PERFORMANCE BOUNDS IN L_P-NORM FOR APPROXIMATE VALUE ITERATION

RÉMI MUNOS*

Abstract.
Approximate Value Iteration (AVI) is a method for solving large Markov Decision Problems by approximating the optimal value function with a sequence of value function representations V_n processed according to the iterations $V_{n+1} = \mathcal{A}\mathcal{T}V_n$ where $\mathcal{T}$ is the so-called *Bellman operator* and $\mathcal{A}$ an *approximation operator*, which may be implemented by a *Supervised Learning* (SL) algorithm.

Usual bounds on the asymptotic performance of AVI are established in terms of the L_∞-norm approximation errors induced by the SL algorithm. However, most widely used SL algorithms (such as least squares regression) return a function (the best fit) that minimizes an empirical approximation error in L_p-norm ($p \geq 1$).

In this paper, we extend the performance bounds of AVI to weighted L_p-norms, which enables to directly relate the performance of AVI to the approximation power of the SL algorithm, hence assuring the tightness and pratical relevance of these bounds. The main result is a performance bound of the resulting policies expressed in terms of the L_p-norm errors introduced by the successive approximations. The new bound takes into account a concentration coefficient that estimates how much the discounted future-state distributions starting from a probability measure used to assess the performance of AVI can possibly differ from the distribution used in the regression operation.

We illustrate the tightness of the bounds on an optimal replacement problem.

Key words. Markov Decision Processes, Dynamic programming, Optimal control, Function approximation, Error analysis, Iterative learning control, Reinforcement learning

AMS subject classifications. 49L20, 90C40, 90C59, 93E20.

1. Introduction. We consider the problem of solving large state-space *Markov Decision Processes* (MDPs) [29] in an infinite time horizon, discounted reward setting.

The *Value Iteration* algorithm is a method for computing the optimal value function V^* by processing a sequence of value function representations V_n according to the iterations $V_{n+1} = \mathcal{T}V_n$, where $\mathcal{T}$ is the so-called *Bellman operator*. Due to a contraction property -in $L_\infty-$norm- of the Bellman operator, the iterates V_n converge to V^* as $n \to \infty$. However, this method is intractable when the number of states is so large that an exact representation of the values is impossible. We therefore need to represent the functions with a moderate number of coefficients and use methods for finding an approximate solution.

A very popular algorithm is the **Approximate Value Iteration** (AVI) algorithm. It has long been implemented in many different settings in Dynamic Programming (DP) [32, 5] with online variants in the field of Reinforcement Learning (RL) [7, 33]. It is defined by a sequence of value function representations V_n that are processed recursively by means of the iterations

$$V_{n+1} = \mathcal{A}\mathcal{T}V_n, \tag{1.1}$$

where $\mathcal{T}$ is the *Bellman operator* and $\mathcal{A}$ an *approximation operator*, which may be sampled-based implemented by a *Supervised Learning* (SL) algorithm (see e.g. [15]).

Since we will make use of different norms, let us remind now their definition: Let $u \in I\!R^N$. Its supremum (L_∞) norm is defined by $||u||_\infty := \sup_{1\leq x\leq N} |u(x)|$. Now, for μ being a probability measure on $\{1, \dots, N\}$, the weighted L_p-(semi) norm (for

*Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau, France (remi.munos@polytechnique.fr).

$p \geq 1$) -denoted by $L_{p,\mu}$- of u is $||u||_{p,\mu} := \left[\sum_{1\leq x\leq N} \mu(x)|u(x)|^p\right]^{1/p}$. In addition, we write $||\cdot||_p$ the unweighted L_p-norm (i.e. when μ is uniform).

At typical implementation of AVI is *Fitted Value Iteration* which, given a function space $\mathcal{F}$, computes at each iteration a new value representation $V_{n+1} \in \mathcal{F}$ by projecting onto $\mathcal{F}$ the Bellman image of the current estimate V_n. For illustration, a sampling-based version of this algorithm could be defined as follows: At stage n, we draw a set of independent states $\{x_k \sim \mu\}_{1\leq k\leq K}$, where μ is some probability measure on the state space, compute the Bellman values $\{v_k := \mathcal{T}V_n(x_k)\}_{1\leq k\leq K}$ for the current approximation V_n at those states, then we make a call to a SL algorithm with the data $\{(x_k, v_k)\}_{1\leq k\leq K}$ (the $\{x_k\}$ being the input and $\{v_k\}$ the desired output). The SL algorithm would return a function V_{n+1} (the best fit) that minimizes some empirical loss

$$V_{n+1} := \arg\min_{g\in\mathcal{F}} \frac{1}{K} \sum_{1\leq k\leq K} l(g(x_k) - v_k),$$

where the *loss function* l is usually a square or an absolute function (or variants, such as the ϵ-insensitive loss used in Support Vectors [36]).

This is a sampled-based version of the minimization problem in a weighted (by μ) absolute or quadratic norm ($L_{p,\mu}$-norm with $p = 1$ or 2 respectively)

$$\arg\min_{g\in\mathcal{F}} ||g - \mathcal{T}V_n||_{p,\mu}.$$

The field of *Statistical Learning* analyses the difference between the minimized empirical loss $\frac{1}{K}\sum_{1\leq k\leq K} l(V_{n+1}(x_k) - v_k)$ and the corresponding $L_{p,\mu}$-norm approximation error $||V_{n+1} - \mathcal{T}V_n||_{p,\mu}$ in terms of the number of samples K and a capacity measure of the function space $\mathcal{F}$ (such as the *covering number* or the *Vapnik-Chervonenkis (VC) dimension* [28, 36] of $\mathcal{F}$).

It is therefore natural to search for bounds on the performance of AVI that rely on weighted L_p- norms ($p \geq 1$) of the approximation errors $||V_{n+1} - \mathcal{T}V_n||_{p,\mu}$.

Unfortunately, the main field of investigation so far in Approximate DP makes use of the supremum norm [4, 5, 6, 29, 7, 16, 13]. For example, the asymptotic performance of the policies deduced by the AVI algorithm may be bounded in terms of the L_∞-norm of the approximation errors $||V_{n+1} - \mathcal{T}V_n||_\infty$ (see Section 2). However, this bound is not very useful since this uniform approximation error is difficult to control in general and is not very practical because most currently known SL algorithms solve an empirical minimization problem in L_p-norm (like least squares regression, neural networks, Support Vector and Kernel regression). Since most approximation operators provides good approximations in L_p-norm but a poor performance with respect to the L_∞-norm, it would be relevant to measure the algorithm performance with respect to the former norm.

The purpose of this paper is to extend error bounds for AVI to L_p-norms. The performance of AVI can therefore be directly related to the approximation power of the SL algorithm.

To begin with, let us mention that of course, norms are equivalent (in the case of finite dimensional spaces) since $||\cdot||_p \leq ||\cdot||_\infty \leq N^{1/p}||\cdot||_p$ (with $p \geq 1$ and N being the number of states), thus the usual L_∞ bound for AVI (detailed in Section 2) may also be used to derive an L_p norm bound. However, because of the $N^{1/p}$ factor, this yields a very loose bound for large scale problems.

The bounds derived here (see Theorem 5.2 in Section 5) depend on a new concentration (or stability) measure of the MDP: The *concentration coefficient* $C(\nu,\mu)$ measures how much the discounted average future-state distribution starting from some distribution ν used to assess the performance of AVI (through the weighting of the L_p−norm of the algorithm's performance) can possibly diverge from the distribution μ used in the regression step (by the SL algorithm). This concentration coefficient is defined as an upper-bound, taken for any non-stationary policy, of the derivative of the discounted future-state distribution (starting from ν and following a policy) with respect to (w.r.t.) the regression distribution μ.

This coefficient is related to the so-called *top-Lyapunov exponent*, which is commonly used to analyse the stability of stochastic processes. Further discussion about this concept in continuous spaces (where this coefficient is defined in terms of the Radon-Nykodim derivative of the related probability measures) can be found in [27].

A sufficient condition for the concentration coefficient to be small is when the MDP is "smooth" (i.e. when the transition probabilities are strongly stochastic, e.g. close to uniform distribution). Actually, we derive another bound, this time on the L_∞ performance of the AVI algorithm (but still in terms of the L_p approximation errors) using another concentration coefficient $C(\mu)$ that relates the immediate transition probabilities of the MDP to the regression distribution μ. For a uniform μ, a smooth MDP will define a small $C(\mu)$ value, and our bound will be sharp. However, for a MDP with deterministic transitions, the coefficient $C(\mu)$ could heavilly depend on the number of states N, making our new bounds no more informative than a usual L_∞−norm bound. This is illustrated in the *chain walk* MDP (for which $C(\mu) = N$) described in Subsection 5.5. However, even for deterministic MDPs, the concentration coefficient $C(\nu,\mu)$ may be small, and independent of N, as illustrated in the same example. For such cases, the new L_p bound is arbitrarily better than the usual L_∞ one.

The main intuition underlying this extension of usual L_∞ bounds to L_p-norms is actually simple (see the first paragraph of Section 5) and is a consequence of the componentwise bounds obtained in Section 4.

To the best of our knowledge, this weighted L_p-norm analysis of AVI is new. Previous L_p analyses in *Approximate Dynamic Programming* (ADP) include *Temporal Difference learning* (for the evaluation of a fixed policy) with linear approximation [35] and *Approximate Policy Iteration* [26] (and [1] in the continuous space, sampled-based case). Let us mention that there is an important body of literature in the domain of weighted L_∞-norm analysis of ADP [7, 17], especially for the linear programming approach [10]. Let us also remark that there exists an important related field concerned with stability, ergodicity and convergence properties of future state distributions w.r.t. the invariant probability measure (in Markov chains [19] or MDPs [18, 25]). This is not the direction followed in this paper since we are interested in the discounted reward case (with a fixed discount factor) and not the average reward case.

The paper is organized as follows: In Section 2, we remind some approximation results in L_∞-norm. Section 3 is a rough survey of approximation operators and SL algorithms. The main tool used in this paper is the derivation of the componentwise bounds for AVI, detailed in Section 4. The performance bounds in L_p-norms are stated in Section 5 and the main result of this paper is given in Theorem 5.2. A subsection provides some intuition on these results in case AVI algorithm would converge, which leads to bounds expressed in terms of the L_p Bellman residual. Section 6 details practical implementations of AVI (a sampling-based method using state-action value function approximation). The case of a continuous measurable state space is treated in Section 7 and a numerical experiment on an optimal replacement problem is detailed.

Preliminaries. We now describe the framework of MDPs in the infinite-time horizon, discounted reward setting, considered here.

Let X be the state space, assumed to be finite with N states and A a finite action space. The results given in this paper extend to infinite state spaces (either countable spaces or continuous spaces, the latter case being illustrated in Section 7). Let $p(x, a, y)$ be the probability that the next state is y given that the current state is x and the action a. Let $r(x, a, y)$ be the (deterministic) reward received when a transition $(x, a) \to y$ occurs.

We call a *(Markov* or *stationary) policy* π a mapping from X to A. We write P^π the $N \times N-$matrix with elements $P^\pi(x, y) := p(x, \pi(x), y)$ and r^π the N-vector with components $r^\pi(x) := \sum_y p(x, \pi(x), y) r(x, \pi(x), y)$.

For a given policy π, the *value function* V^π (considered as a vector with N components) is defined as the expected sum of discounted rewards:

$$V^\pi(x) := \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t, x_{t+1}) | x_0 = x, a_t = \pi(x_t)\Big],$$

where $\gamma \in [0, 1)$ is the *discount factor*. It is well known that V^π is the fixed-point of the operator $\mathcal{T}^\pi : I\!R^N \to I\!R^N$ defined, for any vector $W \in I\!R^N$, by $\mathcal{T}^\pi W := r^\pi + \gamma P^\pi W$.

The *optimal value function* $V^* := \sup_\pi V^\pi$ is the fixed-point of the Bellman operator $\mathcal{T}$ defined, for any $W \in I\!R^N$, $x \in X$, by

$$\mathcal{T}W(x) = \max_{a \in A} \sum_{y \in X} p(x, a, y)[r(x, a, y) + \gamma W(y)].$$

We say that a policy π is *greedy with respect to* $W \in I\!R^N$, if for all $x \in X$,

$$\pi(x) \in \arg\max_{a \in A} \sum_{y \in X} p(x, a, y)[r(x, a, y) + \gamma W(y)].$$

The goal is to find an optimal policy π^*, which is such that for all $x \in X$, $V^{\pi^*}(x) = \max_\pi V^\pi(x)$. It is easy to see that a policy greedy w.r.t. V^* is optimal. Since $\mathcal{A}$ is finite, such an optimal policy always exits.

2. Approximation results in L_∞-norm. Consider the **AVI algorithm** defined by (1.1) and define

$$\varepsilon_n := \mathcal{T}V_n - V_{n+1} \in I\!R^N \tag{2.1}$$

the **approximation error** at stage n. In general, AVI does not converge, but nevertheless its asymptotic behavior may be analyzed. If the approximation errors are uniformly bounded $||\varepsilon_n||_\infty \le \varepsilon$, then a bound on the difference between the asymptotic performance of policies π_n greedy w.r.t. V_n and the optimal policy is (see e.g. [7]):

$$\limsup_{n \to \infty} ||V^* - V^{\pi_n}||_\infty \le \frac{2\gamma}{(1-\gamma)^2}\varepsilon. \tag{2.2}$$

Since the proof is very simple, it is reminded here.

Proof. From the triangle inequality, the γ-contraction of the Bellman operators $\mathcal{T}$ and $\mathcal{T}^{\pi_n}$, and the fact that π_n is greedy w.r.t. V_n (i.e. $\mathcal{T}^{\pi_n}V_n = \mathcal{T}V_n$), we have

$$\begin{aligned}||V^* - V^{\pi_n}||_\infty &\leq ||\mathcal{T}V^* - \mathcal{T}^{\pi_n}V_n||_\infty + ||\mathcal{T}^{\pi_n}V_n - \mathcal{T}^{\pi_n}V^{\pi_n}||_\infty \\ &\leq \gamma||V^* - V_n||_\infty + \gamma(||V_n - V^*||_\infty + ||V^* - V^{\pi_n}||_\infty),\end{aligned}$$

thus

$$||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{1-\gamma}||V^* - V_n||_\infty. \tag{2.3}$$

Moreover, $||V^* - V_{n+1}||_\infty \leq ||\mathcal{T}V^* - \mathcal{T}V_n||_\infty + ||\mathcal{T}V_n - V_{n+1}||_\infty \leq \gamma||V^* - V_n||_\infty + \varepsilon$. Now, taking the upper limit yields $\limsup_{n\to\infty} ||V^* - V_n||_\infty \leq \varepsilon/(1-\gamma)$, which combined with (2.3) yields (2.2).□

This L_∞-bound is expressed in terms of the uniform approximation error over all states, which is difficult to guarantee, especially for large state-space problems. Moreover, it is not very useful in practice since most current approximation operators and supervised learning methods perform a minimization problem in L_1 or L_2 norm (although some exceptions of L_∞ function approximation in the framework of DP exist, see e.g. [12, 14]).

3. Approximation operators and Supervised Learning algorithms. In this section we present an overview the problem of function approximation in the context of *Statistical Learning* (see e.g. [36, 15]). To illustrate, an example of a supervised learning (SL) algorithm would take as input some data $\{(x_k, v_k)\}_{1\leq k\leq K}$, where the states $\{x_k \in X\}$ are drawn according to some distribution μ on X, and the values $\{v_k \in \mathbb{R}\}$ are unbiased estimates of some (unknown) random function with mean $f(x_k)$. This SL algorithm would return a function (called the *best fit*) that minimizes (within a given class of functions $\mathcal{F}$) the empirical loss, solving:

$$\inf_{g\in\mathcal{F}} \frac{1}{K}\sum_{k=1}^{K} l(v_k - g(x_k)),$$

where the loss function l is usually an absolute or a quadratic function (or variants, such as the *ϵ-insensitive loss function* used in Support Vectors or *Huber loss function* used for robust regression [36]).

If the unknown function is deterministic (i.e. $v_k = f(x_k)$), $\mathcal{A}$ may be considered as an approximation operator that returns a compact representation $g \in \mathcal{F}$ of an unknown function f by minimizing some empirical L_p-norm ($p = 1$ or 2) based on the data. This is a sampling-based version of a minimization problem in weighted norm $L_{p,\mu}$. Statistical Learning theory establishes bounds on the error between the minimized empirical loss $\frac{1}{K}\sum_{k=1}^{K} l(f(x_k) - g(x_k))$ and the $L_{p,\mu}-$norm difference $||f - g||_{p,\mu}$ in terms of the number of samples K and the capacity (or complexity) measure of the function space $\mathcal{F}$, characterized e.g. by the *covering number* or the *Vapnik-Chervonenkis dimension* [28, 36] of $\mathcal{F}$.

The projection onto the span of a fixed family of functions (often called *features*) is called *linear approximation* and include *Splines*, *Radial Basis, Fourier* or *Wavelet decomposition.* It is often the case that a better approximation is reached when choosing the features according to f (i.e. *feature selection*). This *non-linear approximation* is particularly efficient when f has piecewise regularities (e.g. in adaptive wavelet basis [24] such functions are compactly represented with few non-zero coefficients). Greedy algorithms for selecting the best features among a given dictionary of functions include the *Matching Pursuit* and variants [9]. Approximation theory studies the approximation error in terms of the smoothness of f [11].

In Statistical Learning, supervised learning algorithms include *Neural Network*, *Locally Weighted Learning* and *Kernel Regression* [2], *Support-Vectors* and *Reproducing Kernels* [37, 36].

Hence, given the fact that we may always bound the empirical minimized error using statistical learning tools, in the sequel, we will establish our bounds using the $L_{p,\mu}-$norm of the approximation errors. An extension of these results to sampling-based AVI is described in [27] and a policy iteration algorithm with Bellman residual minimization using a single sample-path is described in [1].

4. Componentwise performance bounds. In this section, we formulate componentwise performance bounds, from which L_p bounds will be derived in the next section. The L_∞ bound previously stated (2.2) is also an immediate consequence of a componentwise bound.

4.1. Performance bound for AVI. A componentwise bound on the asymptotic performance of the policies π_n greedy w.r.t. V_n is provided now.

LEMMA 4.1. *Consider the AVI algorithm defined by (1.1) and write $\varepsilon_n = \mathcal{T}V_n - V_{n+1} \in \mathbb{R}^N$ the approximation error at stage n. Let π_n be a greedy policy w.r.t. V_n. We have*

$$\limsup_{n\to\infty} V^* - V^{\pi_n} \le \limsup_{n\to\infty} (I-\gamma P^{\pi_n})^{-1} \Big(\sum_{k=0}^{n-1} \gamma^{n-k}\big[(P^{\pi^*})^{n-k} + P^{\pi_n}P^{\pi_{n-1}}\dots P^{\pi_{k+2}}P^{\pi_{k+1}}\big]|\varepsilon_k|\Big), \tag{4.1}$$

where $|\varepsilon_k|$ denotes the vector of absolute values of ε_k.

In order to prove this lemma, we first need this preliminary result.

LEMMA 4.2. *Let A be an invertible matrix such that all the elements of its inverse are positive. Then the solutions to the inequality $Au \le b$ are also solutions to $u \le A^{-1}b$.*

Proof of Lemma 4.2. Let u be a solution to $Au \le b$. This means that there exists a vector c with positive components s.t. $Au = b - c$, thus $u = A^{-1}b - A^{-1}c$. Since all components of $A^{-1}c$ are positive, we deduce that $u \le A^{-1}b$. □

Proof of Lemma 4.1. From the definitions of $\mathcal{T}$ and $\mathcal{T}^\pi$ we have componentwise $\mathcal{T}V_k \ge \mathcal{T}^{\pi^*}V_k$ and $\mathcal{T}V^* \ge \mathcal{T}^{\pi_k}V^*$, thus

$$V^* - V_{k+1} = \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_k + \mathcal{T}^{\pi^*}V_k - \mathcal{T}V_k + \varepsilon_k \le \gamma P^{\pi^*}(V^* - V_k) + \varepsilon_k$$
$$V^* - V_{k+1} = \mathcal{T}V^* - \mathcal{T}^{\pi_k}V^* + \mathcal{T}^{\pi_k}V^* - \mathcal{T}V_k + \varepsilon_k \ge \gamma P^{\pi_k}(V^* - V_k) + \varepsilon_k,$$

where in the second line, we used the definition of π_k as a greedy policy w.r.t. V_k, i.e. $\mathcal{T}^{\pi_k}V_k = \mathcal{T}V_k$. We deduce by induction

$$V^* - V_n \le \sum_{k=0}^{n-1} \gamma^{n-k-1}(P^{\pi^*})^{n-k-1}\varepsilon_k + \gamma^n (P^{\pi^*})^n (V^* - V_0), \tag{4.2}$$

$$V^* - V_n \ge \sum_{k=0}^{n-1} \gamma^{n-k-1}(P^{\pi_{n-1}}P^{\pi_{n-2}}\dots P^{\pi_{k+1}})\varepsilon_k + \gamma^n(P^{\pi_n}P^{\pi_{n-1}}\dots P^{\pi_1})(V^* - V_0). \tag{4.3}$$

Now, using again the definition of π_n and the fact that $\mathcal{T}V_n \geq \mathcal{T}^{\pi^*}V_n$, we have:

$$\begin{aligned} V^* - V^{\pi_n} &= \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_n + \mathcal{T}^{\pi^*}V_n - \mathcal{T}V_n + \mathcal{T}V_n - \mathcal{T}^{\pi_n}V^{\pi_n} \\ &\leq \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V_n + \mathcal{T}V_n - \mathcal{T}^{\pi_n}V^{\pi_n} \\ &= \gamma P^{\pi^*}(V^* - V_n) + \gamma P^{\pi_n}(V_n - V^{\pi_n}) \\ &= \gamma P^{\pi^*}(V^* - V_n) + \gamma P^{\pi_n}(V_n - V^* + V^* - V^{\pi_n}), \end{aligned}$$

thus $(I - \gamma P^{\pi_n})(V^* - V^{\pi_n}) \leq \gamma(P^{\pi^*} - P^{\pi_n})(V^* - V_n)$. Now, since $(I - \gamma P^{\pi_n})$ is invertible and its inverse $\sum_{k\geq 0}(\gamma P^{\pi_n})^k$ has positive elements, we use Lemma 4.2 to deduce that

$$V^* - V^{\pi_n} \leq \gamma(I - \gamma P^{\pi_n})^{-1}(P^{\pi^*} - P^{\pi_n})(V^* - V_n).$$

This, combined with (4.2) and (4.3), and after taking the absolute value (note that the vector $V^* - V^{\pi_n}$ is non-negative), yields

$$\begin{aligned} V^* - V^{\pi_n} \leq\ & (I - \gamma P^{\pi_n})^{-1} \\ & \Big\{ \sum_{k=0}^{n-1} \gamma^{n-k}\big[(P^{\pi^*})^{n-k} + (P^{\pi_n}P^{\pi_{n-1}}\dots P^{\pi_{k+1}})\big]|\varepsilon_k| \\ & + \gamma^{n+1}\big[(P^{\pi^*})^{n+1} + (P^{\pi_n}P^{\pi_n}P^{\pi_{n-1}}\dots P^{\pi_1})\big]|V^* - V_0|\Big\}. \end{aligned} \tag{4.4}$$

We deduce (4.1) by taking the upper limit. □

4.2. Performance bound based on the Bellman residual. In this section, we derive a componentwise performance bound of a policy π greedy w.r.t. some function $V \in I\!R^N$ in terms of the Bellman residual of V. This result extends the L_∞-bound (see a proof in [38]):

$$||V^* - V^\pi||_\infty \leq \frac{2}{1-\gamma}||\mathcal{T}V - V||_\infty. \tag{4.5}$$

The componentwise counterpart of this bound is stated now.

LEMMA 4.3. *Let $V \in I\!R^N$ and π a policy greedy w.r.t. V. Then*

$$V^* - V^\pi \leq \Big[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^{\pi})^{-1}\Big]|\mathcal{T}V - V|. \tag{4.6}$$

We immediately notice that (4.5) is a direct consequence of this result, since for any stochastic matrix P, $||(I-\gamma P)^{-1}||_\infty = 1/(1-\gamma)$.

Proof of Lemma 4.3. We use the fact that $\mathcal{T}V \geq \mathcal{T}^{\pi^*}V$ and the definition of π (i.e. $\mathcal{T}V = \mathcal{T}^\pi V$) to derive

$$\begin{aligned} V^* - V^\pi &= \mathcal{T}^{\pi^*}V^* - \mathcal{T}^{\pi^*}V + \mathcal{T}^{\pi^*}V - \mathcal{T}V + \mathcal{T}V - \mathcal{T}^\pi V^\pi \\ &\leq \gamma P^{\pi^*}(V^* - V^\pi + V^\pi - V) + \gamma P^\pi(V - V^\pi), \end{aligned}$$

hence $(I - \gamma P^{\pi^*})(V^* - V^\pi) \leq \gamma(P^{\pi^*} - P^\pi)(V^\pi - V)$. Again, since $(I - \gamma P^{\pi^*})$ is invertible and its inverse has positive elements, from Lemma 4.2, we deduce

$$V^* - V^\pi \leq \gamma(I - \gamma P^{\pi^*})^{-1}(P^{\pi^*} - P^\pi)(V^\pi - V).$$

Moreover,

$$\begin{aligned}(I-\gamma P^\pi)(V^\pi - V) &= V^\pi - V - \gamma P^\pi V^\pi + \gamma P^\pi V\\ &= r^\pi + \gamma P^\pi V - (r^\pi + \gamma P^\pi V^\pi) + V^\pi - V\\ &= \mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi + V^\pi - V = \mathcal{T} V - V,\end{aligned}$$

thus

$$\begin{aligned}V^* - V^\pi &\le \gamma (I-\gamma P^{\pi^*})^{-1}(P^{\pi^*} - P^\pi)(I-\gamma P^\pi)^{-1}(\mathcal{T}V - V)\\ &= (I-\gamma P^{\pi^*})^{-1}\Big[(I-\gamma P^\pi) - (I-\gamma P^{\pi^*})\Big](I-\gamma P^\pi)^{-1}(\mathcal{T}V-V)\\ &= \Big[(I-\gamma P^{\pi^*})^{-1} - (I-\gamma P^\pi)^{-1}\Big](\mathcal{T}V-V)\\ &\le \Big[(I-\gamma P^{\pi^*})^{-1} + (I-\gamma P^\pi)^{-1}\Big]|\mathcal{T}V-V|. \qquad \square\end{aligned}$$

5. Approximation results in L_p-norms. In this section, we generalize the previously mentioned L_∞ bounds to L_p-norms. The main intuition behing this extension is simple and relies on the componentwise results described in the previous section.

Indeed, assume that there exists two vectors u and v with positive components, such that, componentwise $u \le Qv$, where Q is a stochastic matrix. Of course, we may deduce that $||u||_\infty \le ||v||_\infty$, but in addition, if ν and μ are probability measures on X such that componentwise $\nu Q \le C\mu$, where $C \ge 1$ is a constant (and using usual matrix notations with the probability measures being considered as row vectors), then we deduce that

$$||u||_{p,\nu} \le C^{1/p}||v||_{p,\mu}.$$

Indeed we have

$$\begin{aligned}||u||^p_{p,\nu} = \sum_{x\in X}\nu(x)|u(x)|^p &\le \sum_{x\in X}\nu(x)\big[\sum_{y\in X}Q(x,y)v(y)\big]^p\\ &\le \sum_{x\in X}\nu(x)\sum_{y\in X}Q(x,y)v(y)^p\\ &\le C\sum_{y\in X}\mu(y)|v(y)|^p = C||v||^p_{p,\mu},\end{aligned}$$

using Jensen's inequality.

For example, if the Markov chain induced by Q has an invariant probability measure ν, then we have $||u||_{p,\nu} \le ||v||_{p,\nu}$ (i.e. the constant $C=1$). This is the main tool used in [35] to derive an L_p-norm bound for temporal difference learning with linear function approximation, where one policy only is considered.

Now, in an MDP, there are several policies, thus several stochastic matrices to be considered in order to relate $||u||_{p,\nu}$ to $||v||_{p,\mu}$. The next subsection defines the *concentration coefficients* $C_1(\nu,\mu)$, $C_2(\nu,\mu)$, and $C(\mu)$ that generalize the constant C used here to the case when several policies are considered.

A simple case for which the above idea may apply is the case of Bellman residual bounds: Choose $u = V^* - V^\pi$ and $v = \frac{2}{1-\gamma}|\mathcal{T}V - V|$, and notice that the L_∞ bound (4.5) is a consequence of (4.6). The above idea will yield an L_p-norm performance bound (this will be done in Subsection 5.3).

This same idea also holds for deriving performance bounds for AVI. We notice that the L_∞ bound (2.2) may be deduced from the componentwise bounds (4.1) and extension to L_p-norms is possible with an adequate constant, to be defined now.

5.1. Definition of the concentration coefficients. We now define the concentration coefficients $C(\mu)$, $C_1(\nu,\mu)$, and $C_2(\nu,\mu)$, that depend on the MDP, under which the distributions ν and μ may be related. Let ν and μ be two probability measures on X.

DEFINITION 5.1. *We call $C(\mu) \in I\!R^+ \cup \{+\infty\}$ the **transition probabilities concentration coefficient**, defined by*

$$C(\mu) = \max_{x,y\in X,\, a\in A} \frac{p(x,a,y)}{\mu(y)}$$

(with the convention that $0/0 = 0$, and we set $C(\mu) = \infty$ if $\mu(y) = 0$ and $p(x,a,y) > 0$ for some x, y, a). Now, let $\pi_1, \pi_2, \dots$ denotes any sequence of policies. For all integer $m \geq 1$, we define $c(m) \in I\!R^+ \cup \{+\infty\}$ by

$$c(m) = \max_{\pi_1,\dots,\pi_m,\, y\in X} \frac{(\nu P^{\pi_1} P^{\pi_2} \dots P^{\pi_m})(y)}{\mu(y)}, \tag{5.1}$$

(with the same convention as above) and write $c(0) = 1$. Note that these constants depend on ν and μ.

*We define $C_1(\nu,\mu)$ and $C_2(\nu,\mu) \in I\!R^+ \cup \{+\infty\}$, the **first** and **second order discounted future state distribution concentration coefficients**, by*

$$C_1(\nu,\mu) := (1-\gamma) \sum_{m\geq 0} \gamma^m c(m), \tag{5.2}$$

$$C_2(\nu,\mu) := (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} c(m). \tag{5.3}$$

Note that since these coefficients will appear in our bounds we are interested in the cases of finite values, for which it is sufficient that the distribution μ be strictly positive.

The transition probability concentration coefficient $C(\mu)$ was introduced in [26] to derive performance bounds for approximate policy iteration. $C(\mu)$ provides information about the relative smoothness of the immediate transition probabilities w.r.t. μ, whereas $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ give information about the worst discounted average future state distribution when starting from ν and following any policy. Informally, the future state transition is a probability measure over the state space induced by the state visitation frequency of the Markov chain resulting from the MDP when following a policy.

The coefficients $c(m)$ measure how much the future state distributions $\nu P^{\pi_1} \dots P^{\pi_m}$ may possibly differ from the distribution μ. The definition of $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ introduces an exponential discounting (first order discounting weight of γ^m for $C_1(\nu,\mu)$, and second order discounting weight of $(m+1)\gamma^m$ for $C_2(\nu,\mu)$, where m is the horizon time). The discounting makes these coefficients small for a reasonably large class of MDPs. For any sequence of policies $\pi_1,\dots,\pi_m$, the (first and second order) discounted future state distributions starting from ν and using this sequence of policies (i.e. $\{x_i \sim p(x_{i-1}, \pi_i(x_{i-1}), \cdot)\}_{1\leq i\leq m}$) is bounded by these coefficients

($C_1(\nu,\mu)$ and $C_2(\nu,\mu)$) times μ: for all x_0, y in X,

$$(1-\gamma)\sum_{m\geq 0}\gamma^m \Pr\big(x_m = y \big| x_0\sim\nu, \pi_1,\dots,\pi_m\big) \leq C_1(\nu,\mu)\mu(y),$$

$$(1-\gamma)^2\sum_{m\geq 1} m\gamma^{m-1} \Pr\big(x_m = y \big| x_0\sim\nu, \pi_1,\dots,\pi_m\big) \leq C_2(\nu,\mu)\mu(y).$$

These coefficients are related to the so-called *top-Lyapunov exponent* Γ, which play a fundamental role in the stability analysis of stochastic processes. It turns out that the stability of a stochastic system, as related to the top-Lyapunov condition $\Gamma \leq 0$ [8], is equivalent to the finiteness of the concentration coefficients. Hence, a small value of these coefficients can be interpreted as a stability condition too. Further discussion about this concept can be found in the report [27].

5.2. L_p-norm performance bounds for AVI. The next result establishes performance bounds for AVI in terms of the $L_{p,\mu}$-norm of the approximation errors $\varepsilon_n = V_{n+1} - \mathcal{T}V_n$.

THEOREM 5.2. *Let μ and ν be two probability measures on X. Consider the AVI algorithm defined by (1.1), write π_n a policy greedy w.r.t. V_n, and $\varepsilon_n = V_{n+1} - \mathcal{T}V_n \in I\!R^N$ the approximation error. Let $\varepsilon > 0$ and assume that $\mathcal{A}$ returns $\varepsilon-$approximations V_{n+1} in $L_{p,\mu}$-norm ($p \geq 1$) of $\mathcal{T}V_n$, i.e. $||\varepsilon_n||_{p,\mu} \leq \varepsilon$, for $n \geq 0$. Then:*

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\big[C(\mu)\big]^{1/p}\varepsilon, \tag{5.4}$$

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{p,\nu} \leq \frac{2\gamma}{(1-\gamma)^2}\big[C_2(\nu,\mu)\big]^{1/p}\varepsilon. \tag{5.5}$$

Notice that the l.h.s. of the first result (5.4) evaluates the performance in terms of a L_∞-norm whereas the l.h.s. of the second result (5.5) makes use of a L_p norm (although the r.h.s. of both results is expressed in L_p norm). The first result does not depend on the distribution ν and may directly be compared to the L_∞ bound (2.2). Actually (5.4) directly implies (2.2) when $p \to \infty$ (for any strictly positive measure μ).

Proof of Theorem 5.2. First, notice that the coefficient $C(\mu)$ is always larger than $C_2(\nu,\mu)$ for any distribution ν. Indeed, for all $m \geq 1$, $c(m) \leq C(\mu)$. Thus $C_2(\nu,\mu) \leq (1-\gamma)^2\sum_{m\geq 1} m\gamma^{m-1}C(\mu) = C(\mu)$. Thus, if the bound (5.5) holds for any ν, choosing ν to be a Dirac at each state implies that (5.4) also holds. Therefore, we only need to prove (5.5). We may rewrite (4.4) as

$$V^* - V^{\pi_n} \leq \frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2}\Big[\sum_{k=0}^{n-1}\alpha_k A_k|\varepsilon_k| + \alpha_n A_n|V^* - V_0|\Big],$$

with the positive coefficients $\{\alpha_k\}_{0\leq k\leq n}$

$$\alpha_k := \frac{(1-\gamma)\gamma^{n-k-1}}{1-\gamma^{n+1}}, \text{ for } 0 \leq k < n$$

$$\text{and } \alpha_n := \frac{(1-\gamma)\gamma^n}{1-\gamma^{n+1}},$$

(we notice that the sum $\sum_{k=0}^{n} \alpha_k = 1$), and the stochastic matrices $\{A_k\}_{0\le k\le n}$:

$$A_k := \frac{1-\gamma}{2}(I-\gamma P^{\pi_n})^{-1}\big[(P^{\pi^*})^{n-k} + (P^{\pi_n}P^{\pi_{n-1}}\dots P^{\pi_{k+1}})\big], \text{ for } 0\le k<n$$
$$A_n := \frac{1-\gamma}{2}(I-\gamma P^{\pi_n})^{-1}\big[(P^{\pi^*})^{n+1} + (P^{\pi_n}P^{\pi_n}\dots P^{\pi_1})\big].$$

Since the two sides of this componentwise bound are positive, we may take the $L_{p,\nu}$ norm of those two vectors:

$$\begin{aligned}
||V^* - V^{\pi_n}||_{p,\nu}^p &\\
&\le \Big[\frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2}\Big]^p \sum_{x\in X}\nu(x)\Big[\sum_{k=0}^{n-1}\alpha_k A_k|\varepsilon_k| + \alpha_n A_n|V^*-V_0|\Big]^p(x)\\
&\le \Big[\frac{2\gamma(1-\gamma^{n+1})}{(1-\gamma)^2}\Big]^p \sum_{x\in X}\nu(x)\Big[\sum_{k=0}^{n-1}\alpha_k A_k|\varepsilon_k|^p + \alpha_n A_n|V^*-V_0|^p\Big](x), \qquad (5.6)
\end{aligned}$$

using two times Jensen's inequality (since the coefficients $\{\alpha_k\}_{0\le k\le n}$ sum to 1 and the matrix A_k are stochastic) (i.e. convexity of $x\to|x|^p$). The second term in the brackets disappears when taking the upper limit. Now, from the definition of the coefficients $c(m)$, $\nu A_k \le (1-\gamma)\sum_{m\ge 0}\gamma^m c(m+n-k)\mu$, thus the first term in (5.6) satisfies

$$\begin{aligned}
\sum_x \nu(x)\sum_{k=0}^{n-1}\alpha_k A_k|\varepsilon_k|^p(x) &\le \sum_{k=0}^{n-1}\alpha_k(1-\gamma)\sum_{m\ge 0}\gamma^m c(m+n-k)||\varepsilon_k||_{p,\mu}^p\\
&\le \frac{(1-\gamma)^2}{1-\gamma^{n+1}}\sum_{m\ge 0}\sum_{k=0}^{n-1}\gamma^{m+n-k-1}c(m+n-k)\varepsilon^p\\
&\le \frac{1}{1-\gamma^{n+1}}C_2(\nu,\mu)\varepsilon^p,
\end{aligned}$$

where we replaced α_k by their values, and used the fact that $||\varepsilon_k||_{p,\mu}\le\varepsilon$. By taking the upper limit in (5.6), we deduce (5.5). □

What if AVI converges ?. We know that there is no guarantee that AVI converges. However, experimentally, we observe that in some cases convergence occurs. It is interesting to notice that in such cases, better bounds may be derived (in any norm) whenever $\gamma > 1/2$. Indeed, convergence of AVI would mean that there exists $V \in I\!R^N$ such that $\lim_{n\to\infty} V_n = V$. Thus, by taking the limit in (1.1), we deduce that V is a fixed-point of the operator $\mathcal{AT}$, i.e. $V = \mathcal{AT}V$, and the approximation error (2.1) tends to the residual $\mathcal{T}V - V$ of V.

We deduce that the asymptotic performance of AVI is the performance of a policy π greedy w.r.t. V, thus may be expressed in terms of the residual $\mathcal{T}V-V$. Hence, the bounds based on the Bellman residual (the L_∞-norm bound (4.5) or the componentwise bound (4.6)), which yields a coefficient $2/(1-\gamma)$ instead of $2\gamma/(1-\gamma)^2$ (for AVI bounds), provides a better bound whenever $\gamma > 1/2$. The next subsection provides an extension of Bellman residual bounds to L_p-norms.

5.3. L_p-norm bounds based on the Bellman residual. Here, we relate the performance of a policy π greedy w.r.t. V (where $V\in I\!R^N$) in terms of the $L_{p,\mu}$-norm of its residual $\mathcal{T}V - V$.

THEOREM 5.3. *Let V be a vector of size N and π a policy greedy w.r.t. V. Let μ and ν be two probability measures on X. Then*

$$||V^* - V^\pi||_\infty \leq \frac{2}{(1-\gamma)}\big[C(\mu)\big]^{1/p}||\mathcal{T}V - V||_{p,\mu}, \tag{5.7}$$

$$||V^* - V^\pi||_{p,\nu} \leq \frac{2}{(1-\gamma)}\big[C_1(\nu,\mu)\big]^{1/p}||\mathcal{T}V - V||_{p,\mu}. \tag{5.8}$$

Here also, the first result (5.7) provides a L_∞-norm bound on the performance, which may directly be compared to the L_∞ bound (4.5) (letting $p \to \infty$) whereas a L_p norm performance bound is stated in the second result (5.8).

Proof of Theorem 5.3. We may rewrite (4.6) as

$$V^* - V^\pi \leq \frac{2}{1-\gamma}A|\mathcal{T}V - V|,$$

where A is the stochastic matrix

$$A = \frac{1-\gamma}{2}\Big[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^\pi)^{-1}\Big].$$

Using the idea described in the introduction of this section, we have

$$\begin{aligned}||V^* - V^\pi||_{p,\nu}^p &\leq \Big[\frac{2}{1-\gamma}\Big]^p \sum_{x\in X}\nu(x)\Big[A|\mathcal{T}V - V|\Big]^p(x) \\ &\leq \Big[\frac{2}{1-\gamma}\Big]^p \sum_{x\in X}\nu(x)\Big[A|\mathcal{T}V - V|^p\Big](x), \end{aligned} \tag{5.9}$$

from Jensen's inequality. Now, from the definition of the coefficients $c(m)$, $\nu A \leq (1-\gamma)\sum_{m\geq 0}\gamma^m c(m)\mu = C_1(\nu,\mu)\mu$, thus

$$||V^* - V^\pi||_{p,\nu}^p \leq \Big[\frac{2}{1-\gamma}\Big]^p C_1(\nu,\mu)\mu|\mathcal{T}V - V|^p = \Big[\frac{2}{1-\gamma}\Big]^p C_1(\nu,\mu)||\mathcal{T}V - V||_{p,\mu}^p,$$

which proves (5.8). Now, since $C(\mu) \geq C_1(\nu,\mu)$ for any ν, choosing ν to be a Dirac at each state yields (5.7). □

For intuition purpose, the components $A(x,y)$ of the matrix A indicates a bound on the contribution of the (absolute value of the) residual at a state y to the performance error at the state x. Indeed,

$$V^*(x) - V^\pi(x) \leq \frac{2}{1-\gamma}\sum_{y\in X}A(x,y)|\mathcal{T}V - V|(y).$$

It is clear from (5.9) that if we chose $\mu = \nu A$, then the L_p bound becomes

$$||V^* - V^\pi||_{p,\nu} \leq \frac{2}{(1-\gamma)}||\mathcal{T}V - V||_{p,\mu}. \tag{5.10}$$

This bound may inspire us for solving a direct Bellman residual minimization problem, in some given function space $\mathcal{F}$:

$$\min_{V\in\mathcal{F}} ||\mathcal{T}V - V||_{p,\mu}^p$$

where the distribution μ now depends on V, through the policy π greedy w.r.t. V, i.e. $\mu = \nu A = \frac{1-\gamma}{2}\nu\Big[(I-\gamma P^{\pi^*})^{-1} + (I-\gamma P^{\pi})^{-1}\Big]$. We write $\mu = (\mu^\pi + \mu^*)/2$ with $\mu^\pi = (1-\gamma)\nu(I-\gamma P^\pi)^{-1}$ being the discounted future state distribution starting from ν and following policy π, and $\mu^* = (1-\gamma)\nu(I-\gamma P^{\pi^*})^{-1}$, similarly defined from the optimal policy π^*.

Thus the $L_{p,\mu}$-norm of the residual to be minimized is composed of two contributions:

$$||\mathcal{T}V - V||^p_{p,\mu} = \frac{1}{2}\Big(||\mathcal{T}V - V||^p_{p,\mu^\pi} + ||\mathcal{T}V - V||^p_{p,\mu^*}\Big). \tag{5.11}$$

One may consider an iterative optimization method, such as a gradient method, where at each iteration an empirical residual would be computed and minimized. Minimization of the first term in (5.11) is easy to implement by designing a sampling device from μ^π (i.e. start from an initial state $x \sim \nu$ and follow transitions using the current policy π during a horizon time that is a exponential random variable with coefficient γ). The second term is more difficult to deal with because there is no sampling device from μ^* since π^* is unknown; one may consider a somehow uniform density instead or use a discounted future state distribution using a stochastic policy (where each action has a strict positive probability to be chosen).

5.4. Some intuition about the coefficients $C(\mu)$, $C_1(\nu,\mu)$, and $C_2(\nu,\mu)$. Let us give some more insight about these coefficients in the case of a uniform distribution $\mu = (\frac{1}{N}\dots\frac{1}{N})$. In that case, from its definition, the coefficient $C(\mu)$ is always smaller than the number of states N. $C(\mu)$ equals N if there exists at least a deterministic transition (i.e. for some $x, y \in X$, $a \in A$, we have $p(x,a,y) = 1$). In that case, the L_p (say, for $p = 1$) bound (5.4) would be not better than the L_∞ one (2.2) combined with the simple norm comparison result $||\cdot||_\infty \le N||\cdot||_1$.

Hence, the L_p bound (5.4) (resp. (5.7)) is more informative than the usual L_∞ one (2.2) (resp. (4.5)) whenever the concentration coefficient $C(\mu)$ is smaller than the number of states. An interesting case for which this happens is when the state space is continuous and the transition kernel admits a density w.r.t. μ, for which case, $C(\mu)$ is the upper bound of this density. This continuous space case will be considered in Section 7 and illustrated on an optimal replacement problem.

Now, consider the coefficients $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ when ν and μ are both uniform.

- Their largest possible value is obtained in a MDP where for a specific policy π, all states jump to a given state -say state 1- with probability 1. Thus, for any ν, for all m, $\nu(P^\pi)^m = (1\,0\,\dots\,0) \le c(m)\mu$ holds with $c(m) = N$ (with equality in state 1), and therefore $C_1(\nu,\mu) = C_2(\nu,\mu) = N$. This is the worst case because the future state distribution accumulates on a single state. In that case, the L_p bound (5.5) (resp. (5.8)) may actually be derived from the L_∞ one (2.2) (resp. (4.5)) since $||\cdot||_p \le ||\cdot||_\infty$ and $||\cdot||_\infty \le N^{1/p}||\cdot||_p$.
- Their lowest possible value is obtained in a MDP with uniform transition probabilities $p(x,a,y) = 1/N$, for all $x, y \in X$ and $a \in A$. When ν and μ are both uniform then $c(m) = 1$ and $C_1(\nu,\mu) = C_2(\nu,\mu) = 1$ (this is the lowest possible value since for a uniform ν and any stochastic matrix P, we have $\max_y \sum_x \nu(x)P(x,y) \ge 1/N$).

Notice however that any deterministic MDP would not necessarily lead to a high value of the coefficients $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$ (contrarily to the case of $C(\mu)$). Indeed, in an MDP where the policies consist in permutations of the states (for which each

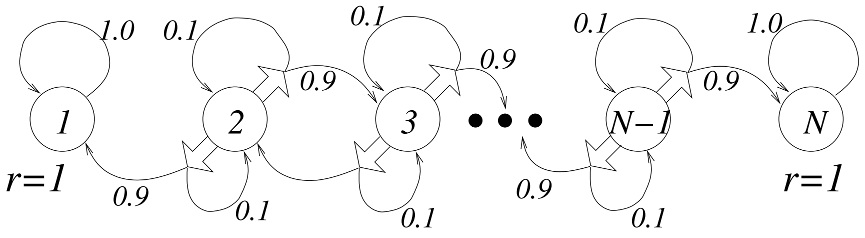

FIGURE 5.1. *The chain walk MDP.*

state has a unique successor and unique predecessor), then $C(\mu) = N$ (since the transitions are deterministic, as seen previously), but $C_1(\nu, \mu) = C_2(\nu, \mu) = 1$ for uniform distributions ν and μ (since for all $m \geq 0$, $c(m) = 1$). Another example where the discounted future state distribution concentration coefficients is low (and independent of the number of states N) is provided in the chain walk MDP described in the next subsection.

The concentration coefficients $C_1(\nu, \mu)$ and $C_2(\nu, \mu)$ express how the (first and second order) discounted future state distribution, starting from the initial distribution ν, may possibly differ from μ. A low value of these coefficients means that the mass of the discounted future state distribution starting from ν does not accumulate on few specific states for which the distribution μ is low. For the purpose of obtaining low values of these coefficients (thus probably good performance for AVI), it is desirable that μ be somehow uniformly distributed (this condition was already mentioned in [22, 20, 26] to secure the policy improvement steps in approximate policy iteration).

5.5. Illustration on the *chain walk MDP*. We illustrate the fact that the L_p-norm bound (5.5) given in Theorem 5.2 is tighter than the L_∞−norm (2.2) (combined with the norm comparison $||\cdot||_\infty \leq N^{1/p}||\cdot||_p$) on the *chain walk* MDP defined in [23] (see Figure 5.1). This case provides an example for which the coefficient $C(\mu)$ is high (its value is the number of states N) but $C_1(\nu, \mu)$ and $C_2(\nu, \mu)$ are low (independent of N).

This is a linear chain with N states with two dead-end states: states 1 and N. On each of the interior states $2 \leq x \leq N-1$ there are two possible actions: right or left, which moves the state in the intended direction with probability 0.9, and fails with probability 0.1, leaving the state unchanged. The reward simply depends on the current state and is 1 at boundary states and 0 elsewhere: $r = (1\,0\ldots0\,1)'$.

We consider an approximation of the value function in the two dimensional function space $\mathcal{F} := \{f_\alpha(x) = \alpha_1 + \alpha_2 x\}_{\alpha \in I\!R^2}$ where $x \in \{1, \ldots, N\}$ is the state index. Assume that the initial approximation is zero: $V_0 = (0\ldots0)'$. Then $\mathcal{T}V_0 = (1\,0\ldots0\,1)'$. The best fit (in L_∞-norm) of $\mathcal{T}V_0$ in $\mathcal{F}$ is the constant function $V_1 = (\frac{1}{2}\ldots\frac{1}{2})'$ which produces an error $||V_1 - \mathcal{T}V_0||_\infty = \frac{1}{2}$.

Let us choose uniform distributions $\nu = \mu = (\frac{1}{N}\ldots\frac{1}{N})$. In L_1-norm, the best fit of $\mathcal{T}V_0$ in $\mathcal{F}$ is $V_1 = (0\ldots0)'$ (for $N > 4$) and the resulting error is $||V_1 - \mathcal{T}V_0||_1 = \frac{2}{N}$. In L_2-norm the best fit is also constant $V_1 = (\frac{2}{N}\ldots\frac{2}{N})'$ and the error is $||V_1 - \mathcal{T}V_0||_2 = \frac{\sqrt{2N-4}}{N}$.

In these three cases, we observe by induction that the successive approximations V_n are constant, thus $\mathcal{T}V_n = r + \gamma V_n$ and the approximation errors remain the same as in the first iteration: for all $n \geq 0$, $||V_{n+1} - \mathcal{T}V_n||_\infty = \frac{1}{2}$, $||V_{n+1} - \mathcal{T}V_n||_1 = \frac{2}{N}$, and $||V_{n+1} - \mathcal{T}V_n||_2 = \frac{\sqrt{2N-4}}{N}$.

Since V_n is constant, any policy π_n is greedy w.r.t. V_n. Hence for $\pi_n = \pi^*$ the l.h.s. of (2.2) and (5.5) are equal to zero. Now, in order to compare the r.h.s. of these inequalities, let us calculate the coefficients $C(\mu)$ and $C_1(\nu,\mu)$ and $C_2(\nu,\mu)$. Since state 1 jumps to itself with probability 1, we have no better coefficient than $C(\mu) = N$.

Now, the maximum in (5.1) is reached when the mass of the future state distribution is mostly concentrated on one specific state -say state 1- which corresponds to a policy π_{Left} that chooses everywhere action left. We see that for $\nu = \mu$,

$$\nu(P^{\pi_{\text{Left}}})^m(x) \leq \nu(P^{\pi_{\text{Left}}})^m(1) \leq (1 + 0.9m)\mu(x),$$

for all $x \geq 0$, thus $c(m) \leq 1 + 0.9m$. We deduce that the coefficients $C_1(\nu,\mu) \leq (1-\gamma)\sum_{m\geq 0}\gamma^m(1+0.9m)$ and $C_2(\nu,\mu) \leq (1-\gamma)^2\sum_{m\geq 1} m\gamma^{m-1}(1+0.9m)$ are upper bounded by a value that is **independent of the number of states** N.

Thus, if we consider the performance of AVI in L_1-norm, the bound (5.5) (for $p = 1$) provides an approximation of order $O(N^{-1})$, whereas the L_1 bound that would be obtained from the usual L_∞ result (2.2) combined with the norm comparison $||\cdot||_\infty \leq N||\cdot||_1$ would provide a $O(1)$ approximation only.

Similarly, the L_2-norm bound is of order $O(N^{-1/2})$, whereas the $L_\infty-$norm bound (2.2) combined with $||\cdot||_\infty \leq N^{1/2}||\cdot||_2$ would only be of order $O(1)$.

Thus, if our supervised learning algorithm returns the best regression function by minimizing an approximation error in L_p-norm (which is usually the case in practice), **the bound (5.5) may be arbitrarily more informative than (2.2) for large values of** N**.**

6. Practical algorithms. Practical implementations of AVI depend on the amount of knowledge available on the state dynamics as well as the way the expectation operation (in the Bellman operator) may be processed.

In the case of a **complete model** (when the state transitions $p(x,a,y)$ are perfectly known) and if the expectation operation is computationally tractable, then a possible implementation of AVI has already been described in the introduction: at each stage n, we select a set of states $\{x_k \in X\}_{1\leq k\leq K}$ drawn according to some distribution μ, compute the backed-up values $\{v_k = \mathcal{T}V_n(x_k)\}_{1\leq k\leq K}$, and make a call to a SL algorithm with the data $\{(x_k; v_k)\}_{1\leq k\leq K}$, which returns an $\varepsilon-$approximation V_{n+1} in $L_{p,\mu}$-norm, i.e. $||V_{n+1} - \mathcal{T}V_n||_{p,\mu} \leq \varepsilon$. Of course, we need additional assumptions on the number of samples K and the complexity of the function space $\mathcal{F}$ (in terms of covering number or VC dimension) to guarantee that the empirical loss $\left(\frac{1}{K}\sum_{k=1}^K |V_{n+1}(x_k) - v_k|^p\right)^{1/p}$ is close to the norm of the approximation error $||V_{n+1} - \mathcal{T}V_n||_{p,\mu}$, but such considerations are omitted here, and we direct the interested reader to [36, 15, 30].

However, it is often the case that no explicit representation of the transition probabilities $p(x,a,y)$ is available, but there exists a sampling device that allows to generate states y according to the distribution $p(x,a,\cdot)$ at any state x and action a of our choice. We call this a **generative model** (see [21] for a survey of several sampling models). One possible way to compute the expectation operation in the Bellman operator is to replace it by an empirical mean using this sampling device. This leads to *sampling based fitted value iteration*, studied in [34].

Another alternative, closer in spirit to Reinforcement Learning (RL) [33], consists in introducing the state-action value function, or Q-function, defined, for each state-action $(x, a) \in X \times A$ by

$$Q^*(x,a) := \sum_{y \in X} p(x,a,y)\big[r(x,a,y) + \gamma V^*(y)\big].$$

We have the properties that $V^*(x) = \max_{a \in A} Q^*(x,a)$, and Q^* is the fixed point of the operator $\mathcal{R}$, mapping from the space of functions $X \times A \to I\!R$ to itself, defined for any $Q : X \times A \to I\!R$ by

$$\mathcal{R}Q(x,a) := \sum_{y \in X} p(x,a,y)\big[r(x,a,y) + \gamma \max_{b \in A} Q(y,b)\big].$$

An AVI algorithm using this representation would consist in defining successive approximations Q_n (with any initial Q_0) according to the recursion

$$Q_{n+1} = \mathcal{A}\mathcal{R}Q_n, \tag{6.1}$$

where $\mathcal{A}$ is a SL algorithm on $X \times A$. A model-free RL algorithm would collect a number of transitions of the form $\{(x_k, a_k) \xrightarrow{r_k} y_k\}_{1 \le k \le K}$, where a_k is an action chosen in state x_k, the next state y_k being generated according to the generative model (i.e. $y_k \sim p(x_k, a_k, \cdot)$), and $r_k = r(x_k, a_k, y_k)$ is the received reward. We then compute the back-up values $v_k = r_k + \gamma \max_{b \in A} Q_n(y_k, b)$ (which provides an unbiased estimate of $\mathcal{R}Q_n(x_k, a_k)$), and make a call to the SL algorithm with the data $\{(x_k, a_k); v_k\}_{1 \le k \le K}$ (the inputs being the couples $\{(x_k, a_k)\}$, and the desired output $\{v_k\}$), which returns the next Q-function Q_{n+1}.

An interesting case is when $\mathcal{A}$ is a linear operator *in the values* $\{v_k\}$ such as in linear approximation, memory-based learning (k-Nearest Neighbors, Locally Weighted Learning [3, 15]) or Support Vector Regression (in the case of a quadratic loss function). In that case, the approximation $\mathcal{A}$ and expectation $\mathbb{E}$ operators commute and the approximation Q_{n+1} returned by the SL algorithm is therefore an unbiased estimate of $\mathcal{A}\mathcal{R}Q_n$. Thus when K is large, such an iteration acts like a (model-based) AVI iteration, and bounds similar to those of Theorem 5.2 may be derived.

Notice that a policy π'_n derived from the approximate Q-function: $\pi'_n(x) \in \arg\max_{a \in A} Q_n(x,a)$ is different from the policy π_n greedy w.r.t. V_n, defined by $V_n(x) = \max_a Q_n(x,a)$. Indeed, the latter satisfies $\pi_n(x) \in \arg\max_{a \in A} \mathcal{R}Q_n(x,a)$. However, bounds similar to (2.2), (5.4), and (5.5) on the performance of such policies π'_n may be derived analogously. An example of such bound in L_∞-norm is provided now. Extension to L_p bounds would follow the same lines as in Sections 4 and 5.

The performance $Q^\pi : X \times A \to I\!R$ of a policy π is defined as follows: $Q^\pi(x,a)$ is the expected sum of rewards when starting from x, choosing action a and using policy π thereafter. Q^π is also the fixed-point of the Bellman operator $\mathcal{R}^\pi$, mapping from the space of functions $X \times A \to I\!R$ to itself, defined by

$$\mathcal{R}^\pi Q(x,a) := \sum_{y \in X} p(x,a,y)\big[r(x,a,y) + \gamma Q(y, \pi(y))\big].$$

THEOREM 6.1. *Consider the AVI algorithm defined by the Q-function iteration (6.1). Let ε be a uniform bound on the L_∞ approximation errors of the Q-functions, i.e. $||Q_{n+1} - \mathcal{R}Q_n||_\infty \le \varepsilon$. The asymptotic performance of the policy π'_n (defined by $\pi'_n(x) \in \arg\max_{a \in A} Q_n(x,a)$) satisfy*

$$\limsup_{n \to \infty} ||Q^* - Q^{\pi'_n}||_\infty \le \frac{2\gamma}{(1-\gamma)^2}\varepsilon.$$

Proof of Theorem 6.1. The proof is similar to that of (2.2); it suffices to replace the V-value by the Q-values, the $\mathcal{T}$ (resp. $\mathcal{T}^\pi$) operator by the $\mathcal{R}$ (resp. Q^π) operators, and notice that $\mathcal{R}^{\pi'_n} Q_n = \mathcal{R} Q_n$.□

7. Numerical experiment in the continuous case. All previous results extend to the case of continuous measurable state spaces. We first redefine the concentration coefficients in this context and illustrate numerically the method on an optimal replacement problem, for which the coefficient $C(\mu)$ is explicitly computed.

Let us write $P(x, a, B)$ the transition probability kernel, where B is any measurable subset of X. For a stationary policy $\pi : X \to A$, we write $P^\pi(x, B) = P(x, \pi(x), B)$, which defines a right linear operator (defined on the space of bounded measurable function V with domain X): $P^\pi V(x) := \int_X V(y) P^\pi(x, dy)$, and a left-linear operator (defined on the space of probability measures μ on X): $\mu P^\pi(B) := \int_X P^\pi(x, B)\mu(dx)$. The product of two kernels P^{π_1} and P^{π_2} is defined by $P^{\pi_1} P^{\pi_2} (x, B) := \int_X P^{\pi_1}(x, dy) P^{\pi_2}(y, B)$.

7.1. Concentration coefficients. With these notations, the concentration coefficients are defined as follows: let ν and μ be two probability distributions on X.

We assume that for all $x \in X$, $a \in A$, $P(x, a, \cdot)$ is absolutely continuous w.r.t. μ and the Radon-Nikodym derivative of $P(x, a, \cdot)$ w.r.t. $\mu(\cdot)$ is bounded uniformly in x and a. Then, the transition probabilities concentration coefficient $C(\mu)$ is defined by

$$C(\mu) := \sup_{x \in X, a \in A} \frac{dP(x, a, \cdot)}{d\mu}.$$

Notice that if μ is the Lebesgue measure over X, and if $P(x, a, \cdot)$ admits a uniformly bounded density, then the concentration coefficient $C(\mu)$ is equal to the upper bound of this density. This case is illustrated in the numerical experiment below. The first and second order discounted future state distribution concentration coefficients $C_1(\nu, \mu)$ and $C_2(\nu, \mu)$ are defined similarly from (5.2) and (5.3).

7.2. An optimal replacement problem. This experiment illustrates the respective tightness of the L_∞, L_1, and L_2 norm bounds on a continuous space control problem excerpted from [31].

A one-dimensional continuous variable $x_t \in [0, x_{\max}]$ measures the accumulated utilization (such as the odometer reading on a car) of a product. $x_t = 0$ denotes a brand new product. At each discrete time t, there are two possible decisions: either keep ($a_t = \mathrm{K}$) or replace ($a_t = \mathrm{R}$), in which case an additional cost $C_{replace}$ (of selling the existing product and replacing it for a new one) occurs. The transition densities are exponential with parameter β with a truncated queue. Moreover, if the next state y is larger than the maximal value $x_{\max}$ (e.g. the car breaks down because it is too damaged) then a new state is immediately redrawn and a penalty $C_{dead} > C_{replace}$ occurs. The transition densities are thus defined as follows: defining $q(x) := \beta e^{-\beta x}/(1 - e^{-\beta x_{\max}})$,

$$p(x, a = \mathrm{R}, y) = \begin{cases} q(y) & \text{if } y \in [0, x_{\max}] \\ 0 & \text{otherwise.} \end{cases}$$

$$p(x, a = \mathrm{K}, y) = \begin{cases} q(y - x) & \text{if } y \in [x, x_{\max}] \\ q(y - x + x_{\max}) & \text{if } y \in [0, x) \\ 0 & \text{otherwise.} \end{cases}$$

The current cost (opposite of a reward) $c(x)$ is the sum of a slowly increasing function (maintenance cost) and a discontinuous punctual cost (e.g. which may represent car insurance fees).

The current cost function and the optimal value function (computed by a discretization on a high resolution grid) are shown on Figure 7.1.

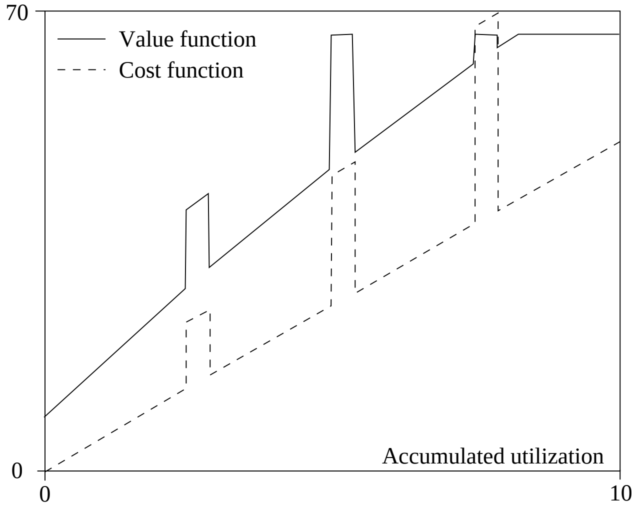

FIGURE 7.1. *Cost and value functions.*

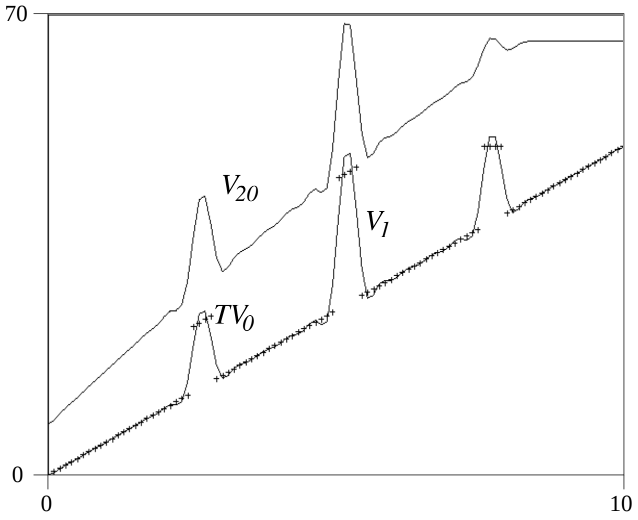

FIGURE 7.2. $\mathcal{T}V_0$ *(crosses),* V_1 *and* V_{20}.

We choose the numerical values $\gamma = 0.6$, $\beta = 0.6$, $C_{replace} = 50$, $C_{dead} = 70$, and $x_{\max} = 10$. We consider a uniform distribution μ on the domain $[0, x_{\max}]$. We choose K points (with $K = 200$ or 2000 points) uniformly located over the domain $\{x_k := kx_{\max}/K\}_{0\le k<K}$ to perform the L_2 minimization fitting problem at each

iteration:

$$V_{n+1} = \arg\min_{f\in\mathcal{F}} \frac{1}{K}\sum_{k=1}^{K}[f(x_k) - \mathcal{T}V_n(x_k)]^2,$$

where $\mathcal{F}$ is the space spanned by a truncated cosine basis (with $M = 20$ or $M = 40$ basis functions):

$$\mathcal{F} := \Big\{f(x) = \sum_{m=1}^{M} \alpha_m \cos(m\pi \frac{x}{x_{\max}})\Big\}_{\alpha\in I\!R^M}.$$

We start with initial values $V_0 = 0$. In Figure 7.2 we show the first iteration (for the grid with $K = 200$ points): the backed-up values $\mathcal{T}V_0$ (indicated with crosses), the corresponding approximation V_1 (best fit of $\mathcal{T}V_0$ in the cosine approximation space $\mathcal{F}$). The approximate value function computed after 20 iterations (when there are no significant improvement of the approximations) is also plotted.

The concentration coefficient $C(\mu)$ is the highest peak of the transition density with respect to the uniform distribution μ, thus $C(\mu) = q(0)x_{\max} = \beta x_{\max}/(1 - e^{-\beta x_{\max}}) \simeq 6$.

	$\lvert\lvert\varepsilon_n\rvert\rvert_\infty$	$C(\mu)\lvert\lvert\varepsilon_n\rvert\rvert_1$	$\sqrt{C(\mu)}\lvert\lvert\varepsilon_n\rvert\rvert_2$
$K = 200$, $M = 20$	12.4	0.367	1.16
$N = 2000$, $M = 40$	12.4	0.0552	0.897

Table 7.1
Comparison of the r.h.s. of the L_∞, L_1 and L_2 bounds.

Table 1 compares the right hand side (up to the constant $2\gamma/(1-\gamma)^2$) of equations (2.2) and (5.4) for $p = 1$ and 2, their left hand side being the same since they use the same L_∞-norm. We notice that the L_1 and L_2 bounds (5.4) are much tighter than the L_∞ one (2.2). Moreover we observe that the L_1 and L_2 approximation errors tend to 0 when the number K of sampling points and the number M of basis functions go to infinity, whereas the L_∞ bound does not. Indeed, since the cost function is discontinuous, the L_∞ approximation error (using continuous function approximation such as the cosine basis used here) will never be smaller than half the value of the largest jump, even for large values of K and M. This example illustrates the fact that the L_p bound (5.4) may be arbitrarily tighter than the L_∞ one (2.2).

8. Conclusion. Theorem 5.2 provides a useful tool to bound the performance of AVI from the L_p-norm of the approximation errors, thus in terms of the approximation power of most SL algorithms. Expressing the performance of AVI in the same norm as the norm used by the supervised learner to solve the regression problem guarantees the tightness and practical application of the bounds.

In order that these bounds be of any use, we need to estimate an upper bound on the concentration coefficients $C(\mu)$, $C_1(\nu,\mu)$, and $C_2(\nu,\mu)$, which may be difficult in general. We illustrate the case of low values of $C_1(\nu,\mu)$, and $C_2(\nu,\mu)$ in the chain walk MDP, and the case of a low value of $C(\mu)$ in the optimal replacement problem. Future work would consider defining classes of problems for which these coefficients may be evaluated.

Extension to other loss functions l, such as ϵ-*insensitive* (used in Support Vectors) or *Huber loss function* (for robust regression) [36] is straightforward (as long as l is an increasing and convex function over $\mathbb{R}^+$). Another possible extension is AVI for Markov games.

Acknowledgements. The author wishes to thank Csaba Szepesvári and the anonymous reviewers who helped improving in a significant manner the clarity of the concepts introduced in the paper.

REFERENCES

[1] A. Antos, C. Szepesvari, and R. Munos, *Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path*, Conference on Learning Theory, (2006).
[2] C. G. Atkeson, A. W. Moore, and S. A. Schaal, *Locally weighted learning*, AI Review, 11 (1997).
[3] ———, *Locally weighted learning for control*, AI Review, 11 (1997).
[4] R. Bellman, *Dynamic Programming*, Princeton Univ. Press, 1957.
[5] R. Bellman and S. Dreyfus, *Functional approximation and dynamic programming*, Math. Tables and other Aids Comp., 13 (1959), pp. 247–251.
[6] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice Hall, 1987.
[7] D. P. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.
[8] P. Bougerol and N. Picard, *Strict stationarity of generalized autoregressive processes*, Annals of Probability, 20 (1992), pp. 1714–1730.
[9] G. Davies, S. Mallat, and M. Avellaneda, *Adaptive greedy approximations*, J. of Constr. Approx., 13 (1997), pp. 57–98.
[10] D. de Farias and B. V. Roy, *The linear programming approach to approximate dynamic programming*, Operations Research, 51 (2003).
[11] R. DeVore, *Nonlinear Approximation*, Acta Numerica, 1997.
[12] G. Gordon, *Stable function approximation in dynamic programming*, Proceedings of the International Conference on Machine Learning, (1995).
[13] G. J. Gordon, *Approximate solutions to Markov Decision Processes*, PhD thesis, CS department, Carnegie Mellon University, Pittsburgh, PA, 1999.
[14] C. Guestrin, D. Koller, and R. Parr, *Max-norm projections for factored mdps*, Proceedings of the International Joint Conference on Artificial Intelligence, (2001).
[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, 2001.
[16] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes, Basic Optimality Criteria*, Springer-Verlag, New-York, 1996.
[17] ———, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New-York, 1999.
[18] O. Hernández-Lerma, R. Montes-De-Oca, and R. Cavazos-Caneda, *Recurrence conditions for Markov decision processes with borel state space: A survey*, Annals of Operation Research, 28 (1991), pp. 29–46.
[19] A. Hordijk and F. Spieksma, *On ergodicity and recurrence properties of a Markov chain with an application to an open Jackson network*, Advances in Applied Probabilities, (1992), pp. 343–376.
[20] S. Kakade and J. Langford, *Approximately optimal approximate reinforcement learning*, Proceedings of the 19th International Conference on Machine Learning, (2002).
[21] S. M. Kakade, *On the Sample Complexity of Reinforcement Learning*, PhD thesis, Gatsby Computationel Neuroscience Unit, University College London, 2003.
[22] D. Koller and R. Parr, *Policy iteration for factored mdps*, Proceedings of the 16th conference on Uncertainty in Artificial Intelligence, (2000).
[23] M. Lagoudakis and R. Parr, *Least-squares policy iteration*, Journal of Machine Learning Research, 4 (2003), pp. 1107–1149.
[24] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1997.
[25] S. P. Meyn, *Stability, performance evaluation, and optimization*, Handbook of Markov Decision Processes: Methods and Applications, (2001), pp. 305–346.
[26] R. Munos, *Error bounds for approximate policy iteration*, 19th International Conference on Machine Learning, (2003).
[27] R. Munos and C. Szepesvári, *Finite time bounds for sampling based fitted value iteration*,

Technical report, Computer and Automation Research Institute of the Hungarian Academy of Sciences, (2006).

[28] D. Pollard, *Convergence of Stochastic Processes*, Springer Verlag, New York, 1984.

[29] M. L. Puterman, *Markov Decision Processes, Discrete Stochastic Dynamic Programming*, A Wiley-Interscience Publication, 1994.

[30] E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2005.

[31] J. Rust, *Numerical Dynamic Programming in Economics*, In Handbook of Computational Economics, Elsevier, North Holland, 1996.

[32] A. Samuel, *Some studies in machine learning using the game of checkers*, IBM Journal on Research and Development, (1959), pp. 210–229. Reprinted in *Computers and Thought*, E.A. Feigenbaum and J. Feldman, editors, McGraw-Hill, New York, 1963.

[33] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, Bradford Book, (1998).

[34] C. Szepesvari and R. Munos, *Finite time bounds for sampling based fitted value iteration*, International Conference on Machine Learning, (2005).

[35] J. Tsitsiklis and B. Van Roy, *An analysis of temporal difference learning with function approximation*, IEEE Transactions on Automatic Control, 42(5) (1997), pp. 674–690.

[36] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[37] V. Vapnik, S. E. Golowich, and A. Smola, *Support vector method for function approximation, regression estimation and signal processing*, In Advances in Neural Information Processing Systems, (1997), pp. 281–287.

[38] R. Williams and L. Baird, *Tight performance bounds on greedy policies based on imperfect value functions*, Technical Report NU-CCS-93-14. Northeastern University, (1993).