UVIP: MODEL-FREE APPROACH TO EVALUATE REINFORCEMENT LEARNING ALGORITHMS

Ilya Levin
HSE University
ivlevin@hse.ru

Denis Belomestny
Duisburg-Essen University
and HSE University
denis.belomestny@uni-due.de

Alexey Naumov
HSE University
anaumov@hse.ru

Sergey Samsonov
HSE University
svsamsonov@hse.ru

ABSTRACT

Policy evaluation is an important instrument for the comparison of different algorithms in Reinforcement Learning (RL). Yet even a precise knowledge of the value function V^π corresponding to a policy π does not provide reliable information on how far is the policy π from the optimal one. We present a novel model-free upper value iteration procedure (UVIP) that allows us to estimate the suboptimality gap $V^\star(x) - V^\pi(x)$ from above and to construct confidence intervals for $V^\star$. Our approach relies on upper bounds to the solution of the Bellman optimality equation via martingale approach. We provide theoretical guarantees for UVIP under general assumptions and illustrate its performance on a number of benchmark RL problems.

1 INTRODUCTION

The key objective of Reinforcement Learning (RL) is to learn an optimal agent's behaviour in an unknown environment. A natural performance metric is given by the value function V^π which is expected total reward of the agent following π. There are efficient algorithms to evaluate this quantity, e.g. temporal difference methods Sutton (1988), Tsitsiklis & Van Roy (1997). Unfortunately, even a precise knowledge of V^π does not provide reliable information on how far is the policy π from the optimal one. To address this issue a popular quality measure is the *regret* of the algorithm that is the difference between the total sum of rewards accumulated when following the optimal policy and the sum of rewards obtained when following the current policy π (see e.g. Jaksch et al. (2010)). In the setting of finite state- and action space Markov Decision Processes (MDP) there is a variety of regret bounds for popular RL algorithms like Q-learning Jin et al. (2018), optimistic value iteration Azar et al. (2017), and many others.

Unfortunately, regret bounds beyond the discrete setup are much less common in the literature. Even more crucial drawback of the regret-based comparison is that regret bounds are typically pessimistic and rely on the unknown quantities of the underlying MDP's. A simpler, but related, quantity is the *suboptimality gap (policy error)* $\Delta_\pi(x) \doteq V^\star(x) - V^\pi(x)$. Since we do not know $V^\star$, the suboptimality gap can not be calculated directly. There is a vast amount of literature devoted to theoretical guarantees for $\Delta_\pi(x)$, see e.g. Antos et al. (2007), Szepesvári (2010), Pires & Szepesvári (2016) and references therein. However, these bounds share the same drawbacks as the regret bounds. Moreover, known bounds does not apply to the general policy π and depends heavily on the particular algorithm which produced it. For instance, in Approximate Policy Iteration (API, Bertsekas & Tsitsiklis (1996)) all existing bounds for $\Delta_\pi(x)$ depend on the one-step error induced by the approximation of the action-value function. This one-step error is difficult to quantify since it depends on the unknown smoothness properties of the action-value function. Similarly, in policy gradient methods (see e.g. Sutton & Barto (2018)), there is always an approximation error due to the choice of family of policies that can be hardly quantified. The approach based on the policy optimism principle (see Efroni et al. (2019)) suggests to initialise the value iteration algorithm using

an upper bound (optimistic value) for $V^\star$, yielding a sequence of upper bounds converging to $V^\star$. However this approach is tailored to finite state- and action space MDPs and is not applicable to evaluate the quality of the general policy π. To summarise, such bounds can not be used to construct tight model-free confidence bounds for $V^\star$.

In this paper we are interested in deriving agnostic (model independent) bounds for the policy error using the concept of upper solutions to the Bellman optimality equation. Our approach is substantially different from the ones known in literature as it can be used to evaluate performance bounds for arbitrary given policy π. The concept of upper solutions is closely related to martingale duality in optimal control and information relaxation approach, see Belomestny & Schoenmakers (2018), Rogers (2007) and references therein. This idea has been successfully used in the recent paper Shar & Jiang (2020). This work proposes to use the duality approach to improve the performance of Q-learning algorithm in finite horizon MDP through the use of "lookahead" upper and lower bounds. Compared to Shar & Jiang (2020), our approach is not restricted to the Q-learning. The concept of upper solutions has also a connection to distributional RL, as it can be formulated pathwise or using distributional Bellman operator, see e.g. Lyle et al. (2019). A further study of this connection is a promising future research area.

Contributions and Organization The contributions of this paper are three-fold:

- We propose a novel approach to construct model free confidence bounds for the optimal value function $V^\star$ based on a notion of upper solutions.
- Given a policy π, we propose an upper value iterative procedure (UVIP) for constructing an (almost sure) upper bound for V^π such that it coincides with $V^\star$ if $\pi = \pi^\star$.
- We study convergence properties of the approximate UVIP in the case of general state and action spaces. In particular, we show that the variance of the resulting upper bound is small if π is close to π^* leading to the tight confidence bounds for $V^\star$.

The paper is organized as follows. First, in section 2, we briefly recall main concepts related with the MDPs, and introduce some notations. Then in sections 3 and 4 we introduce the framework of UVIP and discuss its basic properties. In section 5 we perform theoretical study of the approximate UVIP. Numerical results are collected in section 6. Section 7 concludes the paper. Section A in appendix is devoted to the proof of main theoretical results.

Notations and definitions For $N \in \mathbb{N}$ we define $[N] \doteq \{1, \dots, N\}$. Let us denote the space of bounded measurable functions with domain X by $\mathcal{B}(\mathsf{X})$ equipped with the norm $\|f\|_{\mathsf{X}} = \sup_{x\in\mathsf{X}} |f(x)|$ for any $f \in \mathcal{B}(\mathsf{X})$. In what follows, whenever a norm is uniquely identifiable from its argument, we will drop the index of the norm denoting the underlying space. We denote by P^a an $\mathcal{B}(\mathsf{X}) \to \mathcal{B}(\mathsf{X})$ operator defined by $(\mathrm{P}^a V)(x) = \int V(x')\mathrm{P}^a(dx'|x)$. For an arbitrary metric space $(\mathcal{X}, \rho_\mathcal{X})$ and function $f : \mathcal{X} \to \mathbb{R}$ we define by $\mathrm{Lip}_{\rho_\mathcal{X}}(f) \doteq \sup_{x\neq y} |f(x) - f(y)|/\rho_\mathcal{X}(x, y)$.

2 PRELIMINARY

A *Markov Decision Process* (MDP) is a tuple $(\mathsf{X}, \mathsf{A}, \mathcal{P}, r)$, where X is the state space, A is the action space, $\mathcal{P} = (\mathrm{P}^a)_{a\in\mathsf{A}}$ is the *transition probability kernel* and $r = (r^a)_{a\in\mathsf{A}}$ is the *reward function*. For each state $x \in \mathsf{X}$ and action $a \in \mathsf{A}$, $\mathrm{P}^a(\cdot|x)$ stands for a distribution over the states in X, that is, the distribution over the next states given that action a is taken in the state x. For each action $a \in \mathsf{A}$ and state $x \in \mathsf{X}$, $r^a(x)$ gives a reward received when action a is taken in state x. An MDP describes the interaction of an agent and its environment. When an action $A_t \in \mathsf{A}$ at time t is chosen by the agent, the state X_t transitioned to $X_{t+1} \sim \mathrm{P}^{A_t}(\cdot|X_t)$. The agent's goal is to maximize the *expected total discounted reward*, $\mathsf{E}[\sum_{t=0}^\infty \gamma^t r^{A_t}(X_t)]$, where $0 < \gamma < 1$ is the *discount factor*. A rule describing the way an agent acts given its past actions and observations is called a *policy*. The *value function* of a policy π in a state $x \in \mathsf{X}$, denoted by $V^\pi(x)$, is $V^\pi(x) = \mathsf{E}[\sum_{t=0}^\infty \gamma^t r^{A_t}(X_t)|X_0 = x]$, that is, the expected total discounted reward when the initial state ($X_0 = x$) assuming the agent follows the policy π. Similarly, we define the action-value function $Q^\pi(x, a) = \mathsf{E}[\sum_{t=0}^\infty \gamma^t r^{A_t}(X_t)|X_0 = x, A_0 = a]$. An *optimal policy* is one that achieves the maximum possible value amongst all policies in each state $x \in \mathsf{X}$. The *optimal value* for state x is denoted by $V^\star(x)$. A *deterministic Markov policy* can be identified with a map $\pi : \mathsf{X} \to \mathsf{A}$, and the space of measurable deterministic Markov policies will be denoted by Π. When, in addition, the reward function is bounded, which we assume from now on, all the value functions are bounded and one can always find a deterministic Markov policy that is optimal Puterman (2014). We also define a *greedy policy* w.r.t. action-value function $Q(x, a)$,

which is a deterministic policy $\pi(x) \in \operatorname{argmax}_{a\in\mathsf{A}} Q(x,a)$. The *Bellman return operator* w.r.t. P, $\mathsf{T}_{\mathrm{P}} : \mathcal{B}(\mathsf{X}) \to \mathcal{B}(\mathsf{X}\times\mathsf{A})$, is defined by $(\mathsf{T}_{\mathrm{P}}V)(x,a) = r^a(x) + \gamma \mathrm{P}^a V(x)$ and the *maximum selection* operator $\mathsf{M} : \mathcal{B}(\mathsf{X}\times\mathsf{A}) \to \mathcal{B}(\mathsf{X})$ is defined by $(\mathsf{M}V^{\cdot})(x) = \max_a V^a(x)$. Then $\mathsf{M}\mathsf{T}_{\mathrm{P}}$ corresponds to the *Bellman optimality operator*, see Puterman (2014). The optimal value function $V^\star$ satisfies a non-linear fixed-point equation

$$V^\star(x) = \mathsf{M}\mathsf{T}_{\mathrm{P}} V^\star(x). \tag{1}$$

which is known as the *Bellman optimality equation*. We write $Y^{x,a}, x \in \mathsf{X}, a \in \mathsf{A}$ for a random variable generated according to $\mathrm{P}^a(\cdot|x)$, and define a random Bellman operator $(\widetilde{\mathsf{T}}_{\mathrm{P}}V)(x) \mapsto r^a(x) + \gamma V(Y^{x,a})$. We say that a (deterministic) policy π is *greedy* w.r.t. a function $V \in \mathcal{B}(\mathsf{X})$ if, for all $x \in \mathsf{X}$

$$\pi(x) \in \operatorname{argmax}_{a\in\mathsf{A}} \left\{ r^a(x) + \gamma \mathrm{P}^a V(x) \right\}.$$

3 Upper solutions and the main concept of UVIP

A straightforward approach to bound the policy error $\Delta_\pi(x)$ requires the estimation of the optimal value function $V^\star(x)$. Recall that $V^\star$ is a solution of the Bellman optimality equation (1). If the transition kernel $(\mathrm{P}_a)_{a\in\mathsf{A}}$ is known, the standard solution is the value iteration algorithm, see Bertsekas & Shreve (1978). In this algorithm, the estimates are recursively constructed via $V_{k+1} = \mathsf{M}\mathsf{T}_{\mathrm{P}} V_k$. Due to the Banach's fixed point theorem, $\|V_k - V^\star\|_{\mathsf{X}} \le \gamma^k \|V_0 - V^\star\|_{\mathsf{X}}$, provided that $V_0 \in \mathcal{B}(\mathsf{X})$. Moreover, $V_k(x) \ge V^\star(x)$ for any $x \in \mathsf{X}$ and $k \in \mathbb{N}$, provided that $V_0(x) \ge V^\star(x)$. For example, if $\|r^a\|_{\mathsf{X}} \le R_{\max}$ for all $a \in \mathsf{A}$, we can take $V_0(x) = R_{\max}/(1-\gamma)$.

Unfortunately, (1) does not allow to represent $V^\star$ as an expectation and to reduce the problem of estimating $V^\star$ to a stochastic approximation problem. Moreover, if $(\mathrm{P}^a)_{a\in\mathsf{A}}$ is replaced by its empirical estimate $\hat{\mathrm{P}}^a$ the desired upper biasness property $V_k(x) \ge V^\star(x)$ is lost. Some recent works (e.g. Efroni et al. (2019)) suggested a modification of the optimism-based approach applicable in case of unknown $(\mathrm{P}^a)_{a\in\mathsf{A}}$. Yet this modification contains an additional optimization step, which is unfeasible beyond the tabular state- and action space problems. Therefore the problem of constructing upper bounds for the optimal value function $V^\star$ and policy error remains open and highly relevant. Below we describe our approach, which is based on the following key assumptions:

- we consider infinite-horizon MDPs with discount factor $\gamma < 1$;
- we can sample from the conditional distribution $\mathrm{P}^a(\cdot|x)$ for any $x \in \mathsf{X}$ and $a \in \mathsf{A}$.

The key concept of our algorithm is an *upper solution*, introduced below.

Definition 3.1. *We call a function V^{up} an upper solution to the Bellman optimality equation* (1) *if*

$$V^{\mathrm{up}}(x) \ge \mathsf{M}\mathsf{T}_{\mathrm{P}} V^{\mathrm{up}}(x)\,, \forall x \in \mathsf{X}\,.$$

Upper solutions can be used to build tight upper bounds for the optimal value function $V^\star$. Let $\Phi \in \mathcal{B}(\mathsf{X})$ be a *martingale function* w.r.t. the operator P^a, that is, $\mathrm{P}^a\Phi(x) = 0$ for all $a \in \mathsf{A}, x \in \mathsf{X}$. Define V^{up} as a solution to the following fixed point equation:

$$V^{\mathrm{up}}(x) = \mathsf{E}[\max_a \{ r^a(x) + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi(Y^{x,a}))\}], \quad Y^{x,a} \sim \mathrm{P}^a(\cdot|x). \tag{2}$$

In terms of the random Bellman operator $\widetilde{\mathsf{T}}_{\mathrm{P}}$, we can rewrite (2) as $V^{\mathrm{up}} = \mathsf{E}[\mathsf{M}\widetilde{\mathsf{T}}_{\mathrm{P}}(V^{\mathrm{up}} - \Phi)]$. It is easy to see that (2) defines an upper solution. Indeed, for any $x \in \mathsf{X}$,

$$\begin{aligned} V^{\mathrm{up}}(x) &\ge \max_a \mathsf{E}[r^a(x) + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi(Y^{x,a}))] \\ &= \max_a \{ r^a(x) + \gamma \mathrm{P}^a V^{\mathrm{up}}(x) \} = \mathsf{M}\mathsf{T}_{\mathrm{P}} V^{\mathrm{up}}(x)\,. \end{aligned}$$

Note that unlike the optimal state value function $V^\star$, the upper solution V^{up} is represented as an expectation, which allows us to use various stochastic approximation methods to compute V^{up}. The Banach's fixed-point theorem implies that for iterates

$$V^{\mathrm{up}}_{k+1} = \mathsf{E}[\mathsf{M}\widetilde{\mathsf{T}}_{\mathrm{P}}(V^{\mathrm{up}}_k - \Phi)], \quad k \in \mathbb{N},$$

we have convergence $V^{\mathrm{up}}_k \to V^{\mathrm{up}}$ as $k \to \infty$. Moreover, V^{up} does not depend on V^{up}_0 and $V^{\mathrm{up}}_k(x) \ge V^\star(x)$ for any $k \in \mathbb{N}, x \in \mathsf{X}$, provided that $V^{\mathrm{up}}_0(x) \ge V^\star(x)$. Given a policy π and the

corresponding value function V^π, we set $\Phi_\pi^{x,a}(y) \doteq V^\pi(y) - (\mathrm{P}^a V^\pi)(x)$. It is easy to check that $\mathrm{P}^a \Phi_\pi^{x,a}(x) = 0$. This leads to the upper value iterative procedure (UVIP):

$$V_{k+1}^{\mathrm{up}}(x) = \mathsf{E}[\mathsf{M}\widetilde{\mathsf{T}}_{\mathrm{P}}(V_k^{\mathrm{up}} - \Phi_\pi^{x,\cdot})(x)] = \mathsf{E}\big[\max_a \{r^a(x) + \gamma(V_k^{\mathrm{up}}(Y^{x,a}) - \Phi_\pi^{x,a}(Y^{x,a}))\}\big] \tag{3}$$

with $V_0^{\mathrm{up}} \in \mathcal{B}(\mathsf{X})$. Algorithm 1 contains the pseudocode of the UVIP for MDPs with finite state and action spaces. Several generalizations are discussed in the next section. Further note that by taking

Algorithm 1: UVIP

Input: V^π,V_0^{up}, γ, ε
Result: V^{up}
for $x \in \mathsf{X}, a \in \mathsf{A}$ **do**
  for $y \in \mathsf{X}$ **do**
    $\Phi_\pi^{x,a}(y) = V^\pi(y) - (\mathrm{P}^a V^\pi)(x)$;
  end
end
$k = 1$; **while** $\|V_k^{\mathrm{up}} - V_{k-1}^{\mathrm{up}}\|_{\mathsf{X}} > \varepsilon$ **do**
  for $x \in \mathsf{X}$ **do**
    $V_{k+1}^{\mathrm{up}}(x) = \mathsf{E}[\max_a \{r^a(x) + \gamma(V_k^{\mathrm{up}}(Y^{x,a}) - \Phi_\pi^{x,a}(Y^{x,a}))\}], \quad Y^{x,a} \sim \mathrm{P}^a(\cdot|x)$;
  end
  $k = k + 1$;
end
$V^{\mathrm{up}} = V_k^{\mathrm{up}}$.

$\Phi^{x,a}(y) \doteq V^\star(y) - (\mathrm{P}^a V^\star)(x)$, we get with probability 1 :

$$V^\star(x) = (\mathsf{M}\widetilde{\mathsf{T}}_{\mathrm{P}}(V^\star - \Phi^{x,\cdot}))(x) = \max_a \{r^a(x) + \gamma(V^\star(Y^{x,a}) - \Phi^{x,a}(Y^{x,a}))\}, \tag{4}$$

that is, (4) can be viewed as an almost sure version of the Bellman equation $V^\star = \mathsf{M}\mathsf{T}_{\mathrm{P}} V^\star$.

The upper solutions can be used to evaluate the quality of the policies and to construct confidence intervals for $V^\star$. It is clear that

$$V^\pi(x) \le V^\star(x) \le V_k^{\mathrm{up}}(x)$$

for any $k \in \mathbb{N}$ and $x \in \mathsf{X}$, thus a policy π can be evaluated by computing the difference $\Delta_{\pi,k}^{\mathrm{up}}(x) \doteq V_k^{\mathrm{up}}(x) - V^\pi(x) \ge \Delta_\pi(x)$. Representations (3) and (4) imply

$$\left\|V_{k+1}^{\mathrm{up}} - V^\star\right\|_{\mathsf{X}} \le \gamma \left\|V_k^{\mathrm{up}} - V^\star\right\|_{\mathsf{X}} + 2\gamma \left\|V^\pi - V^\star\right\|_{\mathsf{X}}, \quad k \in \mathbb{N}.$$

Hence, we derive that $\Delta_\pi^{\mathrm{up}} \doteq \lim_{k\to\infty} \Delta_{\pi,k}^{\mathrm{up}}$ satisfies

$$\|\Delta_\pi\|_{\mathsf{X}} \le \|\Delta_\pi^{\mathrm{up}}\|_{\mathsf{X}} \le \left(1 + 2\gamma(1-\gamma)^{-1}\right) \|V^\star - V^\pi\|_{\mathsf{X}}. \tag{5}$$

As a result $\Delta_\pi^{\mathrm{up}} = 0$ if $\pi = \pi^\star$ and the corresponding confidence intervals collapses into one point. Moreover, for a policy π which is greedy w.r.t. an action-value function $Q^\pi(x,a)$, it holds that $V^\pi(x) \ge V^\star(x) - 2(1-\gamma)^{-1}\|Q^\pi - Q^\star\|_{\mathsf{X}\times\mathsf{A}}$ (see Szepesvári (2010)). Thus we can rewrite the bound (5) in terms of action-value functions

$$\|\Delta_\pi^{\mathrm{up}}\|_{\mathsf{X}} \le 2\left(1 + 2\gamma(1-\gamma)^{-1}\right)(1-\gamma)^{-1}\|Q^\pi - Q^\star\|_{\mathsf{X}\times\mathsf{A}}.$$

The quantity $\Delta_{\pi,k}^{\mathrm{up}}$ can be used to measure the quality of policies π obtained by many well-known algorithms like Reinforce (Williams (1992)), API (Bertsekas & Tsitsiklis (1996)), A2C (Mnih et al. (2016)) and DQN (Mnih et al. (2013), Mnih et al. (2015)).

4 Approximate UVIP

In order to implement the approach described in the previous section, we need to construct empirical estimates for the outer expectation and the one-step transition operator P^a in (3). While in tabular

case this boils down to a straightforward Monte Carlo, in case of infinitely many states we need an additional approximation step. Algorithm 2 contains the pseudocode of *Approximate* UVIP algorithm. Our main assumption is that sampling from $\mathrm{P}^a(\cdot|x)$ is available for any $a \in \mathsf{A}$ and $x \in \mathsf{X}$. For simplicity we assume that the value function V^π is known, but it can be replaced by its (lower biased) estimate both in Algorithm 2 and subsequent theoretical results. The proposed algorithm proceeds as follows. At the $(k+1)$th iteration, given a previously constructed approximation $\widehat{V}_k^{\mathrm{up}}$, we compute

Algorithm 2: Approximate UVIP

Input: Sample $(\mathsf{x}_1, \dots, \mathsf{x}_N)$; V^π, $\widetilde{V}_0^{\mathrm{up}}$, M_1, M_2, γ, ε
Result: $\widehat{V}^{\mathrm{up}}$
Generate $r^a(X_i), Y_j^{\mathsf{x}_i,a} \sim \mathrm{P}^a(\cdot|\mathsf{x}_i)$ for all $i \in [N]$, $j \in [M_1+M_2]$, $a \in \mathsf{A}$;
$k=1$; **while** $\|\widetilde{V}_k^{\mathrm{up}} - \widetilde{V}_{k-1}^{\mathrm{up}}\|_{\mathsf{x}_N} > \varepsilon$ **do**
  for $a \in \mathsf{A}$ **do**
    for $i \in [N]$ **do**
      for $j \in [M_1+M_2]$ **do**
        $\widehat{V}_k^{\mathrm{up}}(Y_j^{\mathsf{x}_i,a}) = I[\widetilde{V}_k^{\mathrm{up}}](Y_j^{\mathsf{x}_i,a})$ with $I[\cdot](\cdot)$ defined in (6);
      end
      $\overline{V}^{(i,a)} = M_1^{-1} \sum_{j=1}^{M_1} V^\pi(Y_j^{\mathsf{x}_i,a})$;
    end
  end
  for $i \in [N]$ **do**
    $\widetilde{V}_{k+1}^{\mathrm{up}}(\mathsf{x}_i) = M_2^{-1} \sum_{j=M_1+1}^{M_1+M_2} \max_{a\in\mathcal{A}} \big\{ r^a(\mathsf{x}_i) + \gamma\big(\widehat{V}_k^{\mathrm{up}}(Y_j^{\mathsf{x}_i,a}) - V^\pi(Y_j^{\mathsf{x}_i,a}) + \overline{V}^{(i,a)}\big)\big\}$;
  end
  $k = k+1$;
end
$\widehat{V}^{\mathrm{up}} = \widehat{V}_k^{\mathrm{up}}$.

$$\widetilde{V}_{k+1}^{\mathrm{up}}(\mathsf{x}_i) = M_2^{-1} \sum_{j=M_1+1}^{M_1+M_2} \max_a \Big\{ r^a(\mathsf{x}_i) + \gamma\Big(\widehat{V}_k^{\mathrm{up}}(Y_j^{\mathsf{x}_i,a}) - V^\pi(Y_j^{\mathsf{x}_i,a}) + M_1^{-1} \sum_{l=1}^{M_1} V^\pi(Y_l^{\mathsf{x}_i,a})\Big)\Big\},$$

where $\mathsf{X}_N = \{\mathsf{x}_1, \dots, \mathsf{x}_N\}$ are design points, either deterministic or sampled from some distribution on X. Then the next iterate $\widehat{V}_{k+1}^{\mathrm{up}}$ is obtained via an interpolation scheme based on the points $\widetilde{V}_{k+1}^{\mathrm{up}}(\mathsf{x}_1), \dots, \widetilde{V}_{k+1}^{\mathrm{up}}(\mathsf{x}_N)$ such that $\widehat{V}_{k+1}^{\mathrm{up}}(\mathsf{x}_i) = \widetilde{V}_{k+1}^{\mathrm{up}}(\mathsf{x}_i)$, $i = 1, \dots, N$. Note that interpolation is needed, since $\widehat{V}_{k+1}^{\mathrm{up}}$ has to be calculated at the (random) points $Y_j^{\mathsf{x}_i,a}$, which may not belong to the set X_N. In tabular case when $|\mathsf{X}| < \infty$ is not large one can omit the interpolation and take $\mathsf{X}_N = \mathsf{X}$. In a more general setting when $(\mathsf{X}, \rho_{\mathsf{X}})$ is an arbitrary compact metric space we suggest using an appropriate interpolation procedure. The one described below is particularly useful for our situation where the function to be interpolated is only Lipschitz continuous (due to the presence of the maximum). The *optimal* central interpolant for a function $f \in \mathrm{Lip}_{\rho_{\mathsf{X}}}(L)$ is defined as

$$I[f](x) \doteq (H_f^{\mathrm{low}}(x) + H_f^{\mathrm{up}}(x))/2, \tag{6}$$

where

$$H_f^{\mathrm{low}}(x) = \max_{\ell\in[N]} (f(\mathsf{x}_\ell) - L\rho_{\mathsf{X}}(x, \mathsf{x}_\ell)),\ H_f^{\mathrm{up}}(x) = \min_{\ell\in[N]} (f(\mathsf{x}_\ell) + L\rho_{\mathsf{X}}(x, \mathsf{x}_\ell)).$$

Note that $H_f^{\mathrm{low}}(x) \le f(x) \le H_f^{\mathrm{up}}(x)$, $H_f^{\mathrm{low}}, H_f^{\mathrm{up}} \in \mathrm{Lip}_{\rho_{\mathsf{X}}}(L)$ and hence $I[f] \in \mathrm{Lip}_{\rho_{\mathsf{X}}}(L)$. An efficient algorithm is proposed in Beliakov (2006) to compute the values of the interpolant $I[f]$ without knowing L in advance. The so-constructed interpolant achieves the bound

$$\|f - I[f]\|_{\mathsf{X}} \le L \max_{x\in\mathsf{X}} \min_{\ell\in[N]} \rho_{\mathsf{X}}(x, \mathsf{x}_\ell). \tag{7}$$

The quantity

$$\rho(\mathsf{X}_N, \mathsf{X}) \doteq \max_{x\in\mathsf{X}} \min_{\ell\in[N]} \rho_{\mathsf{X}}(x, \mathsf{x}_\ell) \tag{8}$$

in the r.h.s. of (7) is usually called covering radius (also known as the mesh norm or fill radius) of X_N with respect to X.

5 THEORETICAL RESULTS

In this section, we analyze the distance between $(\widehat{V}_k^{\mathrm{up}})_{k\in\mathbb{N}}$ and $V^\star$, where $\widehat{V}_k^{\mathrm{up}}(x)$ is the k-th iterate of Algorithm 2. Recall that $\mathsf{X}_N = \{\mathsf{x}_1, \dots, \mathsf{x}_N\}$ is a set of design points (random or deterministic) used in the iterations of Algorithm 2. First, note that with $\overline{V}_k^{\mathrm{up}}(x) \doteq \mathsf{E}\big[\widehat{V}_k^{\mathrm{up}}(x)\big]$ we have

$$\overline{V}_k^{\mathrm{up}}(x) \geq \max_a \big\{ r^a(x) + \gamma \mathrm{P}^a \overline{V}_{k-1}^{\mathrm{up}}(x) \big\}, \quad x \in \mathsf{X}_N, \quad k \in \mathbb{N}. \tag{9}$$

Furthermore, if $\widehat{V}_0^{\mathrm{up}}(x) \geq V^\star(x)$ for $x \in \mathsf{X}_N$, then $\overline{V}_k^{\mathrm{up}}(x) \geq V^\star(x)$ for any $x \in \mathsf{X}_N$ and $k \in \mathbb{N}$. Hence $\widehat{V}_k^{\mathrm{up}}$ is an upper-biased estimate of $V^\star$ for any $k \geq 0$.

Before stating our convergence results, we first state a number of technical assumptions.

A1. *We suppose that $(\mathsf{X}, \rho_{\mathsf{X}})$ and $(\mathsf{A}, \rho_{\mathsf{A}})$ are compact metric spaces. Moreover, $\mathsf{X} \times \mathsf{A}$ is equipped with some metric ρ, such that $\rho\big((x,a),(x',a)\big) = \rho_{\mathsf{X}}(x,x')$ for any $x, x' \in \mathsf{X}$ and $a \in \mathsf{A}$.*

We put special emphasis on the cases when X (resp. A) is either finite or $\mathsf{X} \subseteq [0,1]^{d_{\mathsf{X}}}$ with $d_{\mathsf{X}} \in \mathbb{N}$.

A2. *There exists a measurable mapping $\psi : \mathsf{X} \times \mathsf{A} \times \mathbb{R}^m \to \mathsf{X}$ such that $Y^{x,a} = \psi(x,a,\xi)$, where ξ is a random variable with values in $\Xi \subseteq \mathbb{R}^m$ and distribution P_ξ on Ξ, that is, $\psi(x,a,\xi) \sim \mathrm{P}^a(\cdot|x)$.*

A2 is a reparametrization assumption which is popular in RL, see e.g. Ciosek & Whiteson (2020), Heess et al. (2015), Liu et al. (2018) and the related discussions. This assumption is rather mild, since a large class of controlled Markov chains can be represented in the form of random iterative functions, see Douc et al. (2018).

A3. *For some positive constant $R_{\max}$ and all $a \in \mathsf{A}$, $\|r^a\|_{\mathsf{X}} \leq R_{\max}$.*

A4. *For some positive constants $L_\psi \leq 1, L_{\max}, L_\pi$ and all $a \in \mathsf{A}$, $\xi \in \Xi$,*

$$\mathrm{Lip}_{\rho_{\mathsf{X}}}(r^a(\cdot)) \leq L_{\max}, \quad \mathrm{Lip}_\rho(\psi(\cdot,\cdot,\xi)) \leq L_\psi, \quad \mathrm{Lip}_\rho((V^\pi \circ \psi)(\cdot,\cdot,\xi)) \leq L_\pi\,.$$

Remark 5.1. *If $|\mathsf{X}| < \infty$ and $|\mathsf{A}| < \infty$, the assumption A4 holds with $\rho_{\mathsf{X}}(x,x') = \mathbb{1}_{\{x\neq x'\}}$, $\rho((x,a),(x',a')) = \mathbb{1}_{\{(x,a)\neq(x',a')\}}$ and constants $L_\psi = 1$, $L_{\max} = R_{\max}, L_\pi = R_{\max}/(1-\gamma)$.*

The condition $L_\psi \leq 1$ implies a non-explosive behaviour of the Markov chain $(X_i)_{i\geq 0}$. This assumption is common in theoretical RL studies, see e.g. Pires & Szepesvári (2016). If $L_\psi < 1$, the corresponding Markov kernel contracts and there exists a unique invariant probability measure, see e.g. Jarner & Tweedie (2001).

Suppose that for each $k \in [K]$ we use an i.i.d. sample $\boldsymbol{\xi}_k = (\xi_{k,1} \dots, \xi_{k,M_1+M_2}) \sim \mathrm{P}_\xi^{\otimes(M_1+M_2)}$ to generate $Y_j^{x,a} = \psi(x,a,\xi_{k,j}), j \in [M_1+M_2]$ and these samples are independent for different k. For $\varepsilon > 0$, we denote by $\mathcal{N}(\mathsf{X}\times\mathsf{A}, \rho, \varepsilon)$ the covering number of the set $\mathsf{X}\times\mathsf{A}$ w.r.t. metric ρ, that is, the smallest cardinality of an ε-net of $\mathsf{X}\times\mathsf{A}$ w.r.t. ρ. Then $\log\mathcal{N}(\mathsf{X}\times\mathsf{A}, \rho, \varepsilon)$ is the metric entropy of $\mathsf{X}\times\mathsf{A}$ and

$$I_{\mathcal{D}} \doteq \int_0^{\mathsf{D}} \sqrt{\log\mathcal{N}\big(\mathsf{X}\times\mathsf{A}, \rho, u\big)}\,\mathrm{d}u$$

is the Dudley's integral. Here $\mathsf{D} \doteq \mathrm{diam}(\mathsf{X}\times\mathsf{A}) \doteq \max_{(x,a),(x',a')\in\mathsf{X}\times\mathsf{A}} \rho((x,a),(x',a'))$. Recall that $\rho(\mathsf{X}_N, \mathsf{X})$ defined in (8) is the covering radius of the set X_N w.r.t. X. We now state one of our main theoretical results.

Theorem 5.1. *Let A1 – A4 hold and suppose that $\mathrm{Lip}_{\rho_{\mathsf{X}}}(\widehat{V}_0^{\mathrm{up}}) \leq L_0$ with some constant $L_0 > 0$. Then for any $k \in \mathbb{N}$ and $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \|V^\pi - V^*\|_{\mathsf{X}} + \frac{I_{\mathcal{D}} + \mathsf{D}\sqrt{\log(1/\delta)}}{\sqrt{M_1}} + \rho(\mathsf{X}_N, \mathsf{X})\,. \tag{10}$$

In the above bound $\lesssim$ stands for inequality up to a constant depending on $\gamma, L_{\max}, L_\psi, L_\pi, L_0$ and $R_{\max}$. A precise dependence on the aforementioned constants can be found in (21) *in Appendix.*

Proof. The proof is given in Section A.1. □

Below we specify the result of Theorem 5.1 for two particular cases of MDPs, which are common in applications. The first one is an MDP with finite state and action spaces, and the second one is an MDP with the state space $\mathsf{X} \subseteq [0,1]^{d_{\mathsf{X}}}$.

Corollary 5.1. *Let $|\mathsf{X}|, |\mathsf{A}| < \infty$ and assume A2, A3. Then for any $k \in \mathbb{N}$ and $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that*

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \|V^\pi - V^*\|_{\mathsf{X}} + \sqrt{\frac{\log(|\mathsf{X}||\mathsf{A}|/\delta)}{M_1}}\,.$$

The precise expression for the constants can be found in (22) *in Appendix.*

Proof. The proof is given in Section A.2. □

Corollary 5.2. *Let $\mathsf{X} \subseteq [0,1]^{d_{\mathsf{X}}}$, $|\mathsf{A}| < \infty$, and consider $\rho_{\mathsf{X}}(x,x') = \|x-x'\|$, $\rho\big((x,a),(x',a')\big) = \|x-x'\| + \mathbb{1}_{\{a\neq a'\}}$. Assume that A2 – A4 hold and let $\mathsf{X}_N = \{\mathsf{x}_1,\dots,\mathsf{x}_N\}$ be a set of N points independently and uniformly distributed over X. If additionally $\mathrm{Lip}_{\rho_{\mathsf{X}}}(\widehat{V}_0^{\mathrm{up}}) \le L_0$ for some $L_0 > 0$, then for any $k \in \mathbb{N}$ and $\delta \in (0,1/2)$ it holds with probability at least $1-\delta$ that*

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \|V^\pi - V^*\|_{\mathsf{X}} + \sqrt{\frac{d_{\mathsf{X}}\log(d_{\mathsf{X}}|\mathsf{A}|/\delta)}{M_1}} + \sqrt{d_{\mathsf{X}}}\left(N^{-1}\log(1/\delta)\log N\right)^{1/d_{\mathsf{X}}}\,.$$

Precise expression for the constants can be found in (23) *in Appendix.*

Proof. The proof is given in Section A.2. □

Variance of the estimator and confidence bounds. Our next step is to bound the variance of the estimator $\widehat{V}_k^{\mathrm{up}}(x)$. We additionally assume that $\mathsf{X}\times\mathsf{A}$ is a parametric class with the metric entropy satisfying the following assumption:

A5. *There exist a constant $C_{\mathsf{X},\mathsf{A}} > 1$ such that for any $\varepsilon \in (0,\mathsf{D})$,*

$$\log \mathcal{N}(\mathsf{X}\times\mathsf{A},\rho,\varepsilon) \le C_{\mathsf{X},\mathsf{A}}\log(1+1/\varepsilon).$$

Denote the r.h.s. of (10) by σ_k, that is,

$$\sigma_k \doteq \gamma^k\big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \|V^\pi - V^*\|_{\mathsf{X}} + \frac{I_{\mathcal{D}} + \mathsf{D}}{\sqrt{M_1}} + \rho(\mathsf{X}_N,\mathsf{X})\,. \tag{11}$$

The next theorem implies that $\mathsf{Var}\big[\widehat{V}_k^{\mathrm{up}}(x)\big]$ can be much smaller than the standard rate $1/M_2$, provided that V^π is close to V^* and M_1, N, K are large enough.

Theorem 5.2. *Let A1 – A5 hold and assume additionally $\mathrm{Lip}_{\rho_{\mathsf{X}}}(\widehat{V}_0^{\mathrm{up}}) \le L_0$ for some $L_0 > 0$. Then*

$$\max_{x\in\mathsf{X}} \mathsf{Var}\big[\widehat{V}_k^{\mathrm{up}}(x)\big] \le \mathrm{C}\,\sigma_k^2\log(\mathrm{e}\vee\sigma_k^{-1})M_2^{-1}\,, \tag{12}$$

where the constant C depends on $C_{\mathsf{X},\mathsf{A}}, \gamma, L_{\max}, L_\psi, L_\pi, L_0$ and $R_{\max}$. A precise expression for C can be found in (30) *in appendix.*

Proof. The proof is given in Section A.3. □

Corollary 5.3. *Recall that $\widehat{V}_k^{\mathrm{up}}$ is an upper biased estimate of $V^\star$ in a sense that $\overline{V}_k^{\mathrm{up}}(x) \ge V^\star(x)$ provided $\widehat{V}_0^{\mathrm{up}}(x) \ge V^\star(x)$ for $x \in \mathsf{X}_N$. Together with Theorem 5.2, it implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$V^\pi(x) \le V^\star(x) \le \widehat{V}_k^{\mathrm{up}}(x) + \sigma_k\sqrt{\mathrm{C}\log(\mathrm{e}\vee\sigma_k^{-1})\delta^{-1}M_2^{-1}} + L_V\rho(\mathsf{X}_N,\mathsf{X})\mathbb{1}_{\{x\notin\mathsf{X}_N\}}, \quad x\in\mathsf{X}, \tag{13}$$

where the constant L_V is given by (18) *in appendix.*

Note that bounds of type (13) are known in the literature only in the case of specific policies π. For example, Wainwright (2019) proves bounds of this type for greedy policies in tabular Q-learning. At the same time, (13) holds for arbitrary policy π and general state space.

Now we aim to track the dependence of the r.h.s. of (13) on the quantity $\|V^\pi - V^\star\|_{\mathsf{X}}$ for MDPs with finite state and action spaces. The following proposition implies that σ_k scales (almost) linearly with $\|V^\pi - V^\star\|_{\mathsf{X}}$.

Proposition 5.1. *Let* $|\mathsf{X}|, |\mathsf{A}| < \infty$*, assume A2, A3, and* $\left\|\widehat{V}_0^{\mathrm{up}}\right\|_{\mathsf{X}} \leq R_{\max}(1-\gamma)^{-1}$*. Then for* k *and* M_1 *large enough, and* $\delta \in (0,1)$ *it holds with probability at least* $1-\delta$ *that*

$$\sigma_k \lesssim \|V^\pi - V^\star\|_{\mathsf{X}} \,. \tag{14}$$

The precise bounds for k *and* M_1 *can be found in* (32).

Proof. The proof is given in Section A.4. □

6 Numerical results

In this section we demonstrate the performance of Algorithm 2 on several tabular and continuous state-space RL problems. We construct the upper confidence bounds for policy π, coming from the Value iteration procedure and Reinforce algorithm. Recall that the closer policy π is to the optimal one $\pi^\star$, the smaller is the difference between $V^\pi(x)$ and $V^{\mathrm{up},\pi}(x)$.

Discrete state-space MDPs We consider 3 popular tabular environments: Garnet (Archibald et al. (1995)), Chain (Rowland et al. (2020)) and NRooms(Domingues et al. (2021)). Detailed specifications can be found in Appendix B. For each environment we perform K updates of the Value iteration (see Appendix B for details) with known transition kernel P^a. We denote the k-th step estimate of the action-value function as $\hat{Q}_k(x,a)$ and denote π_k the greedy policy w.r.t. $\hat{Q}_k(x,a)$. Then we evaluate the policies π_k with the Algorithm 2 for certain iteration numbers k. We don't use the approximation step due to the small state space size. Experimental details are provided in Table 1 in the appendix. Figure 1 displays the gap between $V^{\pi_k}(x)$ and $V^{\mathrm{up},\pi_k}(x)$, which converges to zero while π^k converges to the optimal policy $\pi^\star$.

For the NRooms environment we learn a sub-optimal policy π using VI algorithm and then in the third room w.p. $1/2$ replace this policy with uniformly random π_c. Indeed, this leads to a worse policy in this specific room and, as expected, should increase our upper bounds. To illustrate the effect, we compute an accurate upper bounds obtained with UVIP algorithm. We can see on the Figure 1(bottom) that UVIP shows the sub-optimality of our policy in the specific room with a slight change of values in other rooms.

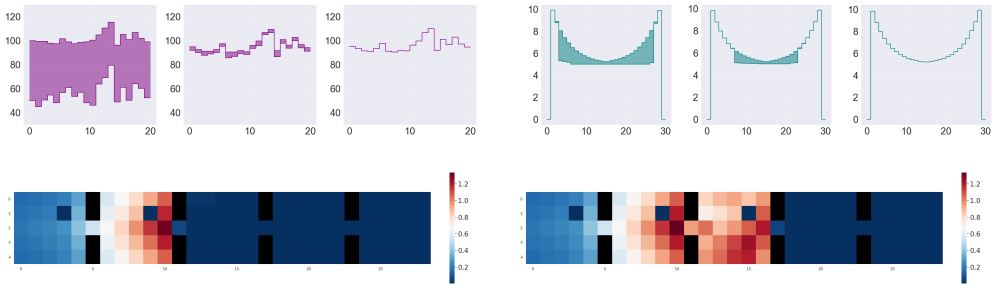

Figure 1: The difference between $V^{\mathrm{up},\pi_i}(x)$ and $V^{\pi_i}(x)$. X-axis represents states in a discrete environment for all pictures. Each group of three pictures of the same color demonstrates the process of learning the policy from the first iteration to the last. *First row*: Evaluation of the policies during the process of Value iteration for Garnet (left) and Chain environments (right). The policies are obtained greedily from $Q_i(x,a)$ function at the i-th step. *Second row*: Comparison of gap between V^π and $V^{\mathrm{up},\pi}$ for the learned policy π and corrupted π_c on NRooms environment. Color on this plot represents the value of $V^{\mathrm{up},\pi} - V^\pi$.

Continuous state-space MDPs We consider the AI Gym CartPole and Acrobot environments (Brockman et al. (2016)). Their descriptions can be found in Appendix B. For CartPole we evaluate A2C algorithm policy π_1 (Mnih et al. (2016)), linear deterministic policy (LD) π_2 described in B and random uniform policy π_3. Figure 2 (left) indicates superior quality of π_2, sort of instability introduced by A2C training in π_1 and low quality of π_3. We also evaluated a policy for Acrobot given by A2C as well as a policy from Dueling DQN (Wang et al. (2016)) (Fig. 2 (right)). From the plots we can conclude that the both policies are good, but far from optimal.

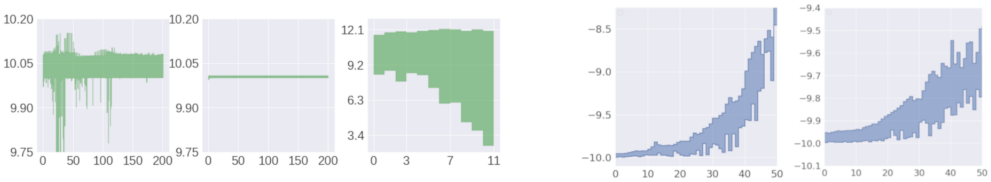

Figure 2: Upper and lower bounds of the three different policies. *Left:* For CartPole π_1, π_2, π_3 policies, respectively. For horizontal axis we sample single trajectory according to the policy. *Right:* For Acrobot Dueling DQN and A2C policies, respectively. We evaluated the bounds for 50 first states of the trajectory for each algorithm.

Additionally, we compare policies in TwinRooms environment from *rlberry* (Domingues et al. (2021)). We obtain two policies π_1 and π_2 after running Kernel- UCBVI(Domingues et al. (2022)) algorithm on 2500 and 5000 iteration steps. The result on Figure 3 shows that the policy π_2 have tighter upper bounds after more learning steps, and thus, we can conclude that it has better performance. Also, our upper bounds highlight the regions of the state space which are less studied with our policy.

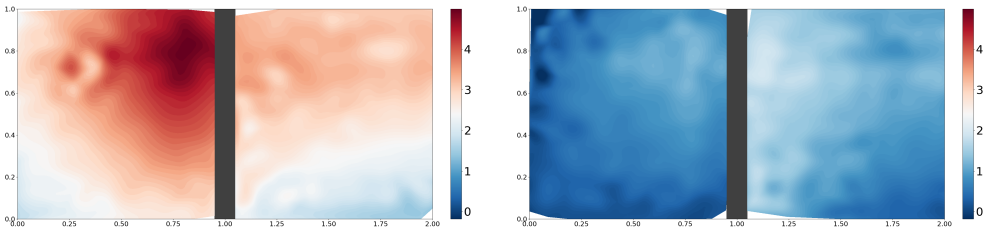

Figure 3: We illustrate the gap between $V^{up,\pi}$ and V^{π} in TwinRooms environment. Color on this plot represents the value of $V^{\mathrm{up},\pi} - V^{\pi}$. On the left and right hand sides we show this quantity for π_1 and π_2 corresponding. We obtain π_1 and π_2 after 2500 and 5000 learning steps of the Kernel-UCBVI algorithm.

7 Conclusion and Future Work

In this work we propose a new approach towards model-free evaluation of the agent's policies in RL, based on upper solutions to the Bellman optimality equation (1). To the best of our knowledge, the UVIP is the first procedure which allows to construct the non-asymptotic confidence intervals for the optimal value function $V^{\star}$ based on the value function corresponding to an arbitrary policy π. In our analysis we consider only infinite-horizon MDPs and assume that sampling from the conditional distribution $\mathrm{P}^a(\cdot|x)$ is feasible for any $x \in \mathsf{X}$ and $a \in \mathsf{A}$. A promising future research direction is to generalize UVIP to the case of finite-horizon MDPs combining it with the idea of Real-time dynamic programming (see Efroni et al. (2019)). Moreover, the plain Monte Carlo estimates are not necessarily the most efficient way to estimate the outer expectation in Algorithm 1. Other stochastic approximation techniques could also be applied to approximate the solution of (2).

8 Acknowledgments

The article was prepared within the framework of the HSE University Basic Research Program. This research was supported in part through computational resources of HPC facilities at HSE University Kostenetskiy et al. (2021).

REFERENCES

András Antos, Rémi Munos, and Csaba Szepesvári. Fitted q-iteration in continuous action-space mdps. 2007.

T. W. Archibald, K. I. M. McKinnon, and L. C. Thomas. On the generation of Markov Decision Processes. *The Journal of the Operational Research Society*, 46(3):354–361, 1995. ISSN 01605682, 14769360. URL http://www.jstor.org/stable/2584329.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. 70:263–272, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/azar17a.html.

Gleb Beliakov. Interpolation of Lipschitz functions. *Journal of computational and applied mathematics*, 196(1):20–44, 2006.

Denis Belomestny and John Schoenmakers. *Advanced Simulation-Based Methods for Optimal Stopping and Control: With Applications in Finance*. Springer, 2018.

Dimitri P Bertsekas and Steven E Shreve. Stochastic optimal control, volume 139 of mathematics in science and engineering, 1978.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996. ISBN Athena Scientific. URL http://www.athenasc.com/ndpbook.html.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Kamil Ciosek and Shimon Whiteson. Expected policy gradients for reinforcement learning. *Journal of Machine Learning Research*, 21(52):1–51, 2020. URL http://jmlr.org/papers/v21/18-012.html.

Omar Darwiche Domingues, Yannis Flet-Berliac, Edouard Leurent, Pierre Ménard, Xuedong Shang, and Michal Valko. rlberry - A Reinforcement Learning Library for Research and Education, 10 2021. URL https://github.com/rlberry-py/rlberry.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-based reinforcement learning: A finite-time analysis, 2022. URL https://arxiv.org/abs/2004.05599.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/25caef3a545a1fff2ff4055484f0e758-Paper.pdf.

Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. 28, 2015. URL https://proceedings.neurips.cc/paper/2015/file/148510031349642de5ca0c544f31b2ef-Paper.pdf.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. URL http://jmlr.org/papers/v11/jaksch10a.html.

SF Jarner and RL Tweedie. Locally contracting iterated functions and stability of markov chains. *Journal of applied probability*, pp. 494–507, 2001.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

PS Kostenetskiy, RA Chulkevich, and VI Kozyrev. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, pp. 012050. IOP Publishing, 2021.

Hao Liu, Yihao Feng, Yi Mao, Dengyong Zhou, Jian Peng, and Qiang Liu. Action-dependent control variates for policy optimization via stein identity. February 2018. URL https://www.microsoft.com/en-us/research/publication/action-dependent-control-variates-policy-optimization-via-stein-identity/.

Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4504–4511, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. *arXiv e-prints*, art. arXiv:1312.5602, December 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/mniha16.html.

Bernardo Ávila Pires and Csaba Szepesvári. Policy error bounds for model-based reinforcement learning with factored linear models. In *Conference on Learning Theory*, pp. 121–151. PMLR, 2016.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

A. Reznikov and E. B. Saff. The covering radius of randomly distributed points on a manifold. *Int. Math. Res. Not. IMRN*, (19):6065–6094, 2016. ISSN 1073-7928. doi: 10.1093/imrn/rnv342. URL https://doi.org/10.1093/imrn/rnv342.

L. Rogers. Pathwise stochastic optimal control. *SIAM J. Control and Optimization*, 46:1116–1132, 01 2007. doi: 10.1137/050642885.

Mark Rowland, Anna Harutyunyan, Hado van Hasselt, Diana Borsa, Tom Schaul, Rémi Munos, and Will Dabney. Conditional importance sampling for off-policy learning, 2020.

Ibrahim El Shar and Daniel Jiang. Lookahead-bounded q-learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8665–8675. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/shar20a.html.

R. S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Richard Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3: 9–44, 08 1988. doi: 10.1007/BF00115009.

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997. ISSN 2334-3303. doi: 10.1109/9.580874.

Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596. An introduction with applications in data science, With a foreword by Sara van de Geer.

Martin J. Wainwright. Variance-reduced q-learning is minimax optimal. 2019.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling network architectures for deep reinforcement learning, 2016.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

A PROOF OF THE MAIN RESULTS

Throughout this section we will use additional notations. Let $\psi_2(x) = e^{x^2} - 1$, $x \in \mathbb{R}$. For r.v. η we denote $\|\eta\|_{\psi_2} \doteq \inf\{t > 0 : \mathsf{E}\left[\exp\{\eta^2/t^2\}\right] \le 2\}$ the Orlicz 2-norm. We say that η is a *sub-Gaussian random variable* if $\|\eta\|_{\psi_2} < \infty$. In particular, this implies that for some constants $C, c > 0$, $\mathrm{P}(|\eta| \ge t) \le 2\exp\{-ct^2/\|\eta\|_{\psi_2}^2\}$ and $\mathsf{E}^{1/p}[|\eta|^p] \le C\sqrt{p}\|\eta\|_{\psi_2}$ for all $p \ge 1$. Consider a random process $(X_t)_{t\in T}$ on a metric space (T, d). We say that the process has *sub-Gaussian increments* if there exists $K \ge 0$ such that

$$\|X_t - X_s\|_{\psi_2} \le K\mathsf{d}(t,s), \quad \forall t, s \in T.$$

We start from the following proposition.

Proposition A.1. *Under A1 – A4 for any $M \in \mathbb{N}$ and $p \ge 1$*

$$\mathsf{E}^{1/p}\Big[\Big\|\frac{1}{M}\sum_{l=1}^{M}[V^\pi(\psi(\cdot,\cdot,\xi_l)) - \mathsf{E}V^\pi(\psi(\cdot,\cdot,\xi_l))]\Big\|_{\mathsf{X}\times\mathcal{A}}^p\Big] \lesssim \frac{L_\pi I_{\mathcal{D}} + \{L_\pi \mathsf{D} + R_{\max}/(1-\gamma)\}\sqrt{p}}{\sqrt{M}}.$$

Proof. We apply the empirical process methods. To simplify notations we denote

$$Z(x,a) = \frac{1}{\sqrt{M}}\sum_{\ell=1}^{M}[V^\pi(\psi(x,a,\xi_\ell)) - \mathsf{E}V^\pi(\psi(x,a,\xi_\ell))],\ (x,a) \in \mathsf{X}\times\mathsf{A},$$

that is, $Z(x,a)$ is a random process on the metric space $(\mathsf{X}\times\mathsf{A}, \rho)$. Below we show that the process $Z(x,a)$ has sub-Gaussian increments. In order to show it, let us introduce for $\ell \in [M]$

$$Z_\ell \doteq [V^\pi(\psi(x,a,\xi_l)) - \mathsf{E}V^\pi(\psi(x,a,\xi_l))] - [V^\pi(\psi(x',a',\xi_l)) - \mathsf{E}V^\pi(\psi(x',a',\xi_l))]\,.$$

Clearly, by A4,

$$\|Z_\ell\|_{\psi_2} \lesssim L_\pi \rho((x,a),(x',a'))\,,$$

that is, Z_ℓ is a sub-Gaussian r.v. for any $\ell \in [M]$. Since $Z(x,a) - Z(x',a') = M^{-1/2}\sum_{\ell=1}^{M} Z_\ell$ is a sum of independent sub-Gaussian r.v, we may apply (Vershynin, 2018, Proposition 2.6.1 and Eq. (2.16))) to obtain that $Z(x,a)$ has sub-Gaussian increments with parameter $K \asymp L_\pi$. Fix some $(x_0, a_0) \in \mathsf{X}\times\mathsf{A}$. By the triangular inequality,

$$\sup_{(x,a)\in\mathsf{X}\times\mathsf{A}} |Z(x,a)| \le \sup_{(x,a),(x',a')\in\mathsf{X}\times\mathsf{A}} |Z(x,a) - Z(x',a')| + Z(x_0,a_0). \tag{15}$$

By the Dudley integral inequality, e.g. (Vershynin, 2018, Theorem 8.1.6), for any $\delta \in (0,1)$,

$$\sup_{(x,a),(x',a')\in\mathsf{X}\times\mathsf{A}} |Z(x,a) - Z(x',a')| \lesssim L_\pi\left[I_{\mathcal{D}} + \mathsf{D}\sqrt{\log(2/\delta)}\right].$$

holds with probability at least $1-\delta$. Again, under A3, $Z(x_0,a_0)$ is a sum of i.i.d. bounded centered random variables with ψ_2-norm bounded by $R_{\max}/(1-\gamma)$. Hence, applying Hoeffding's inequality, e.g. (Vershynin, 2018, Theorem 2.6.2.), for any $\delta \in (0,1)$,

$$|Z(x_0,a_0)| \lesssim R_{\max}\sqrt{\log(1/\delta)}/(1-\gamma)$$

holds with probability $1-\delta$. The last two inequalities and (15) imply the statement. □

A.1 PROOF OF THEOREM 5.1

Fix $p \ge 2$ and denote for any $k \in \mathbb{N}$, $\mathcal{M}_k \doteq \mathsf{E}^{1/p}[\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}}^p]$. For any $x \in \mathsf{X}$, we introduce

$$\widetilde{V}_{k+1}^{\mathrm{up},\pi}(x) = \frac{1}{M_2}\sum_{j=M_1+1}^{M_1+M_2}\max_a\left\{r^a(x) + \gamma\Big(\widehat{V}_k^{\mathrm{up}}(Y_j^{x,a}) - V^\pi(Y_j^{x,a}) + \frac{1}{M_1}\sum_{\ell=1}^{M_1}V^\pi(Y_\ell^{x,a})\Big)\right\}$$

Recall that $Y_j^{x,a} = \psi(x,a,\xi_{k,j})$, $j \in [M_1+M_2]$ for independent random variables $(\xi_{k,j})$, thus we can write

$$\widetilde{V}_{k+1}^{\mathrm{up},\pi}(x) = \frac{1}{M_2}\sum_{j=M_1+1}^{M_1+M_2} R_k^x(\xi_{k,j};\xi_{k,1},\ldots,\xi_{k,M_1})\,. \tag{16}$$

We first calculate $L_{k+1} \doteq \mathrm{Lip}_\rho(\widetilde{V}^{\mathrm{up},\pi}_{k+1})$ for any $k \in \mathbb{N}$. Since under A4, $\mathrm{Lip}_\rho((V^\pi \circ \psi)(\cdot,\cdot,\xi)) \le L_\pi$, and using (16),

$$L_{k+1} \le L_{\max} + \gamma(L_k L_\psi + 2L_\pi)\,. \tag{17}$$

Expanding (17) and using the assumptions of Theorem 5.1, we obtain

$$L_{k+1} \le \frac{L_{\max} + 2\gamma L_\pi}{1-\gamma L_\psi} + (\gamma L_\psi)^k L_0\,, \quad k \in \mathbb{N}\,.$$

Using that $\gamma L_\psi < 1$, the maximal Lipshitz constant of $\widetilde{V}^{\mathrm{up},\pi}_k(x), k \in \mathbb{N}$ is uniformly bounded by

$$L_V = \frac{L_{\max} + 2\gamma L_\pi}{1-\gamma L_\psi} + L_0\,. \tag{18}$$

Using (16) and (4), for any $x \in \mathsf{X}$ and $j = M_1+1,\dots,M_1+M_2$.

$$\begin{aligned}
&\mathsf{E}^{1/p}[|R^x_k(\xi_{k,j};\xi_{k,1},\dots,\xi_{k,M_1}) - V^*(x)|^p] \le \\
&\mathsf{E}^{1/p}\Bigg[\Bigg|\max_a \Big\{r^a(x) + \gamma\Big(\widehat{V}^{\mathrm{up}}_k(Y^{x,a}_j) - V^\pi(Y^{x,a}_j) + M_1^{-1}\sum\nolimits_{\ell=1}^{M_1} V^\pi(Y^{x,a}_\ell)\Big)\Big\} - \\
&\qquad \max_a \{r^a(x) + \gamma(V^\star(Y^{x,a}) - V^\star(Y^{x,a}) + \mathrm{P}^a V^\star(x)\}\Bigg|^p\Bigg]
\end{aligned}$$

Hence, with the Minkowski inequality and $|\mathrm{P}^a V^\star(x) - \mathsf{E}V^\pi(\psi(x,a,\cdot))| \le \|V^\pi - V^*\|_{\mathsf{X}}$, we get

$$\begin{aligned}
\mathsf{E}^{1/p}[|R^x_k(\xi_{k,j};\xi_{k,1},\dots,\xi_{k,M_1}) - V^*(x)|^p] &\le \gamma\mathcal{M}_k + 2\gamma\,\|V^\pi - V^*\|_{\mathsf{X}} \\
&+ \gamma\mathsf{E}^{1/p}\Big[\Big\|M_1^{-1}\sum\nolimits_{\ell=1}^{M_1}[V^\pi(\psi(\cdot,\cdot,\xi_{k,\ell})) - \mathsf{E}V^\pi(\psi(\cdot,\cdot,\xi_{k,\ell}))]\Big\|^p_{\mathsf{X}\times\mathcal{A}}\Big]
\end{aligned} \tag{19}$$

To analyze the last term we use the empirical process methods. By Proposition A.1, we get

$$\mathsf{E}^{1/p}\Big[\Big\|\frac{1}{M_1}\sum_{\ell=1}^{M_1}[V^\pi(\psi(\cdot,\cdot,\xi_{k,\ell})) - \mathsf{E}V^\pi(\psi(\cdot,\cdot,\xi_{k,\ell}))]\Big\|^p_{\mathsf{X}\times\mathcal{A}}\Big] \lesssim \frac{L_\pi I_{\mathcal{D}} + \{L_\pi \mathsf{D} + R_{\max}/(1-\gamma)\}\sqrt{p}}{\sqrt{M_1}}.$$

Furthermore, with (7) we construct Lipshitz interpolant $\widehat{V}^{\mathrm{up}}_{k+1}$, such that

$$|\widehat{V}^{\mathrm{up}}_{k+1}(x) - \widetilde{V}^{\mathrm{up},\pi}_{k+1}(x)| \quad \lesssim \quad L_{k+1}\rho(\mathsf{X}_N,\mathsf{X})\,.$$

Combining the above estimates, we get

$$\mathcal{M}_{k+1} \quad \lesssim \quad \gamma\mathcal{M}_k + \gamma\,\|V^\pi - V^*\|_{\mathsf{X}} + \gamma\frac{L_\pi I_{\mathcal{D}} + \{L_\pi \mathsf{D} + R_{\max}/(1-\gamma)\}\sqrt{p}}{\sqrt{M_1}} + L_{k+1}\rho(\mathsf{X}_N,\mathsf{X})\,.$$

Iterating this inequality,

$$\begin{aligned}
\mathsf{E}^{1/p}[\|\widehat{V}^{\mathrm{up}}_k - V^*\|^p_{\mathsf{X}}] \lesssim \gamma^k\big\|\widehat{V}^{\mathrm{up}}_0 - V^*\big\|_{\mathsf{X}} + \frac{\gamma}{1-\gamma}\,\|V^\pi - V^*\|_{\mathsf{X}} + \\
\frac{\gamma L_\pi I_{\mathcal{D}} + \gamma\{L_\pi \mathsf{D} + R_{\max}/(1-\gamma)\}\sqrt{p}}{\sqrt{M_1}(1-\gamma)} + \frac{L_V}{1-\gamma}\rho(\mathsf{X}_N,\mathsf{X})\,.
\end{aligned} \tag{20}$$

Applying Markov's inequality with $p \asymp \log(1/\delta)$, we get that for any $k \in \mathbb{N}$ and $\delta \in (0,1)$,

$$\begin{aligned}
\|\widehat{V}^{\mathrm{up}}_k - V^*\|_{\mathsf{X}} \lesssim \gamma^k\big\|\widehat{V}^{\mathrm{up}}_0 - V^*\big\|_{\mathsf{X}} + \frac{\gamma}{1-\gamma}\,\|V^\pi - V^*\|_{\mathsf{X}} + \\
\frac{\gamma L_\pi I_{\mathcal{D}} + \gamma\{L_\pi \mathsf{D} + R_{\max}/(1-\gamma)\}\sqrt{\log(1/\delta)}}{\sqrt{M_1}(1-\gamma)} + \frac{L_V}{1-\gamma}\rho(\mathsf{X}_N,\mathsf{X})\,.
\end{aligned} \tag{21}$$

holds with probability at least $1-\delta$, where the constant L_V is given in (18). This yields the statement of the theorem.

A.2 Proof of Corollary 5.1 and Corollary 5.2

Proof of Corollary 5.1. Consider $\rho((x,a),(x',a')) = \mathbb{1}_{\{(x,a)\neq(x',a')\}}$ and $\mathsf{X}_N = \mathsf{X}$, that is, we bypass the approximation step. Then $\mathsf{D} = 1$, $I_{\mathcal{D}} \lesssim \sqrt{\log(|\mathsf{X}||\mathsf{A}|)}$, $\rho(\mathsf{X}_N, \mathsf{X}) = 0$, and $r^a(\cdot)$ is Lipshitz w.r.t. ρ_{X} with $L_{\max} \leq R_{\max}$. Moreover, one can take $L_\psi = 1$ and $L_\pi = R_{\max}/(1-\gamma)$ in Assumption A4. Hence, A1 – A4 are valid and one may apply Theorem 5.1. Bound (21) in this case writes as

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \frac{\gamma}{1-\gamma}\|V^\pi - V^*\|_{\mathsf{X}} + \frac{\gamma R_{\max}(\sqrt{\log(|\mathsf{X}||\mathsf{A}|)}+2)\sqrt{\log(1/\delta)}}{\sqrt{M_1}(1-\gamma)^2}. \quad (22)$$

□

Proof of Corollary 5.2. It is easy to see that $\mathsf{D} \leq \sqrt{d_{\mathsf{X}}}+1$, $I_{\mathcal{D}} \lesssim \sqrt{d_{\mathsf{X}}\log|\mathsf{A}|} + \sqrt{d_{\mathsf{X}}\log d_{\mathsf{X}}}$. Proposition A.3 implies that for any $\delta \in (0,1)$, $\rho(\mathsf{X}_N,\mathsf{X}) \lesssim \sqrt{d_{\mathsf{X}}}\left(N^{-1}\log(1/\delta)\log N\right)^{1/d_{\mathsf{X}}}$. Substituting into (21), we obtain

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \frac{\gamma}{1-\gamma}\|V^\pi - V^*\|_{\mathsf{X}} + \frac{L_V\sqrt{d_{\mathsf{X}}}\left(N^{-1}\log(1/\delta)\log N\right)^{1/d_{\mathsf{X}}}}{1-\gamma} + \frac{\gamma L_\pi(\sqrt{d_{\mathsf{X}}\log|\mathsf{A}|}+\sqrt{d_{\mathsf{X}}\log d_{\mathsf{X}}}) + \gamma\{L_\pi\sqrt{d_{\mathsf{X}}} + R_{\max}/(1-\gamma)\}\sqrt{\log(1/\delta)}}{\sqrt{M_1}(1-\gamma)}, \quad (23)$$

where L_V is given in (18). □

A.3 Proof of Theorem 5.2

We use definition of $\widetilde{V}_{k+1}^{\mathrm{up},\pi}(x)$ and $R_k^x(\xi_{k,j};\xi_{k,1},\ldots,\xi_{k,M_1})$ from Theorem 5.1. To simplify notations we denote $\boldsymbol{\xi}_{k,M_1} = (\xi_{k,1},\ldots,\xi_{k,M_1})$ and $\boldsymbol{\xi}_{k,M_2} = (\xi_{k,M_1+1},\ldots,\xi_{k,M_1+M_2})$. In these notations $\boldsymbol{\xi}_k = (\boldsymbol{\xi}_{k,M_1},\boldsymbol{\xi}_{k,M_2})$. Recall that, by the construction, $\widetilde{V}_{k+1}^{\mathrm{up},\pi}(x)$ can be evaluated only at the points $x \in \{\mathsf{x}_1,\ldots,\mathsf{x}_N\}$. By the definition,

$$\widehat{V}_{k+1}^{\mathrm{up}}(x) = \min_{\ell\in[N]}\left\{\widetilde{V}_{k+1}^{\mathrm{up},\pi}(\mathsf{x}_\ell) + L_V\rho_{\mathsf{X}}(\mathsf{x}_\ell,x)\right\}, \quad (24)$$

where the constant L_V is given in (18). We rewrite $\widetilde{V}_{k+1}^{\mathrm{up},\pi}(x)$ as follows

$$\begin{aligned}\widetilde{V}_{k+1}^{\mathrm{up},\pi}(x) &= \frac{1}{M_2}\sum_{j=M_1+1}^{M_1+M_2}\{R_k^x(\xi_{k,j};\boldsymbol{\xi}_{k,M_1}) - \mathsf{E}[R_k^x(\xi_{k,j};\boldsymbol{\xi}_{k,M_1})]\} \\ &+ \mathsf{E}[R_k^x(\xi;\boldsymbol{\xi}_{k,M_1})] =: T_k^x(\boldsymbol{\xi}_k) + \mathsf{E}[R_k^x(\xi;\boldsymbol{\xi}_{k,M_1})],\end{aligned}$$

where ξ is an i.i.d. copy of $\xi_{k,j}$. Conditioned on $\overline{\mathcal{G}}_k \doteq \mathcal{G}_{k-1}\cup\sigma\{\boldsymbol{\xi}_{k,M_1}\}$, $T_k^x(\boldsymbol{\xi}_{k,M_1},\boldsymbol{\xi}_{k,M_2})$ is the sum of i.i.d. centered random variables. In what follows we will often omit the arguments $\boldsymbol{\xi}_{k,M_1}$ and/or $\boldsymbol{\xi}_{k,M_2}$ from the notations of functions T_k^x. Using representation (24),

$$\begin{aligned}\mathrm{Var}[\widehat{V}_{k+1}^{\mathrm{up}}(x)] &= \mathrm{Var}\left[\min_{\ell\in[N]}\left\{\widetilde{V}_{k+1}^{\mathrm{up},\pi}(\mathsf{x}_\ell) + L_V\rho_{\mathsf{X}}(\mathsf{x}_\ell,x)\right\}\right] \\ &\leq \mathsf{E}\left[\left(\min_{\ell\in[N]}\left\{\widetilde{V}_{k+1}^{\mathrm{up},\pi}(\mathsf{x}_\ell) + L_V\rho_{\mathsf{X}}(\mathsf{x}_\ell,x)\right\} - \min_{\ell\in[N]}\left\{\mathsf{E}[R_k^x(\xi;\boldsymbol{\xi}_{k,M_1})] + L_V\rho_{\mathsf{X}}(\mathsf{x}_\ell,x)\right\}\right)^2\right].\end{aligned}$$

Hence, from the previous inequality and the definition of $\widetilde{V}_{k+1}^{\mathrm{up},\pi}(x)$,

$$\mathrm{Var}[\widehat{V}_{k+1}^{\mathrm{up}}(x)] \leq \mathsf{E}[\sup_{\ell\in[N]}|T_k^{\mathsf{x}_\ell}(\boldsymbol{\xi}_{k,M_1},\boldsymbol{\xi}_{k,M_2})|^2] \leq \mathsf{E}[\sup_{x\in\mathsf{X}}|T_k^x(\boldsymbol{\xi}_{k,M_1},\boldsymbol{\xi}_{k,M_2})|^2].$$

To estimate the r.h.s. of the previous inequality we again apply the empirical process method. We first note that for any $x, x' \in \mathsf{X}$,

$$\sup_{\boldsymbol{\xi} \in \Xi^{M_1+M_2}} |T_k^x(\boldsymbol{\xi}) - T_k^{x'}(\boldsymbol{\xi})| \leq L_T \rho_{\mathsf{X}}(x, x'), \tag{25}$$

where

$$L_T = L_{\max} + \gamma(L_V + 2L_\pi)\,. \tag{26}$$

Now we freeze the coordinates $\boldsymbol{\xi}_{M_1}$ and consider $T_k^x(\boldsymbol{\xi}_{M_1}, \cdot)$ as a function on Ξ^{M_2}, parametrized by $x \in \mathsf{X}$. Introduce a parametric class of functions

$$\mathcal{T}_{k,\boldsymbol{\xi}_{M_1}} \doteq \left\{T_k^x(\boldsymbol{\xi}_{M_1}, \cdot) : \Xi^{M_2} \to \mathbb{R}, x \in \mathsf{X}\right\}.$$

For notational simplicity we will omit dependencies on k and $\boldsymbol{\xi}_{M_1}$, and simply write $T^x(\cdot) = T_k^x(\boldsymbol{\xi}_{M_1}, \cdot)$. Note that the functions in $\mathcal{T}_{k,\boldsymbol{\xi}_{M_1}}$ are Lipshitz w.r.t. the uniform metric

$$\rho_{\mathcal{T}_{k,\boldsymbol{\xi}_{M_1}}}(T^x(\cdot), T^{x'}(\cdot)) = \sup_{\boldsymbol{\xi}_{M_2} \in \Xi^{M_2}} |T^x(\boldsymbol{\xi}_{M_2}) - T^{x'}(\boldsymbol{\xi}_{M_2})|, \quad T^x(\cdot),\, T^{x'}(\cdot) \in \mathcal{T}_{k,\boldsymbol{\xi}_{M_1}}\,.$$

To estimate $\operatorname{diam}(\mathcal{T}_{k,\boldsymbol{\xi}_{M_1}})$ we proceed as follows. Denote $\tilde{R}_k^x(\xi; \boldsymbol{\xi}_{k,M_1}) = R_k^x(\xi; \boldsymbol{\xi}_{k,M_1}) - \mathsf{E}\left[R_k^x(\xi; \boldsymbol{\xi}_{k,M_1})\right]$. Using (19), we get an upper bound

$$\begin{aligned}\left|\tilde{R}_k^x(\xi; \boldsymbol{\xi}_{k,M_1})\right| \lesssim \gamma \|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} + 2\gamma \left\|V^\pi - V^*\right\|_{\mathsf{X}} + \\ \gamma \Big\| M_1^{-1} \sum_{l=1}^{M_1} [V^\pi(\psi(\cdot,\cdot,\xi_{k,l})) - \mathsf{E}V^\pi(\psi(\cdot,\cdot,\xi_{k,l}))] \Big\|_{\mathsf{X}\times\mathcal{A}}.\end{aligned} \tag{27}$$

We denote the r.h.s. of this inequality by $R_k^\star$. Clearly, $R_k^\star$ is $\overline{\mathcal{G}}_k$ – measurable function (recall that $\overline{\mathcal{G}}_k = \mathcal{G}_{k-1} \cup \sigma\{\boldsymbol{\xi}_{k,M_1}\}$). We may conclude that $\operatorname{diam}(\mathcal{T}_{k,\boldsymbol{\xi}_{M_1}}) \leq 2R_k^\star$. Furthermore, by (25) its covering number can be bounded as

$$\mathcal{N}(\mathcal{T}_{k,\boldsymbol{\xi}_{M_1}}, \rho_{\mathcal{T}}, \varepsilon) \leq \mathcal{N}(\mathsf{X} \times \mathsf{A}, \rho, \varepsilon/L_T).$$

It is also easy to check that $\left(T^x(\boldsymbol{\xi}_{k,M_2})\right), T^x \in \mathcal{T}_{k,\boldsymbol{\xi}_{M_1}}$ is a sub-Gaussian process on $(\mathcal{T}_{k,\boldsymbol{\xi}_{M_1}}, \rho_T)$ with

$$\|T^x - T^{x'}\|_{\psi_2} \lesssim \rho_{\mathcal{T}}(T^x, T^{x'}).$$

Applying the tower property we get

$$\mathsf{E}[\sup_{x\in\mathsf{X}} |T_k^x(\boldsymbol{\xi}_{k,M_2})|^2] \leq \mathsf{E}[\mathsf{E}[\sup_{x\in\mathsf{X}} |T_k^x(\boldsymbol{\xi}_{k,M_2})|^2 | \overline{\mathcal{G}}_k]]$$

Using the Dudley integral inequality, e.g. (Vershynin, 2018, Theorem 8.1.6) and assumption A5,

$$\begin{aligned}\mathsf{E}[\sup_{x\in\mathsf{X}} |T_k^x(\boldsymbol{\xi}_{k,M_1}, \boldsymbol{\xi}_{k,M_2})|^2 | \overline{\mathcal{G}}_k] &= \mathsf{E}[\sup_{T^x \in \mathcal{T}_{k,\boldsymbol{\xi}_{M_1}}} |T^x(\boldsymbol{\xi}_{k,M_2})|^2 | \overline{\mathcal{G}}_k] \\ &\lesssim \mathsf{E}[|T^{x_0}(\boldsymbol{\xi}_{k,M_2})|^2 | \overline{\mathcal{G}}_k] \\ &+ \frac{1}{M_2}\Big\{L_T\sqrt{C_{\mathsf{X},\mathsf{A}}} \int_0^{2R_k^\star/L_T} \sqrt{\log(1+1/\varepsilon)}\mathrm{d}\varepsilon + R_k^\star\Big\}^2,\end{aligned}$$

where $x_0 \in \mathsf{X}$ is some fixed point. To estimate the first term in the r.h.s. of the previous inequality we apply Hoeffding's inequality. We obtain

$$\mathsf{E}[|T^{x_0}(\boldsymbol{\xi}_{k,M_2})|^2 | \overline{\mathcal{G}}_k] \lesssim \frac{(R_k^\star)^2}{M_2}.$$

Applying Proposition A.4 we get

$$\int_0^{2R_k^\star/L_T} \sqrt{\log(1+1/\varepsilon)}\mathrm{d}\varepsilon \lesssim (R_k^\star\sqrt{\log(1+1/R_k^\star)} + R_k^\star)/L_T.$$

The last two inequalities imply

$$\mathsf{E}[\sup_{x\in\mathsf{X}} |T_k^x(\boldsymbol{\xi}_{k,M_2})|^2 | \overline{\mathcal{G}}_k] \lesssim C_{\mathsf{X},\mathsf{A}} \frac{(R_k^\star)^2 + (R_k^\star)^2 \log(1+1/R_k^\star)}{M_2}.$$

Since for $x > 0$ and $\varepsilon \in (0, 1]$

$$\log(1 + x) \le \varepsilon^{-1} x^{\varepsilon},$$

we obtain

$$\mathsf{E}[\sup_{x \in \mathsf{X}} |T_k^x|^2] \lesssim C_{\mathsf{X},\mathsf{A}} \frac{\mathsf{E}[(R_k^\star)^2] + \mathsf{E}[(R_k^\star)^{2-\varepsilon}]/\varepsilon}{M_2} .$$

Using (27), we get for any $p \ge 1$

$$\begin{aligned}\mathsf{E}^{1/p}\big[(R_k^\star)^p\big] \le \gamma \mathsf{E}^{1/p}\big[\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}}^p\big] + 2\gamma \|V^\pi - V^*\|_{\mathsf{X}} + \\ \gamma \mathsf{E}^{1/p}\left[\Big\| M_1^{-1} \sum_{l=1}^{M_1} [V^\pi(\psi(\cdot,\cdot,\xi_{k,l})) - \mathsf{E} V^\pi(\psi(\cdot,\cdot,\xi_{k,l}))] \Big\|_{\mathsf{X}\times\mathcal{A}}^p\right].\end{aligned}$$

Thus, applying (20) and Proposition A.1, for any $p \ge 1$,

$$\mathsf{E}^{1/p}\big[(R_k^\star)^p\big] \le 3\,\mathrm{C}_0\,\sigma_k\,,$$

where the quantity σ_k is defined as

$$\sigma_k = \gamma^k \big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \|V^\pi - V^*\|_{\mathsf{X}} + \frac{I_{\mathcal{D}} + \mathsf{D}}{\sqrt{M_1}} + \rho(\mathsf{X}_N, \mathsf{X})\,, \tag{28}$$

and the constant C_0 is given by

$$\mathrm{C}_0 = \max\left\{\frac{\gamma L_\pi}{1-\gamma} + \frac{R_{\max}}{(1-\gamma)^2}, \frac{(L_{\max} + 2\gamma L_\pi)\sqrt{2}}{(1-\gamma L_\psi)(1-\gamma)} + \frac{L_0}{1-\gamma}, \frac{\gamma}{1-\gamma}\right\}. \tag{29}$$

This yields the final bound

$$\mathrm{Var}[\widehat{V}_{k+1}^{\mathrm{up}}(x)] \le \mathsf{E}[\sup_{x\in\mathsf{X}} |T_k^x|^2] \le 9 C_{\mathsf{X},\mathsf{A}}\,\mathrm{C}_0^2\,\frac{\sigma_k^2 + \sigma_k^{2-\varepsilon}/\varepsilon}{M_2} \doteq \mathrm{C}\,\frac{\sigma_k^2 + \sigma_k^{2-\varepsilon}/\varepsilon}{M_2}\,. \tag{30}$$

Now the statement follows by the choice $\varepsilon = \log^{-1}(\mathrm{e} \vee \sigma_k^{-1})$.

A.4 Proof of Proposition 5.1

The corrected statement of Proposition 5.1 is given below:

Proposition A.2. *Let $|\mathsf{X}|, |\mathsf{A}| < \infty$, assume A2, A3, and $\big\|\widehat{V}_0^{\mathrm{up}}\big\|_{\mathsf{X}} \le R_{\max}(1-\gamma)^{-1}$. Then for k and M_1 large enough, it holds that*

$$\sigma_k \lesssim \|V^\pi - V^\star\|_{\mathsf{X}}\,. \tag{31}$$

The precise bounds for k and M_1 are given in (32).

Proof. Applying (28) with $I_{\mathcal{D}} \lesssim \sqrt{\log|\mathsf{X}||\mathsf{A}|}$, $\mathsf{D} = 1$, we obtain that

$$\sigma_k \lesssim \gamma^k \big\|\widehat{V}_0^{\mathrm{up}} - V^*\big\|_{\mathsf{X}} + \|V^\pi - V^*\|_{\mathsf{X}} + \frac{\gamma R_{\max}(\sqrt{\log(|\mathsf{X}||\mathsf{A}|)} + 1)}{\sqrt{M_1}(1-\gamma)^2}.$$

Note that, under assumption A3, $\big\|V^\star\big\|_{\mathsf{X}} \le R_{\max}(1-\gamma)^{-1}$. Hence, the previous bound implies $\sigma_k \lesssim \|V^\pi - V^\star\|_{\mathsf{X}}$, provided that k and M_1 are large enough to guarantee

$$\gamma^{k-1} R_{\max} \le \|V^\pi - V^\star\|_{\mathsf{X}}\,, \quad R_{\max}(\sqrt{\log(|\mathsf{X}||\mathsf{A}|)} + 1) M_1^{-1/2} (1-\gamma)^{-2} \le \|V^\pi - V^\star\|_{\mathsf{X}}\,.$$

Thus, it is enough to choose

$$\begin{aligned} &k \ge \log \|V^\pi - V^\star\|_{\mathsf{X}} (\log(1/\gamma))^{-1}\,, \\ &M_1 \ge R_{\max}^2 (\sqrt{\log(|\mathsf{X}||\mathsf{A}|)} + 1)^2 ((1-\gamma)^2 \|V^\pi - V^\star\|_{\mathsf{X}})^{-2}\,. \end{aligned} \tag{32}$$

□

A.5 THE COVERING RADIUS OF RANDOMLY DISTRIBUTED POINTS OVER A CUBE

The following proposition is a particular case of the result (Reznikov & Saff, 2016, Theorem 2.1). We repeat the arguments from that paper and give explicit expressions for the constants.

Proposition A.3. *Let $\mathsf{X} = [0,1]^{d_\mathsf{X}}$ and μ be a uniform distribution on X. Suppose that $\mathsf{X}_N = \{X_1, \dots, X_N\}$ is a set of N points independently distributed over X w.r.t. μ. Denote by $\rho(\mathsf{X}_N, \mathsf{X}) \doteq \max_{x\in\mathsf{X}} \min_{k\in[N]} |x - X_k|$ the covering radius of the set X_N w.r.t. X. Then for any $p \geq 1$,*

$$\mathsf{E}\left[\rho^p(\mathsf{X}_N, \mathsf{X})\right]^{1/p} \lesssim \sqrt{d_\mathsf{X}} \left(\frac{p \log N}{N}\right)^{1/d_\mathsf{X}}. \tag{33}$$

Moreover, for any $\delta \in (0,1)$

$$\rho(\mathsf{X}_N, \mathsf{X}) \lesssim \sqrt{d_\mathsf{X}} \left(\frac{\log(1/\delta) \log N}{N}\right)^{1/d_\mathsf{X}} \tag{34}$$

holds with probability at least $1 - \delta$.

Proof. Let $\mathcal{E}_n = \mathcal{E}_n(\mathsf{X})$ be a maximal set of points such that for any $y, z \in \mathcal{E}_n$ we have $|y - z| \geq 1/n$. Then for any $x \in \mathsf{X}$ there exists a point $y \in \mathcal{E}_n$ such that $|x - y| \leq 1/n$. Denote by $B(x, r)$ a ball centred at $x \in \mathsf{X}$ of radius r (w.r.t. $|\cdot|$) and

$$\Phi(r) = \frac{r^{d_\mathsf{X}} \pi^{d_\mathsf{X}/2}}{2^{d_\mathsf{X}} \Gamma(d_\mathsf{X}/2 + 1)},\ r \in [0, \infty)$$

Since for any $x \in \mathsf{X}$, $\mu(B(x, r)) \geq \Phi(r)$,

$$1 = \mu(\mathsf{X}) \geq \sum_{x\in\mathcal{E}_n} \mu(B(x, (1/(3n)))) \geq |\mathcal{E}_n| \Phi(1/(3n)).$$

Hence,

$$|\mathcal{E}_n| \leq \{\Phi(1/(3n))\}^{-1}. \tag{35}$$

Suppose that $\rho(\mathsf{X}_N, \mathsf{X}) > 2/n$. Then there exists a point $y \in \mathsf{X}$ such that $\mathsf{X}_N \cap B(y, 2/n) = \emptyset$. Choose a point $x \in \mathcal{E}_n$ such that $|x - y| < 1/n$. Then $B(x, 1/n) \subset B(y, 2/n)$, and so the ball $B(x, 1/n)$ doesn't intersect X_N. Hence, $\mathsf{X}_N \cap B(x, 1/(3n)) = \emptyset$. Therefore,

$$\begin{aligned} \mathrm{P}(\rho(\mathsf{X}_N, \mathsf{X}) > 2/n) \leq \mathrm{P}(\exists x \in \mathcal{E}_n : \mathsf{X}_N \cap B(x, 1/(3n)) = \emptyset) \leq \\ \leq |\mathcal{E}_n|(1 - \Phi(1/(3n))^N \leq |\mathcal{E}_n| \mathrm{e}^{-N\Phi(1/(3n))}. \end{aligned} \tag{36}$$

Let $1/(3n) = \Phi^{-1}(\alpha \log N/N)$ for some $\alpha > 0$ to be chosen later. Then $\Phi(1/(3n)) = \alpha \log N/N$. Inequalities (35) and (36) imply

$$\mathrm{P}(\rho(\mathsf{X}_N, \mathsf{X}) > 2/n) \leq \frac{N^{1-\alpha}}{\alpha \log N}. \tag{37}$$

Let us fix any $p \geq 1$. Then

$$\mathsf{E}\left[\rho^p(\mathsf{X}_N, \mathsf{X})\right]^{1/p} \leq \frac{2}{n} + \sqrt{d_\mathsf{X}} \left(\frac{N^{1-\alpha}}{\alpha \log N}\right)^{1/p} = 6\Phi^{-1}(\alpha \log N/N) + \sqrt{d_\mathsf{X}} \left(\frac{N^{1-\alpha}}{\alpha \log N}\right)^{1/p}. \tag{38}$$

Since

$$\Phi^{-1}(r) = \frac{2}{\sqrt{\pi}} \Gamma^{1/d_\mathsf{X}}(d_\mathsf{X}/2 + 1) r^{1/d_\mathsf{X}} \leq 2\sqrt{\mathrm{e}d_\mathsf{X}/\pi}(\mathrm{e}r)^{1/d_\mathsf{X}}$$

we get

$$\mathsf{E}\left[\rho^p(\mathsf{X}_N, \mathsf{X})\right]^{1/p} \leq 12\sqrt{\mathrm{e}d_\mathsf{X}/\pi} \left(\frac{\alpha \mathrm{e} \log N}{N}\right)^{1/d_\mathsf{X}} + \sqrt{d_\mathsf{X}} \left(\frac{N^{1-\alpha}}{\alpha \log N}\right)^{1/p}.$$

It remains to take $\alpha = 1 + p/d_\mathsf{X}$ to obtain bound

$$\mathsf{E}\left[\rho^p(\mathsf{X}_N, \mathsf{X})\right]^{1/p} \leq 48\sqrt{d_\mathsf{X}} \left(\frac{p \log N}{N}\right)^{1/d_\mathsf{X}}.$$

Hence, (33) follows. To prove (34) it remains to apply Markov's inequality. □

A.6 AUXILIARY RESULTS

Proposition A.4. *For any* $\Delta > 0$*,*

$$\int_0^{\Delta} \sqrt{\log(1+1/x)}\mathrm{d}x \lesssim \Delta\sqrt{\log(1+1/\Delta)} + \Delta.$$

Proof. Consider the case $\Delta < 1$. In this case

$$\begin{aligned}
\int_0^{\Delta} \sqrt{\log(1+1/x)}\mathrm{d}x &= \int_0^{\Delta^{100}/2} \sqrt{\log(1+1/x)}\mathrm{d}x + \int_{\Delta^{100}/2}^{\Delta} \sqrt{\log(1+1/x)}\mathrm{d}x \\
&\lesssim \int_0^{\Delta^{100}/2} x^{-1/2}\mathrm{d}x + \int_{\Delta^{100}/2}^{\Delta} \sqrt{\log(1+1/x)}\mathrm{d}x \\
&\lesssim \Delta^{50} + (\Delta - \Delta^{100/2})\sqrt{\log(1+2/\Delta^{100})} \\
&\lesssim \Delta\sqrt{\log(1+1/\Delta)}.
\end{aligned}$$

If $\Delta > 1$,

$$\begin{aligned}
\int_0^{\Delta} \sqrt{\log(1+1/x)}\mathrm{d}x &= \int_0^{1} \sqrt{\log(1+1/x)}\mathrm{d}x + \int_1^{\Delta} \sqrt{\log(1+1/x)}\mathrm{d}x \\
&\lesssim 2 + (\Delta - 1)\sqrt{\log 2} \lesssim \Delta.
\end{aligned}$$

□

B EXPERIMENT SETUP

B.1 ENVIRONMENTS DESCRIPTION

Garnet Garnet example is an MDP with randomly generated transition probability kernel P^a with finite X - state space and A - action space. This example is described with a tuple $\langle N_S, N_A, N_B\rangle$. The first two parameters specify the number of states and actions respectively. The last parameter is responsible for the number of states an agent can go to from state $x \in \mathsf{X}$ by performing action $a \in \mathsf{A}$. In our case we used $N_S = 20, N_A = 5, N_B = 2, \gamma = 0.9$. The reward matrix $r^a(x)$ is set according to the following principle: first, for all state-action pairs, the reward is set uniformly distributed on [0, 1]. Then the pairs are randomly chosen, for which the reward will be increased several times.

Frozen Lake The agent moves in a grid world, where some squares of the lake are walkable, but others lead to the agent falling into the water, so the game restarts. Additionally, the ice is slippery, so the movement direction of the agent is uncertain and only partially depends on the chosen direction. The agent receives 10 points only for finding a path to a goal square, for falling into the hole it doesn't receive anything. We used the built-in 4×4 map and 4 actions for the agent to perform on each state if available (`right, left, up` and `down`). For this experiment we assume that the reward matrix $r^a(x)$ is known, γ-factor is set to be 0.9.

Chain Chain is a finite MDP where agent can move only to 2 adjacent states performing 2 actions from each state (`right` and `left`). Every chain has two terminal states at the ends. For transition to the terminal states agent receives 10 points and episode ends, otherwise the reward is equal to +1. Also, there is p% noise in the system, that is the agent performs a uniformly-random action with probability p. For experiments with chains we set γ-factor to 0.8, to be sure that Pichard iterations converge.

CartPole CartPole is an example of the environment with a finite action space and infinitely large state space. In this environment agent can push cart with pole on it to the left or right direction and the target is to hold the pole up as long as possible. Reward equals to 1 is gain every time until failing or the end of episode. In fact, CartPole hasn't any specific stochastic dynamic, because transitions are deterministic according to actions, so for non degenerate case we should add some noise and we apply normally distributed random variable to the angle. LD(Linear Deterministic) policy can be expressed as $\mathrm{I}\{3 \cdot \theta + \dot{\theta} > 0\}$, where θ is an angle between pole and normal to cart.

Acrobot The environment consists of two joints, or two links. The torque is applied to the binding between the joints. The action space is discrete and consists of three kinds of torques: `left`, `right` and `none`. The state space is 6 dimensional, representing two angles (sine and cosine) characterizing the links position and the angles' velocities. Each episode starts with the small perturbations of the parameters near the resting state having both of the joints in a downward position. The agents' goal is to reach a given boundary from above in a minimum amount of time with its lower point of the link. At each timestep a robot has a reward equal to -1, and it gets 0 in a terminal state, when the boundary has been reached. Also to make the environment stochastic, a random uniform torque from -1 to 1 is added to the force at each step.

B.2 Experimental setup

For the sake of completeness, we provide below hyperparameters for the experiments run in Section 6.

Table 1: Experimental hyperparameters

Environment	M_1	M_2	discount γ	N
Garnet	3000	3000	0.9	–
Frozen Lake	1000	1000	0.9	–
Chain	1000	1000	0.8	–
Cart Pole	150	150	0.9	1500
Acrobot	150	100	0.9	4000

B.3 Auxiliary algorithms

For the sake of completeness we provide here the Value iteration algorithm (Szepesvári (2010), Chapter 1), used in Section 6 for tabular environments.

Algorithm 3: Value Iteration algorithm

Input: P^a, r, ε
Result: $V^\star$
Initialize: $\forall x \in \mathsf{X},\ V_0^\star(x) = 0$
$k = 1$
while $\left\|V_k^\star - V_{k-1}^\star\right\| > \varepsilon$ **do**
  for $x \in \mathsf{X},\ a \in \mathsf{A}$ **do**
    $Q(x,a) = r^a(x) + \gamma \sum_{y \in \mathsf{X}} \mathrm{P}^a(y|x) V_{k-1}^\star(y)$
    $V_k^\star(x) = \max_{a \in \mathsf{A}} Q(x,a)$
  end
  k = k + 1
end