{T}}W\|_{p,\hat{\eta}} > \epsilon_1 \right\} \leq \delta_1.$$
$$(28)$$

The probability that at least one of the inequalities in (27) 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 9

Proof Let us set up the following chain of inequalities:

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

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

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

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

$= \Big\| \max_{a\in\mathcal{A}} \mathbf{E}\Big[\mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\backslash K}(x_{k+1}) \left(\mathsf{T}^{N_t-k-1}\hat{W}_{N_t}^* - \hat{W}_{k+1}^* + \hat{W}_{k+1}^* \right)(x_{k+1}) \big| x_{k+1} \sim T_x\left(\cdot \mid x_k, a \right) \Big]$
$\quad - \hat{W}_k^* \Big\|_{p,\eta}$

$= [\ \max \mathbf{E}[\xi_1 + \xi_2] \leq \max \mathbf{E}|\xi_1| + \max \mathbf{E}|\xi_2|\ ]$

$\leq \Big\| \max_{a\in\mathcal{A}} \mathbf{E}\Big| \mathbf{1}_{A\backslash K}(x_{k+1}) \left(\mathsf{T}^{N_t-k-1}\hat{W}_{N_t}^* - \hat{W}_{k+1}^* \right)(x_{k+1}) \big| x_{k+1} \sim T_x\left(\cdot \mid x_k, a \right) \Big|$
$\quad + \max_{a\in\mathcal{A}} \mathbf{E}\Big| \mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A\backslash K}(x_{k+1}) \hat{W}_{k+1}^*(x_{k+1}) \big| x_{k+1} \sim T_x\left(\cdot \mid x_k, a \right) \Big| - \hat{W}_k^* \Big\|_{p,\eta}$

$= [\text{ Triangular inequality}]$

$\leq \Big\| \max_{a\in\mathcal{A}} \mathbf{E}\Big| \mathbf{1}_{A\backslash K}(x_{k+1}) \left(\mathsf{T}^{N_t-k-1}\hat{W}_{N_t}^* - \hat{W}_{k+1}^* \right)(x_{k+1}) \big| x_{k+1} \sim T_x\left(\cdot \mid x_k, a \right) \Big| \Big\|_{p,\eta}$

$$+ \left\| \max_{a \in \mathcal{A}} \mathbf{E} \Big| \mathbf{1}_K(x_{k+1}) + \mathbf{1}_{A \setminus K}(x_{k+1}) \hat{W}^*_{k+1}(x_{k+1}) | x_{k+1} \sim T_x \left(\cdot \mid x_k, a \right) \Big| - \hat{W}^*_k \right\|_{p, \eta}$$

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

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

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

$$= [\text{ Introduce density function } t_x \left(x_{k+1} \mid x_k, a \right) \text{ for kernel } T_x \ ]$$

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

Let us now show that the first term is bounded by $B^{\frac{1}{p}} \left\| \mathsf{T}^{N_t - k - 1} \hat{W}^*_{N_t} - \hat{W}^*_{k+1} \right\|_{p, \eta}$:

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

$$= [\text{ monotonicity of } L_p\text{-norms with respect to a probability measure }]$$

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

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

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

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

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

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

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

$$= \left(\int_{A \setminus K} \int_{A \setminus K} \max_{a \in \mathcal{A}} \left(\Big| \mathsf{T}^{N_t - k - 1} \hat{W}^*_{N_t} - \hat{W}^*_{k+1} \Big|^p t_x(x_{k+1}|x_k, a) \tfrac{\eta(x_k)}{\eta(x_{k+1})} dx_k \eta(x_{k+1}) dx_{k+1} \right) \right)^{\frac{1}{p}}$$

$$= [\text{ Recall that } \Big| \mathsf{T}^{N_t - k - 1} \hat{W}^*_{N_t} - \hat{W}^*_{k+1} \Big|^p \text{ is only a function of } x_{k+1}]$$

$$= \left(\int_{A \setminus K} \Big| \mathsf{T}^{N_t - k - 1} \hat{W}^*_{N_t} - \hat{W}^*_{k+1} \Big|^p \int_{A \setminus K} \left(\max_{a \in \mathcal{A}} \tfrac{t_x(x_{k+1}|x_k, a) \eta(x_k)}{\eta(x_{k+1})} \right) dx_k \eta(x_{k+1}) dx_{k+1} \right)^{\frac{1}{p}}.$$

Introduce now the upper bound on $\int_{A \setminus K} \left(\max_{a \in \mathcal{A}} \tfrac{t_x(x_{k+1}|x_k, a) \eta(x_k)}{\eta(x_{k+1})} \right) dx_k$ over the domain $A \setminus K$ as $B = \sup_{x_{k+1} \in A \setminus K} \int_{A \setminus K} \max_{a \in \mathcal{A}} \tfrac{t_x(x_{k+1}|x_k, a) \eta(x_k)}{\eta(x_{k+1})} dx_k$, obtaining

$$\leq \left(\int_{A \setminus K} \Big| \mathsf{T}^{N_t - k - 1} \hat{W}^*_{N_t} - \hat{W}^*_{k+1} \Big|^p B \eta(x_{k+1}) dx_{k+1} \right)^{\frac{1}{p}} = B^{\frac{1}{p}} \| \mathsf{T}^{N_t - k - 1} \hat{W}^*_{N_t} - \hat{W}^*_{k+1} \|_{p, \eta}.$$

We have finally shown that $\left\| \mathsf{T}^{N_t-k}\hat{W}^*_{N_t} - \hat{W}^*_k \right\|_{p,\eta} \leq \left\| \mathsf{T}\hat{W}^*_{k+1} - \hat{W}^*_k \right\|_{p,\eta} + B^{\frac{1}{p}} \left\| \mathsf{T}^{N_t-k-1}\hat{W}^*_{N_t} - \hat{W}^*_{k+1} \right\|_{p,\eta}$.
∎

Appendix E. Proof of Theorem 10

Proof If we estimate the quantity $r^*_{x_0}(K,A) = \left(\mathsf{T}^{N_t} W^*_{N_t} \right)(x_0) = \left(\mathsf{T}^{N_t}\hat{W}^*_{N_t} \right)(x_0)$ by $\left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0)$, then we have that $\hat{r}^*_{x_0}(K,A) = \mathbf{1}_K(x_0) + \mathbf{1}_{A\setminus K}(x_0) \left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0)$. The absolute deviation of the approximated $\hat{r}^*_{x_0}(K,A)$ from the exact $r^*_{x_0}(K,A)$ is given as

$$\left| \hat{r}^*_{x_0}(K,A) - r^*_{x_0}(K,A) \right| = \left| \left(\mathsf{T}^{N_t}\hat{W}^*_{N_t} \right)(x_0) - \left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0) \right|.$$

The objective is to present this error as a function of the errors introduced by the approximate mappings $\left| \left(\mathsf{T}^{N_t}\hat{W}^*_{N_t} \right)(x_0) - \left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0) \right|$, as well as of the quantities $\|\hat{W}^*_1 - \mathsf{T}\hat{W}^*_2\|_{p,\eta}$, $\|\hat{W}^*_2 - \mathsf{T}\hat{W}^*_3\|_{p,\eta}$, $\ldots$, $\|\hat{W}^*_{N_t-1} - \mathsf{T}\hat{W}^*_{N_t}\|_{p,\eta}$.

To this end, we first express a bound on $\left| \left(\mathsf{T}^{N_t}\hat{W}^*_{N_t} \right)(x_0) - \left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0) \right|$ as a function of $\left| \left(\hat{\mathsf{T}}\left(\mathsf{T}\hat{W}^*_1 \right)(x_0) - \hat{W}^*_1 \right)(x_0) \right|$ and of $\left\| \mathsf{T}^{N_t-1}\hat{W}^*_{N_t} - \hat{W}^*_1 \right\|_{p,\eta}$. Then Lemma 9 is used to express $\|T^{N_t-1}\hat{W}^*_{N_t} - \hat{W}^*_1\|_{p,\eta}$ as a function of the errors introduced by the approximate mappings. Similar to the first chain of inequality in the proof of Lemma 9 applied at step $k=0$ and point x_0, we obtain that

$$\left| \left(\mathsf{T}^{N_t}\hat{W}^*_{N_t} \right)(x_0) - \left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0) \right|$$
$$\leq \max_{a\in\mathcal{A}} \int_{A\setminus K} \left| \mathsf{T}^{N_t-1}\hat{W}^*_{N_t}(x_1) - \hat{W}^*_1(x_1) \right| t_x(x_1 \mid x_0, a)\, dx_1 + \left| \left(\mathsf{T}\hat{W}^*_1 \right)(x_0) - \left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0) \right|.$$

Let us now introduce a measure for the maximum concentration of the density function $t_x(x_1 \mid x_0, a)$ over $x_1 \in A \setminus K$, for any $a \in \mathcal{A}$, defined relative to the density of the distribution η in (16), as $B_0 = \sup_{x_1 \in A\setminus K} \max_{a\in\mathcal{A}} \frac{t_x(x_1|x_0,a)}{\eta(x_1)}$. Since $B_0\eta(x_1) \geq t_x(x_1 \mid x_0, a)$, it follows that

$$\max_{a\in\mathcal{A}} \int_{A\setminus K} \left| \mathsf{T}^{N_t-1}\hat{W}^*_{N_t}(x_1) - \hat{W}^*_1(x_1) \right| t_x(x_1 \mid x_0, a)\, dx_1 \leq B_0 \int_{A\setminus K} \left| \mathsf{T}^{N_t-1}\hat{W}^*_{N_t}(x_1) - \hat{W}^*_1(x_1) \right| \eta(x_1) dx_1.$$

The last expression corresponds to a 1-norm with respect to a probability measure η 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

$$\left| \left(\mathsf{T}^{N_t}\hat{W}^*_{N_t} \right)(x_0) - \hat{\mathsf{T}}\hat{W}^*_1(x_0) \right| \leq \left| \left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0) - \left(\mathsf{T}\hat{W}^*_1 \right)(x_0) \right| + B_0 \left\| \left(\mathsf{T}^{N_t-1}\hat{W}^*_{N_t} \right) - \hat{W}^*_1 \right\|_{p,\eta}.$$

The second term can be expressed as a function of the weighted p-norm of the approximations by applying Lemma 9. This leads to the expression for an upper bound on the approximation error as

$$\left| \hat{r}^*_{x_0}(K,A) - r^*_{x_0}(K,A) \right| \leq \left| \left(\hat{\mathsf{T}}\hat{W}^*_1 \right)(x_0) - \left(\mathsf{T}\hat{W}^*_1 \right)(x_0) \right| + B_0 \sum_{k=1}^{N_t-1} B^{\frac{k-1}{p}} \left\| \hat{W}^*_k - \mathsf{T}\hat{W}^*_{k+1} \right\|_{p,\eta}.$$

From the above expression, a sufficient condition the accuracy in (3) to hold is

$$\mathbf{P}\Big\{ \Big|\big(\hat{\mathsf{T}}\hat{W}_1^*\big)(x_0) - \big(\mathsf{T}\hat{W}_1^*\big)(x_0)\Big| + B_0 \sum_{k=1}^{N_t-1} B^{\frac{k-1}{p}} \Big\|\hat{W}_k^* - \mathsf{T}\hat{W}_{k+1}^*\Big\|_{p,\eta} > \boldsymbol{\Delta} \Big\} \le \delta_{\boldsymbol{\Delta}}.$$

∎

Appendix F. Sample Complexities

Given $\epsilon_{0,1,2}$ and α, select $\delta_{0,1,2} > 0$ such that $1 - \alpha = \delta_0 + (N_t - 1)\delta_1 + (N_t - 1)\delta_2$, and let us pick values for N, M, M_0 such that

$$\delta_0 \le 2|\mathcal{A}|e^{-2M_0(\epsilon_0)^2}, \quad \delta_1 \le 2|\mathcal{A}|Ne^{-2M(\epsilon_1)^2}, \quad \delta_2 \le 4e(d+1)\Big(\frac{32e}{\epsilon_2^p}\Big)^d e^{-\frac{N\epsilon_2^{2p}}{128}}.$$

Note that the first two inequalities are approximated with first order approximation for which we know that $1 - (1 - 2e^{-2M_0(\epsilon_0)^2})^{|\mathcal{A}|} \le 2|\mathcal{A}|e^{-2M_0(\epsilon_0)^2}$ and $1 - (1 - 2e^{-2M(\epsilon_1)^2})^{|\mathcal{A}|N} \le 2|\mathcal{A}|Ne^{-2M(\epsilon_1)^2}$. The obtained integer values for N, M, M_0 are given as

$$\begin{cases} N &= \Big\lceil 128\left(\ln(4e(d+1)) + d\ln(32e)\right)\Big(\frac{1}{\epsilon_2}\Big)^{2p} + 128dp\Big(\frac{1}{\epsilon_2}\Big)^{2p}\ln\Big(\frac{1}{\epsilon_2}\Big) + 128\Big(\frac{1}{\epsilon_2}\Big)^{2p}\ln\Big(\frac{1}{\delta_2}\Big)\Big\rceil, \\ M &= \Big\lceil \frac{1}{2}\Big(\frac{1}{\epsilon_1}\Big)^2\Big(\ln(2|\mathcal{A}|) + \ln(\frac{1}{\delta_1}) + \ln(N)\Big)\Big\rceil, \\ M_0 &= \Big\lceil \frac{1}{2}\Big(\frac{1}{\epsilon_0}\Big)^2\Big(\ln(2|\mathcal{A}|) + \ln(\frac{1}{\delta_0})\Big)\Big\rceil, \end{cases}$$

The use of the obtained M, M_0, N in (17) leads to a confidence of at least α.

Appendix G. Proof of Theorem 12

Proof The proof of Theorem 12 is built observing that (a.) the single step error $\|\hat{W}_k^* - \mathsf{T}\hat{W}_{k+1}^*\|_{1,\eta}$ is bounded by the sum of the expectations of (19) and (21); that (b.) the propagation of the single step errors gives a bound on the overall approximation error, see Theorem 10 – 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.)

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

Define a set of i.i.d. random variables $\vec{y}_1 = [y_1^{a,1}, y_1^{a,2}, \ldots, y_1^{a,\tilde{M}}]$ drawn from the distribution $y_1^{a,j} \sim T_x\left(\cdot \mid x, a\right)$. Introduce $\mathbf{E}_{\vec{y}_1}\Big[\max_{a\in\mathcal{A}}\hat{\mathsf{T}}_1^a\hat{W}_{k+1}^*(x)|x\Big]$ as an auxiliary variable with $\hat{\mathsf{T}}_1^a$ the estimated operator as defined in (20) and computed over the $\vec{y}_1$.

$$= \mathbf{E}_x\Big[\ \Big|\hat{W}_k^*(x) - \mathbf{E}_{\vec{y}_1}\Big[\max_{a\in\mathcal{A}}\hat{\mathsf{T}}_1^a\hat{W}_{k+1}^*(x)|x\Big] + \mathbf{E}_{\vec{y}_1}\Big[\max_{a\in\mathcal{A}}\hat{\mathsf{T}}_1^a\hat{W}_{k+1}^*(x)|x\Big] - \mathsf{T}\hat{W}_{k+1}^*(x)\Big|\ \Big]$$

33

$$\leq \mathbf{E}_x \left[\left| \hat{W}_k^*(x) - \mathbf{E}_{\vec{y_1}} \left[\max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x)|x \right] \right| \right] + \mathbf{E}_x \left[\left| \mathbf{E}_{\vec{y_1}} \left[\max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x)|x \right] - \mathsf{T}\hat{W}_{k+1}^*(x) \right| \right]$$

$$\leq \underbrace{\mathbf{E}_{x,\vec{y_1}} \left[\left| \hat{W}_k^*(x) - \max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) \right| \right]}_{\text{[Single step error]}} + \underbrace{\mathbf{E}_x \left[\left| \mathbf{E}_{\vec{y_1}} \left[\max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x)|x \right] - \mathsf{T}\hat{W}_{k+1}^*(x) \right| \right]}_{\text{[Bias term]}}.$$

Observe that the *single step error*, $\mathbf{E}_{x,\vec{y_1}} \left[\left| \hat{W}_k^*(x) - \max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) \right| \right]$ is equal to $\|\hat{W}_k^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\|_{1,\eta}$ and $\mathbf{E}\|\hat{W}_k^* - \hat{\mathsf{T}}\hat{W}_{k+1}^*\|_{1,\tilde{\eta}}$. The *bias term* gives the bias introduced by using an estimate of the operator and it can be rewritten as the expected value of (21). Note that $\max_{a \in \mathcal{A}} \mathbf{E}_y \left[V_{k+1}(y)|x,a \right]$ is a function of x, a and $|\mathbf{E}_{\vec{y_1}} [f(\vec{y})]| \leq \mathbf{E}_{\vec{y_1}} [|f(\vec{y})|]$, thus it follows that

$$\text{[Bias term]} \leq \mathbf{E}_x \mathbf{E}_{\vec{y_1}} \left[\left| \max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) - \mathsf{T}\hat{W}_{k+1}^*(x) \right| \, \Big| \, x \right] \leq \mathbf{E}_x \mathbf{E}_{\vec{y_1}} \left[\max_{a \in \mathcal{A}} \left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) - \mathbf{E}_{\vec{y_2}} \left[\hat{\mathsf{T}}_2^a \hat{W}_{k+1}^*(x)|x \right] \right| \, \Big| \, x \right].$$

The second inequality follows from introducing a secondary set of random variables $\vec{y_2} = [y_2^{a,1}, y_2^{a,2}, \dots, y_2^{a,1\tilde{M}}]$ for which the elements are i.i.d. as $y_2^{a,1} \sim T_x(\cdot \mid x,a)$ and which are independent of $\vec{y_1}$. Substituting $\mathsf{T}\hat{W}_{k+1}^*(x)$ with $\max_{a \in \mathcal{A}} \mathbf{E}_{\vec{y_2}} \left[\hat{\mathsf{T}}_2^a \hat{W}_{k+1}^*(x)|x,a \right]$ we have

$$\text{[Bias term]} \leq \mathbf{E}_x \mathbf{E}_{\vec{y_1}} \left[\max_{a \in \mathcal{A}} \mathbf{E}_{\vec{y_2}} \left[\left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^*(x) \right| |x,a,\vec{y_1} \right] |x \right]$$

$$= \mathbf{E}_{x,\vec{y_1},\vec{y_2}} \left[\max_{a \in \mathcal{A}} \left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^*(x) \right| \right].$$

The last equality is equal to the expected value of the estimated bias $\mathbf{E}\big\| \max_{a \in \mathcal{A}} \big| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^* - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^* \big| \big\|_{1,\tilde{\eta}}$. This proves statement (a.).

Part (b.) & (c.) Based on Theorem 10 $\hat{r}_{x_0}^*(K,A)$ has accuracy $\boldsymbol{\Delta}$ with probability $\delta_{\boldsymbol{\Delta}}$ if

$$\mathbf{P}\left\{ |\hat{W}_0^*(x_0) - \mathsf{T}\hat{W}_1^*(x_0)| + B_0 \sum_{k=1}^{N_t-1} B^{k-1} \mathbf{E}\left[\left| \hat{W}_k^*(x) - \max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) \right| + \max_{a \in \mathcal{A}} \left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^*(x) \right| \right] \geq \boldsymbol{\Delta} \right\} < \delta_{\boldsymbol{\Delta}}.$$

Which holds under a union bounding argument if

$$\mathbf{P}\left\{ |\hat{W}_0^*(x_0) - \mathsf{T}\hat{W}_1^*(x_0)| \geq \epsilon_0 \right\} < \delta_0 \tag{29}$$

$$\mathbf{P}\left\{ \begin{array}{c} B_0 \sum_{k=1}^{N_t-1} B^{k-1} \mathbf{E}\left[\left| \hat{W}_k^*(x) - \max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) \right| + \max_{a \in \mathcal{A}} \left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^*(x) \right| \right] \geq B_0 \epsilon \\ + B_0 \sum_{k=1}^{N_t-1} B^{k-1} \left(\left\| \hat{W}_k^* - \hat{\mathsf{T}}\hat{W}_{k+1}^* \right\|_{1,\tilde{\eta}} + \left\| \max_{a \in \mathcal{A}} \left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^* - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^* \right| \right\|_{1,\tilde{\eta}} \right) \end{array} \right\} < e^{-2\frac{\bar{N}\epsilon^2}{L^2}} \tag{30}$$

and $\boldsymbol{\Delta}$ and $\delta_{\boldsymbol{\Delta}}$ are given as (22a) and (22b).
The probabilistic bound (29) follows from Lemma 8 for the estimation error of an empirical norm with accuracy ϵ_0, δ_0 obtained for $p = 1$, $M = M_0$ and $N = 1$ as long as $0 < 2e^{-2M_0\epsilon_0^2} < 1$. The probabilistic bound (30) follows from a one-sided Hoeffding's inequality (Hoeffding, 1963) with random variable

$$\sum_{k=1}^{N_t-1} B^{k-1} \left(\left| \hat{W}_k^*(x) - \max_{a \in \mathcal{A}} \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) \right| + \max_{a \in \mathcal{A}} \left| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^*(x) - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^*(x) \right| \right),$$

obtained from the combination of random variable $x \sim \eta$ and conditional random variables $\vec{y}_1$ and $\vec{y}_2$ and taking values in the range $[0, 2\sum_{k=1}^{N_t-1} B^{k-1}]$. Note that its estimated of interest over $\tilde{N}$ samples can be rewritten in the form of (22a),

$$\sum_{k=1}^{N_t-1} B^{k-1} \Big(\big\| \hat{W}_k^* - \hat{\mathsf{T}}\hat{W}_{k+1}^* \big\|_{1,\tilde{\eta}} + \big\| \max_{a \in \mathcal{A}} \big| \hat{\mathsf{T}}_1^a \hat{W}_{k+1}^* - \hat{\mathsf{T}}_2^a \hat{W}_{k+1}^* \big| \big\|_{1,\tilde{\eta}} \Big).$$

This concludes the proof of Theorem 12. ∎

Appendix H. Scaling factor for case study

Compute B as in (14) using the given density distribution of the transitions (25), as

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

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

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

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

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

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

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

The integral is rewritten as one over a scaled 2-dimension multivariate Gaussian density distribution with mean $\bar{\mu}$ and covariance $\bar{\Sigma}$. With this result, it can be deduced that B is smaller than $\frac{1}{|\mathbf{A}|}|\mathcal{A}|$ as

$$B \leq \sup_{y \in A \setminus K} \frac{1}{|\mathbf{A}|} \left(\sum_{a \in \mathcal{A}} \int_{\mathcal{X}} \frac{1}{\sqrt{|\bar{\Sigma}|(2\pi)^2}} \exp\left(-\frac{1}{2}(x-\bar{\mu}(y,a))^T \bar{\Sigma}^{-1}(x-\bar{\mu}(y,a))\right) dx \right) \qquad (31)$$

$$\leq \sup_{y \in A \setminus K} \frac{1}{|\mathbf{A}|} \sum_{a \in \mathcal{A}} 1 = \frac{1}{|\mathbf{A}|}|\mathcal{A}|.$$