# Error Bounds for Approximate Policy Iteration

**Rémi Munos** REMI.MUNOS@POLYTECHNIQUE.FR
Centre de Mathématiques Appliquées. Ecole Polytechnique. 91128 Palaiseau, France.
http://www.cmap.polytechnique.fr/∼ munos/

## Abstract

In Dynamic Programming, convergence of algorithms such as Value Iteration or Policy Iteration results –in discounted problems– from a contraction property of the back-up operator, guaranteeing convergence to its fixed-point. When approximation is considered, known results in *Approximate Policy Iteration* provide bounds on the closeness to optimality of the approximate value function obtained by successive policy improvement steps as a function of the *maximum norm* of value determination errors during policy evaluation steps. Unfortunately, such results have limited practical range since most function approximators (such as linear regression) select the best fit in a given class of parameterized functions by minimizing some (weighted) *quadratic norm*.

In this paper, we provide error bounds for Approximate Policy Iteration using quadratic norms, and illustrate those results in the case of feature-based linear function approximation.

## 1. Introduction

We consider a *Markov Decision Process* (MDP) (Puterman, 1994; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998) evolving on a state space $X$ with $N$ states. Its dynamics is governed by the *transition probability* function $P(i,a,j)$ which gives the probability that the next state is $j \in X$ knowing that the current state is $i \in X$ and the chosen action is $a \in A$, where $A$ is the (finite) set of possible actions. A *policy* $\pi$ is a mapping from $X$ to $A$. We write $P^\pi$ the $N \times N-$matrix whose elements are $P^\pi(i,j) = p(i,\pi(i),j)$. Let $r(i,a,j)$ be the reward received when a transition from state $i$, action $a$, to state $j$ occurs. Write $r^\pi$ the vector whose components are $r^\pi(i) = \sum_j P^\pi(i,j)\, r(i,\pi(i),j)$. Here, we consider discounted, infinite horizon problems.

The *value function* $V^\pi(i)$ for a policy $\pi$ is the expected sum of discounted future rewards when starting from state $i$ and using policy $\pi$:

$$V^\pi(i) = \mathbb{E}_{i,\pi}[\sum_{t=0}^{\infty} \gamma^t\, r_t]$$

where $r_t$ is the reward received at time $t$ and $\gamma \in [0,1)$ a discount factor. It is known that $V^\pi$ solves the Bellman equation

$$V^\pi(i) = r^\pi(i) + \gamma \sum_{j\in X} P^\pi(i,j)\, V^\pi(j).$$

Thus $V^\pi$ (considered as a vector of size $N$) is the fixed-point of the *back-up operator* $T^\pi$ defined by $T^\pi\cdot = r^\pi + \gamma P^\pi\cdot$. Since $P^\pi$ is a stochastic matrix, it possesses eigenvalues with module less than or equal to one, thus $(I-\gamma P^\pi)$ is invertible, and we write $V^\pi = (I-\gamma P^\pi)^{-1} r^\pi$.

The *optimal value function* $V^*$ is the expected gain when using an optimal policy $\pi^*$: $V^* = V^{\pi^*} = \sup_\pi V^\pi$. We are interested in problems with large state spaces ($N$ is very large, possibly infinite), which prevents us from using exact resolution methods such as Value Iteration or Policy Iteration with look-up tables. Instead, we consider the **Approximate Policy Iteration** algorithm (Bertsekas & Tsitsiklis, 1996) defined iteratively by the two steps:

- *Approximate policy evaluation*: for a given policy $\pi_k$, generate an approximation $V_k$ of the value function $V^{\pi_k}$
- *Policy improvement*: generate a new policy $\pi_{k+1}$ greedy with respect to $V_k$:

$$\pi_{k+1}(i) = \arg\max_{a\in A} \sum_{j\in X} [r(i,a,j) + \gamma\, p(i,a,j) V_k(j)]$$

These steps are repeated until no more improvement of the policies is noticed (using some evaluation criterion). Empirically, the value functions $V^{\pi_k}$ rapidly improve in the first iterations of this algorithm, then oscillations occur with no more performance increase. The behavior in the transitional phase is due to the

*Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003.

relatively good approximation of the value function ($||V_k - V^{\pi_k}||$ is low) in comparison to the closeness to optimality $||V^{\pi_k} - V^*||$, which produces greedy policies (with respect to the approximate $V_k$) that are better than the current policies. Then, once some closeness to optimality is reached, the error in the value approximation prevents the policy improvement step from being efficient: the stationary phase is attained. Hence, this algorithm does not converge (there is no stabilization to some policy) but it is very fast and from the intuition above, we can expect to quantify the closeness to optimality at the stationary phase as a function of the value approximation errors. And indeed, a known result (Bertsekas & Tsitsiklis, 1996, chap. 6.2) provides bounds on the loss $V^* - V^{\pi_k}$ of using policy $\pi_k$ instead of using the optimal one, as a function of the *maximum norm* of the approximation errors $V_k - V^{\pi_k}$:

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \sup_k ||V_k - V^{\pi_k}||_\infty \tag{1}$$

However, this result is difficult to use in many approximation architectures (exceptions include (Gordon, 1995; Guestrin et al., 2001)) since it is very costly to control the maximum norm; the weighted quadratic norms are more commonly used. We recall that a distribution $\mu$ on $X$ defines an inner-product $\langle f, h\rangle_\mu = \sum_{i=1}^N \mu(i)f(i)h(i)$ and a quadratic (semi-) norm $||h||_\mu = \langle h, h\rangle_\mu^{1/2}$. Of course, equivalency between norms implies that $||h|| \leq ||h||_\infty \leq \sqrt{N}||h||$ (where $||\cdot||$ denotes the norm defined by the uniform distribution $\rho \equiv \frac{1}{N}$). But then, the bound (1), rewritten in quadratic norm will include the factor $\sqrt{N}$, which is too large for being of any use in most cases.

Our main result, stated in Section 2 and proved in Appendix A, is to derive analogous bounds in *quadratic norms:* the loss $||V^* - V^{\pi_k}||_\mu$ (for any distribution $\mu$) is bounded by a function of the approximation error $||V_k - V^{\pi_k}||_{\mu_k}$ (for some distribution $\mu_k$ related to $\mu$ and the policies $\pi_k$ and $\pi^*$), as well as by the Bellman residual (Baird, 1995) $||V_k - T^{\pi_k}V_k||_{\widetilde{\mu}_k}$ (for another distribution $\widetilde{\mu}_k$).

In Section 3, we apply those results to the feature-based linear function approximation (where the parameterized functions are weighted linear combinations of basis functions −the features), which have been considered in Temporal Difference learning TD($\lambda$) (Tsitsiklis & Van Roy, 1996) and Least-Squares Temporal Difference: LSTD(0) (Bradtke & Barto, 1996), LSTD($\lambda$) (Boyan, 1999), and LS-Q-learning (Lagoudakis & Parr, 2001).

Both the approximations obtained by minimizing the quadratic Bellman residual and by finding the TD solution (the fixed-point of a combined operator) are considered in sections 3.2 and 3.3. Under the assumption of *uniform stochasticity* of the MDP (Hypothesis 2), bounds on $||V^* - V^{\pi_k}||_\infty$ are derived based on the minimum possible approximation error $\inf_\alpha ||V_\alpha - V^\pi||_{\rho_\pi}$. Proofs are given in Appendix B.

These linear approximation architectures combined with policy improvement still lack theoretical analysis but have produced very promising experimental results on large scale control and reinforcement learning problems (Lagoudakis & Parr, 2001); we hope that this paper will help better understand their behavior.

## 2. Quadratic Norm Bounds

Consider the Approximate Policy Iteration algorithm described in the introduction. $\pi_k$ represents the policy at iteration $k$, and $V_k$ the approximation of the value function $V^{\pi_k}$. The main result of this paper is stated in this theorem.

**Theorem 1** *For any distribution $\mu$ (considered as a row vector) on $X$, define the stochastic matrices*

$$\begin{aligned} Q_k &= \frac{(1-\gamma)^2}{2}(I-\gamma P^{\pi^*})^{-1}[P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1} \\ &\quad + P^{\pi^*}(I-\gamma P^{\pi_k})^{-1}] \\ \widetilde{Q}_k &= \frac{(1-\gamma)^2}{2}(I-\gamma P^{\pi^*})^{-1}[P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1} \\ &\quad (I+\gamma P^{\pi_k}) + P^{\pi^*}] \end{aligned}$$

*Write $\mu_k = \mu Q_k$ and $\widetilde{\mu}_k = \mu\widetilde{Q}_k$. Then $\mu_k$ and $\widetilde{\mu}_k$ are distributions on $X$, and*

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} ||V_k - T^{\pi_k}V_k||_{\mu_k} \tag{2}$$

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} ||V_k - V^{\pi_k}||_{\widetilde{\mu}_k} \tag{3}$$

Some intuition about this result as well as its proof may be found in Appendix A.

Notice that this result is stronger than the bound in max-norm (1), since from (3) and using the fact that $||\cdot||_{\widetilde{\mu}_k} \leq ||\cdot||_\infty$, we deduce that $\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2} \sup_k ||V_k - V^{\pi_k}||_\infty$ for any distribution $\mu$, which implies (1).

Moreover, it provides information about what parts of the state-space are responsible (in terms of local approximation error $V_k - V^{\pi_k}$) for the loss $V^*(i) - V^{\pi_k}(i)$ at any state $i$. This information indicates the areas of the state space where we should focus our efforts in the value approximation (e.g. by locally reallocating computational resources, such as in variable resolution discretization (Munos & Moore, 2002)).

In the next section we describe how to use this result to derive error bounds on the loss $V^* - V^{\pi_k}$ in the case of linear approximation architectures.

## 3. Approximate Policy Evaluation

### 3.1. Linear feature-based approximation

We consider a class of functions $V_\alpha = \Phi\alpha$ linearly parameterized by a parameter $\alpha$ (vector of size $K$, usually much smaller than $N$), where $\Phi$ is the set of basis functions, called *features* (a $N \times K$ matrix in which each column represents a feature).

We assume that the columns of $\Phi$ are linearly independent. Such linear architectures include state aggregation methods, CMACs, polynomial or wavelet regression techniques, radial basis function networks with fixed bases, and finite-element methods. They have been used in incremental Temporal Difference TD($\lambda$) (Tsitsiklis & Van Roy, 1996) or Least-Squares TD (LSTD) (Bradtke & Barto, 1996), (Boyan, 1999). These LSTD methods which "makes efficient use of training samples collected in any arbitrary manner" have recently been extended to model-free LS-Q-learning (Lagoudakis & Parr, 2001). They have demonstrated very good efficiency in reinforcement learning and control of large scale problems.

The space of parameterized functions is written $[\Phi]$ (the span of the columns of $\Phi$). At iteration $k$, the approximate policy evaluation step selects a "good" approximation $V_{\alpha_k}$ (written $V_k$ for simplicity) of the value function $V^{\pi_k}$, in the sense that some (semi-) norm $||V_k - V^{\pi_k}||_{\rho_k}$ be minimized, as much as possible. Several approaches for this minimization problem are possible (Bertsekas & Tsitsiklis, 1996; Schoknecht, 2002; Judd, 1998):

- Find the **optimal approximate solution**, which is the best possible approximation in $[\Phi]$: $V_k$ is the orthogonal projection $\Pi_{\rho_k} V^{\pi_k}$ of $V^{\pi_k}$ onto $[\Phi]$ with respect to the norm $||\cdot||_{\rho_k}$. This regression problem is very costly since $V^{\pi_k}$ is unknown, but estimations may be obtained by Monte-Carlo simulations.

- Find the **minimal quadratic residual (QR) solution**, which is the function $V_k$ that minimizes the quadratic Bellman residual $||V_\alpha - T^{\pi_k} V_\alpha||_{\rho_k}$. This problem is easy to solve since it reduces to the resolution of a linear system of size $K$: Find $\alpha$ such that

$$A\alpha = b \text{ with} \begin{cases} A = \Phi^T (I - \gamma P^{\pi_k})^T D_{\rho_k} (I - \gamma P^{\pi_k}) \Phi \\ b = \Phi^T (I - \gamma P^{\pi_k})^T D_{\rho_k} r^{\pi_k} \end{cases} \tag{4}$$

  where $D_{\rho_k}$ is the $N \times N$ diagonal matrix whose elements are $D_{\rho_k}(i,i) = \rho_k(i)$. This problem always admits a solution since $A$ is invertible.

- Find the **Temporal Difference (TD) solution**, which is the fixed-point of the conjugate operator $\Pi_{\rho_k} T^{\pi_k}$ – the back-up operator followed by the projection onto $[\Phi]$ w.r.t $||\cdot||_{\rho_k}$– i.e. $V_k$ satisfies $V_k = \Pi_{\rho_k} T^{\pi_k} V_k$. Again, this problem reduces to a linear system of size $K$: Find $\alpha$ such that

$$A\alpha = b \text{ with} \begin{cases} A = \Phi^T D_{\rho_k} (I - \gamma P^{\pi_k}) \Phi \\ b = \Phi^T D_{\rho_k} r^{\pi_k} \end{cases} \tag{5}$$

  Here, $A$ is not always invertible.

The matrix $A$ and vector $b$ of the QR and TD solutions may be estimated from transition data coming from arbitrary sources, e.g. incrementally (Boyan, 1999) from the observation of trajectories induced by a given policy or by random policies (Lagoudakis & Parr, 2001), or by archived data coming from prior knowledge.

Thus, one needs to specify the distribution $\rho_k$ used in the minimization problem, which usually depends on the policy $\pi_k$. A steady-state distribution $\overline{\rho}_{\pi_k}$, which would weight more the states that are frequently visited, would be desirable for purely value determination. However, the policy improvement step may perform badly since, from Lemma 3 (see Appendix A), the gain in policy improvement depends on the value approximation at states reached by policy $\pi_{k+1}$ as well as their successors (for policy $\pi_k$), which may be poorly approximated if they are ill-represented in $\overline{\rho}_{\pi_k}$. A more uniform distribution $\rho_k$ would give weight to all states thus insuring a more secure policy improvement step (Koller & Parr, 2000; Kakade & Langford, 2002). We consider these possible choices for $\rho_k$:

- Steady-state distribution $\overline{\rho}_{\pi_k}$ (if a such exists). It satisfies the property $\overline{\rho}_{\pi_k} = \overline{\rho}_{\pi_k} P^{\pi_k}$.

- Constant distribution $\overline{\mu}$ (does not depend on $\pi_k$).

- Mixed distribution $\rho^\lambda_{\pi_k} = \overline{\mu}(I - \lambda P^{\pi_k})^{-1}(1-\lambda)$ (for $0 \leq \lambda < 1$), which starts from an initial distribution $\overline{\mu}$, then transitions induced by $\pi$ occur for a period of time that is a random variable that follows an exponential law $\lambda^t(1-\lambda)$. Thus $\rho^\lambda_{\pi_k}$ corresponds to the distribution of a Markov chain that starts from a state sampled according to $\overline{\mu}$

- and which, at each iteration, either follows policy $\pi$ with probability $\lambda$ or restarts to a new state with probability $1-\lambda$. Notice that when $\lambda$ tends to 0 then $\rho^{\lambda}_{\pi_k}$ tends to the constant distribution $\overline{\mu}$, and when $\lambda$ tends to 1, $\rho^{\lambda}_{\pi_k}$ tends to the steady-state distribution.

- Convex combination of constant and steady-state distributions: $\rho^{\delta}_{\pi_k} = (1-\delta)\overline{\mu} + \delta\,\overline{\rho}_{\pi_k}$.

Now, in order to bound the approximation error $||V_k - V^{\pi_k}||_{\widetilde{\mu}_k}$ and Bellman residual $||V_k - T^{\pi_k}V_k||_{\mu_k}$ (to be used in Theorem 1) as a function of the minimum possible approximation error $\inf_\alpha ||V_\alpha - V^{\pi}||_{\rho_k}$, we need some assumption about the representational power of the approximation architecture.

**Hypothesis 1 (Approximation hypothesis)** *For any policy $\pi$, there exists, in the class of parameterized functions, an $\epsilon-$approximation (in $\rho_\pi-$norm) of the value function $V^\pi$: for some $\varepsilon > 0$, for all policies $\pi$,*

$$\inf_{\alpha} ||V_\alpha - V^\pi||_{\rho_\pi} \leq \varepsilon$$

*where $\rho_\pi$ may depend on the policy $\pi$.*

Next, we study the cases where the approximate function $V_k$ is chosen to be the QR solution (subsection 3.2) and the TD solution (subsection 3.3).

### 3.2. The Quadratic Residual solution

Consider $\alpha_k$ the parameter that minimizes the Bellman residual in quadratic $\rho_k-$norm (solution to (4)). Write $V_k = V_{\alpha_k} = \Phi\alpha_k$ the corresponding value function:

$$||V_k - T^{\pi_k}V_k||_{\rho_k} = \inf_{\alpha} ||V_\alpha - T^{\pi_k}V_\alpha||_{\rho_k}$$

Since, for all $\alpha$, $V_\alpha - T^{\pi_k}V_\alpha = (I-\gamma P^{\pi_k})(V_\alpha - V^{\pi_k})$, we deduce that

$$\begin{aligned} ||V_k - T^{\pi_k}V_k||_{\rho_k} &= \inf_{\alpha} ||(I-\gamma P^{\pi_k})(V_\alpha - V^{\pi_k})||_{\rho_k} \\ &\leq |||I-\gamma P^{\pi_k}|||_{\rho_k}\,\varepsilon \qquad (6) \end{aligned}$$

where $|||\cdot|||_{\rho_k}$ is the matrix norm induced by $||\cdot||_{\rho_k}$ (i.e. $|||A|||_\rho := \sup_{||x||_\rho = 1} ||Ax||_\rho$). Now we have a bound on the residual $V_k - T^{\pi_k}V_k$ in $\rho_k-$norm, but in Theorem 1 we actually need such a bound in $\mu_k-$norm. A crude (but somehow unavailable) bound is

$$||V_k - T^{\pi_k}V_k||^2_{\mu_k} \leq ||\frac{\mu_k}{\rho_k}||_\infty\, ||V_k - T^{\pi_k}V_k||^2_{\rho_k} \qquad (7)$$

where $||\frac{\mu_k}{\rho_k}||_\infty$ express the mismatch between the rather unknown distribution $\mu_k = \mu Q_k$ and the distribution $\rho_k$ used in the minimization problem. In order to bound this ratio, we now provide conditions for which a upper-bound for $\mu_k$ and a lower-bound for $\rho_k$ are possible.

We make the following assumption on the MDP.

**Hypothesis 2 (Uniform stochasticity)** *Let $\overline{\mu}$ be some distribution, for example a uniform distribution. There exists a constant $C$, such that for all policies $\pi$, for all $i, j \in X$,*

$$P^\pi(i,j) \leq C\overline{\mu}(j) \qquad (8)$$

Notice that this hypothesis can always be satisfied for $\overline{\mu}(i) = 1/N$ by choosing $C = N$. However, we are actually interested in finding a constant $C << N$, which requires, intuitively, that each state possesses many successors with rather small corresponding transition probabilities.

**Remark 1** *An interesting case for which this assumption is satisfied is when the MDP has continuous-space (thus $N = \infty$ but all ideas in previous analysis remain valid). In such case, if the continuous problem has a transition probability kernel $P^\pi(x,B)$ (probability that the next state belongs to the subset $B \subset X$ when the current state is $x \in X$ and the chosen action is $\pi(x)$), then the hypothesis reads that there exists a measure $\overline{\mu}$ on $X$ (with $\overline{\mu}(X) = 1$) such that $P^\pi(x,B) \leq C\overline{\mu}(B)$ for all $x$ and all subset $B$. This is true as long as the transition probabilities admit a pdf representation: $p^\pi(x,B) = \int_B p^\pi(y|x)dy$ with bounded density $p^\pi(\cdot|x)$.*

From this assumption, we derive a bound for $\mu_k$:

**Lemma 1** *Assume Hypothesis 2. Then $\mu_k \leq C\overline{\mu}$.*

**Remark 2** *An assumption on the Markov process, other than Hypothesis 2, that would guarantee an upper-bound for $\mu_k$ is that the matrix $P^\pi$ and the resolvent $(I-\gamma P^\pi)^{-1}(1-\gamma)$ have bounded entrant probabilities: there exists two constants $C_1 << N$ and $C_2 << N$ such that for all $\pi$ and all $j \in X$*

$$\begin{aligned} \sum_{i\in X} P^\pi(i,j) &\leq C_1 \\ (1-\gamma)\sum_{i\in X} [(I-\gamma P^\pi)^{-1}](i,j) &\leq C_2 \end{aligned}$$

*then, a bound is $\mu_k \leq C_1 C_2^2 \mu$ (we will not prove this result here). An important case for which this assumption is satisfied is when the MDP is built from a discretization of a (continuous-time) Markov diffusion process for which the state-dynamics are governed by stochastic differential equations with non-degenerate diffusion coefficients.*

The distributions $\rho^{\lambda}_{\pi_k}$ and $\rho^{\delta}_{\pi_k}$ previously defined, which mix the steady-state distribution to a rather uniform distribution $\overline{\mu}$, can be lower-bounded when $\lambda < 1$ or $\delta < 1$, which allows to use inequalities (7) and (2) to derive an error bound on the loss $V^* - V^{\pi_k}$ when using the QR solution:

**Theorem 2** *Assume that Hypothesis 2 holds with some distribution $\overline{\mu}$ and constant $C$.*

- *Assume that Hypothesis 1 holds with the distribution $\rho^{\lambda}_{\pi_k} = \overline{\mu}(I - \lambda P^{\pi_k})^{-1}(1-\lambda)$ (with $0 \le \lambda < 1$), then*

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \le \frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{C}{1-\lambda}}\left(1 + \gamma\sqrt{\min(\frac{C}{1-\lambda}, \frac{1}{\lambda})}\right)\varepsilon$$

- *Assume that Hypothesis 1 holds with the distribution $\rho^{\delta}_{\pi_k} = (1-\delta)\overline{\mu} + \delta\overline{\rho}_{\pi_k}$ (with $0 \le \delta < 1$) (where $\overline{\rho}_{\pi_k}$ is the steady-state distribution for $\pi_k$), then*

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \le \frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{C}{1-\delta}}\left(1 + \gamma\sqrt{C}\right)\varepsilon$$

**Remark 3** *Note that if the steady-state distribution $\overline{\rho}_\pi$ is itself lower-bounded by some constant $\overline{\rho} > 0$, then the bounds on $\rho^{\lambda}_{k}$, $\rho^{\delta}_{k}$ can be tightened for $\lambda$ and $\delta$ close to $1$, which would suppress the terms $1-\lambda$ and $1-\delta$ in the denominators of the right hand side of the above inequalities.*

### 3.3. Temporal Difference solution

Now, we consider that $V_k$ is the Temporal Difference solution, i.e. the fixed-point of the combined operator $\Pi_{\rho_k} T^{\pi_k}$. We notice that $V_k$ solves the system (equivalent to (5) because $\Phi$ has full column rank):

$$(I - \gamma\Pi_{\rho_k}P^{\pi_k})V_k = V_k - \Pi_{\rho_k}(T^{\pi_k}V_k - r^{\pi_k}) = \Pi_{\rho_k}r^{\pi_k} \tag{9}$$

which has a solution if the matrix $(I - \gamma\Pi_{\rho_k}P^{\pi_k})$ is invertible. The *approximation error* $e_k = V_k - V^{\pi_k}$ solves the system

$$\begin{aligned}(I - \gamma\Pi_{\rho_k}P^{\pi_k})e_k &= \Pi_{\rho_k}r^{\pi_k} - V^{\pi_k} + \Pi_{\rho_k}(T^{\pi_k}V^{\pi_k} - r^{\pi_k}) \\ &= \Pi_{\rho_k}V^{\pi_k} - V^{\pi_k} := \varepsilon_k \end{aligned} \tag{10}$$

where $\varepsilon_k$ is the *optimal approximation error.*

#### 3.3.1. Which distribution?

If $\rho_k$ is the steady-state distribution $\overline{\rho}_{\pi_k}$ for policy $\pi_k$, then we have $|||P^{\pi_k}|||_{\overline{\rho}_{\pi_k}} = 1$ (Tsitsiklis & Van Roy, 1996). Thus, if Hypothesis 1 is satisfied for the steady-state distribution, then from (10), we deduce a bound on the approximation error

$$\begin{aligned}||V_k - V^{\pi_k}||_{\overline{\rho}_{\pi_k}} &\le |||(I - \gamma\Pi_{\overline{\rho}_{\pi_k}}P^{\pi_k})^{-1}|||_{\overline{\rho}_{\pi_k}}\varepsilon \\ &\le \frac{1}{1 - \gamma|||P^{\pi_k}|||_{\overline{\rho}_{\pi_k}}}\varepsilon \le \frac{\varepsilon}{1-\gamma}\end{aligned} \tag{11}$$

Now, if $\rho_k$ is different from $\overline{\rho}_{\pi_k}$ then $|||P^{\pi_k}|||_{\rho_k}$ (which is always $\ge 1$ since $P^{\pi_k}$ is a stochastic matrix) may be greater than $1/\gamma$ and (11) does not hold any more. Even if we assume that for all policies $\pi_k$ the matrices $I - \gamma\Pi_{\rho_k}P^{\pi_k}$ are invertible (thus, that the $V_k$ are well-defined), which means that the eigenvalues of $\Pi_{\rho_k}P^{\pi_k}$ are all different from $1/\gamma$, it seems difficult to provide bounds on the approximation error $e_k = (I - \gamma\Pi_{\rho_k}P^{\pi_k})^{-1}\varepsilon_k$ because those eigenvalues may be close to $1/\gamma$: we can easily build simple examples for which the ratio of $||e_k||_{\rho_k}$ (as well as the Bellman residual $||V_k - T^{\pi_k}V_k||_{\rho_k} = ||(I - \gamma P^{\pi_k})e_k||_{\rho_k}$) by $\varepsilon$ is as large as desired. Some numerical experiments showed that the TD solution provided better policies than the QR solution although the value functions were not so accurately approximated. The reason argued was that the TD solution "preserved the shape of the value function to some extent rather than trying to fit the absolute values", thus "the improved policy from the approximate value function is "closer" to the improved policy from the corresponding exact value function" (Lagoudakis & Parr, 2001). More formally, this would mean that the difference between the backed-up errors using $\pi_{k+1}$ and another policy $\pi$

$$d^{\pi}_{k} := T^{\pi_{k+1}}(V_k - V^{\pi_k}) - T^{\pi}(V_k - V^{\pi_k})$$

is small for $\pi = \pi^*_{k+1}$, the greedy policy w.r.t. $V^{\pi_k}$. Since $\pi^*_{k+1}$ is unknown, $d^{\pi}_{k}$ would need to be small for any policy $\pi$. We have

$$\begin{aligned} d^{\pi}_{k} &= \gamma(P^{\pi_{k+1}} - P^{\pi})e_k \\ &= \gamma(P^{\pi_{k+1}} - P^{\pi})(I - \gamma\Pi_{\rho_k}P^{\pi_k})^{-1}\varepsilon_k \end{aligned}$$

Thus, there are two possibilities: either $e_k$ belongs to the intersection (for all $\pi$) of the kernels of $(P^{\pi_{k+1}} - P^{\pi})$, in which case $d^{\pi}_{k}$ is zero, or if this is not the case, $d^{\pi}_{k}$ is also unstable whenever the eigenvalues of $\Pi_{\rho_k}P^{\pi_k}$ are close to $1/\gamma$. The first case, which would be ideal (since then, $\pi_{k+1}$ would be equal to $\pi^*_{k+1}$) does not hold in general. Indeed, if it was true, this would mean that $e_k$ is collinear to the unit vector $\mathbf{1} := (1 1 \dots 1)^T$, say $e_k = c_k\mathbf{1}$ for some scalar $c_k$ (then, $V_k$ would be equal to $V^{\pi_k}$ up to an additive constant) and we would have $\varepsilon_k = (I - \gamma\Pi_{\rho_k}P^{\pi_k})e_k = c_k(I - \gamma\Pi_{\rho_k})\mathbf{1}$. But, by definition, $\varepsilon_k$ is orthogonal to $[\Phi]$ w.r.t. the inner product $\langle\cdot,\cdot\rangle_{\rho_k}$ whereas the vector

$(I-\gamma\Pi_{\rho_k})\mathbf{1}$ is not (for $\gamma < 1$) in general (the exception being if $\mathbf{1}$ is orthogonal to $[\Phi]$ w.r.t. $\langle\cdot,\cdot\rangle_{\rho_k}$). Thus, as soon as the eigenvalues of $\Pi_{\rho_k}P^{\pi_k}$ are close to $1/\gamma$, the approximation error $e_k$ as well as the difference in the backed-up errors $d_k^\pi$ becomes large.

Thus, we believe that in general, the TD solution is less stable and predictable (as long as we do not control the eigenvalues of $\Pi_{\rho_k}P^{\pi_k}$) than the QR solution. However, the TD solution may be preferable in model-free Reinforcement Learning, when unbiased estimators of $A$ and $b$ in (4) and (5) need to be derived from observed data (Munos, 2003).

### 3.3.2. Steady-state distribution

If we consider the case of the steady-state distribution and assume that it is bounded from below (for example if all policies induce an irreducible, aperiodic Markov chain (Puterman, 1994)), we are able to derive the following error bound on the loss $V^* - V^{\pi_k}$.

**Theorem 3** *Assume that Hypothesis 2 holds for a distribution $\overline{\mu}$ (for example uniform) and a constant $C$, that Hypothesis 1 holds with the steady-state distributions $\overline{\rho}_\pi$, and that $\overline{\rho}_\pi$ is bounded from below by $\frac{1}{\kappa}\overline{\mu}$ (with $\kappa$ a constant), then*

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \leq \frac{2\gamma}{(1-\gamma)^3}\sqrt{\kappa C}\,\varepsilon$$

## 4. Conclusion

The main contribution of this paper is the error bounds on $||V^* - V^{\pi_k}||_\mu$ derived as a function of the approximation errors $||V_k - V^{\pi_k}||_{\widetilde{\mu}_k}$ and the Bellman residuals $||V_k - T^{\pi_k}V_k||_{\mu_k}$. The distributions $\mu_k$ and $\widetilde{\mu}_k$ indicate the states that are responsible for the approximation accuracy. An application of this result to linear function approximation is derived and error bounds that do not depend on the number of states are given, provided that the MDP satisfies some uniform stochasticity assumption (that leads to an upper-bound for $\mu_k$ and $\widetilde{\mu}_k$) and that the distribution used in the minimization problem is lower-bounded (in order to insure some reliability of the value approximation uniformly over the state-space, which secures policy improvement steps). In the case of the QR solution, this was guaranteed by using a somehow uniform mixed distribution, whereas in the case of the TD solution, we assumed that the steady-state distribution was already bounded from below.

## A. Proof of Theorem 1.

Define the **approximation error**: $e_k = V_k - V^{\pi_k}$, the **gain** between iteration $k$ and $k+1$: $g_k = V^{\pi_{k+1}} - V^{\pi_k}$, the **loss** of using policy $\pi_k$ instead of the optimal policy: $l_k = V^* - V^{\pi_k}$, and the **Bellman residual** of the approximate value function: $b_k = V_k - T^{\pi_k}V_k$. Those $e_k$, $g_k$, $l_k$, and $b_k$ are column vectors of size $N$. We first state and prove the following results:

**Lemma 2** *It is true that:*

$$l_{k+1} \leq \gamma[P^{\pi^*}l_k + P^{\pi_{k+1}}(e_k - g_k) - P^{\pi^*}e_k]$$

Proof: Indeed,

$$\begin{aligned} l_{k+1} &= T^{\pi^*}V^* - T^{\pi^*}V^{\pi_k} + T^{\pi^*}V^{\pi_k} - T^{\pi^*}V_k \\ &\quad + T^{\pi^*}V_k - T^{\pi_{k+1}}V_k + T^{\pi_{k+1}}V_k \\ &\quad - T^{\pi_{k+1}}V^{\pi_k} + T^{\pi_{k+1}}V^{\pi_k} - T^{\pi_{k+1}}V^{\pi_{k+1}} \\ &\leq \gamma[P^{\pi^*}l_k + P^{\pi_{k+1}}(V^{\pi_k} - V^{\pi_{k+1}}) \\ &\quad + (P^{\pi_{k+1}} - P^{\pi^*})(V_k - V^{\pi_k})] \end{aligned}$$

where we used the fact that $T^{\pi^*}V_k - T^{\pi_{k+1}}V_k \leq 0$ since $\pi_{k+1}$ is greedy with respect to $V_k$. □

**Lemma 3** *It is true that:*

$$g_k \geq -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})\,e_k$$

Proof: Indeed,

$$\begin{aligned} g_k &= T^{\pi_{k+1}}V^{\pi_{k+1}} - T^{\pi_{k+1}}V^{\pi_k} + T^{\pi_{k+1}}V^{\pi_k} - T^{\pi_{k+1}}V_k \\ &\quad + T^{\pi_{k+1}}V_k - T^{\pi_k}V_k + T^{\pi_k}V_k - T^{\pi_k}V^{\pi_k} \\ &\geq \gamma P^{\pi_{k+1}}g_k - \gamma(P^{\pi_{k+1}} - P^{\pi_k})\,e_k \\ &\geq -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})\,e_k \end{aligned}$$

since $T^{\pi_{k+1}}V_k - T^{\pi_k}V_k \geq 0$. □

**Lemma 4** *It is true that:*

$$\begin{aligned} l_{k+1} &\leq \gamma P^{\pi^*}(V^* - V^{\pi_k}) + \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} \\ &\quad - P^{\pi^*}(I - \gamma P^{\pi_k})^{-1}]b_k \end{aligned} \tag{12}$$

*Or equivalently*

$$\begin{aligned} l_{k+1} &\leq \gamma P^{\pi^*}(V^* - V^{\pi_k}) + \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} \\ &\quad (I - \gamma P^{\pi_k}) - P^{\pi^*}]e_k \end{aligned} \tag{13}$$

Proof: From Lemma 3, we have

$$\begin{aligned} e_k - g_k &\leq [I - \gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_k} - P^{\pi_{k+1}})]e_k \\ &\leq (I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k})e_k \end{aligned}$$

and (13) follows from Lemma 2. Inequality (12) is derived by factorizing $(I - \gamma P^{\pi_k})$ and by noticing that $(I - \gamma P^{\pi_k})e_k = V_k - V^{\pi_k} - T^{\pi_k}(V_k - V^{\pi_k}) = V_k - T^{\pi_k}V_k = b_k$ is the Bellman residual of the approximate function $V_k$, which terminates the proof. □

Now, from Lemma 4, we derive the following results:

**Corollary 1** *We have*

$$\limsup_{k\to\infty} l_k \leq \gamma(I-\gamma P^{\pi^*})^{-1} \limsup_{k\to\infty} [P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*}(I-\gamma P^{\pi_k})^{-1}]\, b_k \tag{14}$$

*or equivalently that*

$$\limsup_{k\to\infty} l_k \leq \gamma(I-\gamma P^{\pi^*})^{-1} \limsup_{k\to\infty} [P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1} (I-\gamma P^{\pi_k}) - P^{\pi^*}] e_k$$

Proof: Write $f_k = \gamma[P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*}(I-\gamma P^{\pi_k})^{-1}]\, b_k$. Then, from Lemma 4, $l_{k+1} \leq \gamma P^{\pi^*} l_k + f_k$. By taking the limit superior component-wise

$$(I-\gamma P^{\pi^*}) \limsup_{k\to\infty} l_k \leq \limsup_{k\to\infty} f_k$$

And the result follows since $I-\gamma P^{\pi^*}$ is invertible. The proof of the other inequality is similar. □

**Corollary 2** *By defining the stochastic matrices $Q_k$ and $\widetilde{Q}_k$ as in Theorem 1, we have*

$$\begin{aligned}\limsup_{k\to\infty} l_k &\leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} Q_k |b_k| \\ \limsup_{k\to\infty} l_k &\leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \widetilde{Q}_k |e_k|\end{aligned}$$

*where $|b_k|$ and $|e_k|$ are vectors whose components are $|b_k(i)|$ and $|e_k(i)|$.*

Proof: First, the fact that $Q_k$ and $\widetilde{Q}_k$ are stochastic matrices is a consequence of the properties that if $P_1$ and $P_2$ are stochastic matrices, then $P_1P_2$, $\frac{P_1+P_2}{2}$, and $(1-\gamma)(I-\gamma P_1)^{-1}$ are stochastic matrices too (the third property resulting from the two first and the rewriting of $(I-\gamma P_1)^{-1}$ as $\sum_{t\geq 0}\gamma^t P_1^t$). The result follows when taking the absolute value in the inequalities of Corollary 1. □

Now we are able to prove Theorem 1:

The fact that $\mu_k$ and $\widetilde{\mu}_k$ are distributions (positive vectors whose components sum to one) results from $Q_k$ and $\widetilde{Q}_k$ being stochastic matrices. Let us prove (2). For any vector $h$, define $h^2$ the vector whose components are $h_i^2$. We have, from the convexity of the square function and from Corollary 2,

$$\begin{aligned}\limsup_{k\to\infty} ||l_k||_\mu^2 &= \limsup_{k\to\infty} \mu\, l_k^2 \\ &\leq \frac{4\gamma^2}{(1-\gamma)^4} \limsup_{k\to\infty} \mu[Q_k|b_k|]^2 \\ &\leq \frac{4\gamma^2}{(1-\gamma)^4} \limsup_{k\to\infty} \mu Q_k b_k^2 \\ &\leq \frac{4\gamma^2}{(1-\gamma)^4} \limsup_{k\to\infty} \mu_k b_k^2\end{aligned}$$

Thus $\limsup_{k\to\infty} ||l_k||_\mu \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} ||b_k||_{\mu_k}$. Inequality (3) is deduced similarly. □

**Remark 4** *Some intuition about these bounds may be perceived in a specific case: assume that the policy $\pi_k$ were to converge, say to $\widetilde{\pi}$, and write $\widetilde{V}$ the approximation of $V^{\widetilde{\pi}}$. Then from Corollary 1,*

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

*The right hand side of this inequality measures the expected difference between the backed-up approximation errors using $\widetilde{\pi}$ and $\pi^*$with respect to the discounted future state-distribution induced by the optimal policy. Thus here, the states responsible for the approximation accuracy are the states reached by the optimal policy as well as their successors (for policy $\widetilde{\pi}$).*

## B. Proofs of Section 3

### Proof of Lemma 1

First, for two stochastic matrices $P_1$ and $P_2$ satisfying (8), for all $i$ and $j$, we have $(P_1P_2)(i,j) = \sum_k P_1(i,k)P_2(k,j) \leq C\overline{\mu}(j)\sum_k P_1(i,k) = C\overline{\mu}(j)$ and recursively, for all $k$, $(P_1)^k(i,j) \leq C\overline{\mu}(j)$. Thus also $(1-\gamma)(I-\gamma P_1)^{-1}(i,j) = (1-\gamma)\sum_{t\geq 0}\gamma^t(P_1)^t(i,j) \leq C\overline{\mu}(j)$.

We deduce that $Q_k$ defined in Theorem 1 satisfies $Q_k(i,j) \leq C\overline{\mu}(j)$. Thus, $\mu_k(j) = (\mu Q_k)(j) = \sum_i \mu(i)Q_k(i,j) \leq C\overline{\mu}(j)\sum_i \mu(i) = C\overline{\mu}(j)$. □

### Proof of Theorem 2

Let us state and prove the two Lemmas:

**Lemma 5** *Lower bounds for $\rho_k^\lambda$ and $\rho_k^\delta$.*

*We have $\rho_k^\lambda \geq (1-\lambda)\,\overline{\mu}$ and $\rho_k^\delta \geq (1-\delta)\,\overline{\mu}$.*

Proof: We have $\rho_k^\lambda = (1-\lambda)\overline{\mu}\sum_{t\geq 0}\gamma^t(P^{\pi_k})^t \geq (1-\lambda)\overline{\mu}$, and $\rho_k^\delta = (1-\delta)\overline{\mu} + \delta\overline{\rho}_{\pi_k} \geq (1-\delta)\overline{\mu}$. □

**Lemma 6** *Upper bound for $|||P^{\pi_k}|||_{\rho_k^\lambda}$ and $|||P^{\pi_k}|||_{\rho_k^\delta}$.*

*We have*

$$|||P^{\pi_k}|||_{\rho_k^\lambda}^2 \leq \min(\frac{C}{1-\lambda}, \frac{1}{\lambda}) \textit{ and } |||P^{\pi_k}|||_{\rho_k^\delta}^2 \leq C$$

Proof: First consider $\rho_k^\lambda$. From Hypothesis 2,

$$\begin{aligned}||P^{\pi_k}h||_{\rho_k^\lambda}^2 &= \rho_k^\lambda\,(P^{\pi_k}h)^2 \leq \rho_k^\lambda\, P^{\pi_k}h^2 \\ &\leq C\overline{\mu}h^2 = C||h||_{\overline{\mu}}^2\end{aligned}$$

Moreover, $\overline{\mu} = \frac{1}{1-\lambda}\rho_k^\lambda(I - \lambda P^{\pi_k}) \leq \frac{1}{1-\lambda}\rho_k^\lambda$, thus $||h||_{\overline{\mu}}^2 = \overline{\mu}h^2 \leq \frac{1}{1-\lambda}\rho_k^\lambda h^2 = \frac{1}{1-\lambda}||h||_{\rho_k^\lambda}^2$. Therefore, for all $h$, $||P^{\pi_k}h||^2_{\rho_k^\lambda} \leq \frac{C}{1-\lambda}||h||^2_{\rho_k^\lambda}$. Now, it is also true that

$$\begin{aligned}
||P^{\pi_k}h||^2_{\rho_k^\lambda} &= (1-\lambda)\overline{\mu}\sum_{t=0}^{\infty}\lambda^t(P^{\pi_k})^t(P^{\pi_k}h)^2 \\
&\leq (1-\lambda)\overline{\mu}\sum_{t=0}^{\infty}\lambda^t(P^{\pi_k})^{t+1}h^2 \\
&\leq \frac{1-\lambda}{\lambda}\overline{\mu}(\sum_{t=0}^{\infty}\lambda^t(P^{\pi_k})^t h^2 - h^2) \\
&\leq \frac{1}{\lambda}\rho_k^\lambda h^2 = \frac{1}{\lambda}||h||^2_{\rho_k^\lambda}
\end{aligned}$$

Thus $|||P^{\pi_k}|||^2_{\rho_k^\lambda} \leq \min(\frac{C}{1-\lambda}, \frac{1}{\lambda})$.

Now consider $\rho_k^\delta$. For any vector $h$,

$$\begin{aligned}
||P^{\pi_k}h||^2_{\rho_k^\delta} &= \rho_k^\delta(P^{\pi_k}h)^2 \\
&\leq (1-\delta)\overline{\mu}P^{\pi_k}h^2 + \delta\overline{\rho}_k P^{\pi_k}h^2 \\
&\leq C(1-\delta)\overline{\mu}h^2 + \delta\overline{\rho}_k h^2 \\
&\leq C(1-\delta)||h||^2_{\overline{\mu}} + \delta||h||^2_{\overline{\rho}_k}
\end{aligned}$$

(where we used the property of the steady distribution $\overline{\rho}_k = \overline{\rho}_k P^{\pi_k}$). Moreover, $\overline{\mu} = \frac{1}{1-\delta}(\rho_k^\delta - \delta\overline{\rho}_k)$, thus $||h||^2_{\overline{\mu}} = \frac{1}{1-\delta}(||h||^2_{\rho_k^\delta} - \delta||h||^2_{\overline{\rho}_k})$. Thus $||P^{\pi_k}h||^2_{\rho_k^\delta} \leq C(||h||^2_{\rho_k^\delta} - \delta||h||^2_{\overline{\rho}_k}) + \delta||h||_{\overline{\rho}_k} \leq C||h||^2_{\rho_k^\delta}$ since $C \geq 1$. Thus $|||P^{\pi_k}h|||^2_{\rho_k^\delta} \leq C$. □

**Proof of Theorem 2:**

For any distribution $\mu$, putting together (2), (7) and (6), we have $\limsup_{k\to\infty}||l_k||_\mu \leq \frac{2\gamma}{(1-\gamma)^2}\limsup_{k\to\infty}\sqrt{||\frac{\mu_k}{\rho_k}||_\infty}|||I-\gamma P^{\pi_k}|||_{\rho_{\pi_k}}\varepsilon$.

Now, from Lemmas 1, 5, 6, and by using the fact that $|||I-\gamma P^\pi|||_{\rho_\pi} \leq 1+\gamma|||P^\pi|||_{\rho_\pi}$, we deduce the bound in $||\cdot||_\mu$, but since this is true for any distribution $\mu$, the same bound holds in $||\cdot||_\infty$. □

**Proof of Theorem 3**

For any distribution $\mu$, let $\widetilde{\mu}_k = \mu\widetilde{Q}_k$ with $\widetilde{Q}_k$ defined in Theorem 1. Analogously to (7) we have $||e_k||^2_{\widetilde{\mu}_k} \leq ||\frac{\widetilde{\mu}_k}{\overline{\rho}_k}||_\infty\,||e_k||^2_{\overline{\rho}_k}$. Similarly to Lemma 1, we have $\widetilde{\mu}_k \leq C\overline{\mu}$, thus $||\frac{\widetilde{\mu}_k}{\overline{\rho}_k}||_\infty \leq \kappa C$. Since $\overline{\rho}_k$ is the steady-state distribution, $||e_k||_{\overline{\rho}_k} \leq \frac{\varepsilon}{1-\gamma}$, thus $||e_k||_{\widetilde{\mu}_k} \leq \sqrt{\kappa C}\,\frac{\varepsilon}{1-\gamma}$, and from (3),

$$\limsup_{k\to\infty}||l_k||_\mu \leq \frac{2\gamma}{(1-\gamma)^3}\sqrt{\kappa C}\,\varepsilon$$

and since this bound holds for any distribution $\mu$, it also holds in max-norm. □

## References

Baird, L. C. (1995). Residual algorithms : Reinforcement learning with function approximation. *Machine Learning : proceedings of the Twelfth International Conference.*

Bertsekas, D. P., & Tsitsiklis, J. (1996). *Neuro-dynamic programming.* Athena Scientific.

Boyan, J. (1999). Least-squares temporal difference learning. *Proceedings of the 16th International Conference on Machine Learning*, 49–56.

Bradtke, S., & Barto, A. (1996). Linear least-squares algorithms for temporal difference learning. *Journal of Machine Learning, 22*, 33–57.

Gordon, G. (1995). Stable function approximation in dynamic programming. *Proceedings of the International Conference on Machine Learning.*

Guestrin, C., Koller, D., & Parr, R. (2001). Max-norm projections for factored mdps. *Proceedings of the International Joint Conference on Artificial Intelligence.*

Judd, K. (1998). *Numerical methods in economics.* MIT Press.

Kakade, S., & Langford, J. (2002). Approximately optimal approximate reinforcement learning. *Proceedings of the 19th International Conference on Machine Learning.*

Koller, D., & Parr, R. (2000). Policy iteration for factored mdps. *Proceedings of the 16th conference on Uncertainty in Artificial Intelligence.*

Lagoudakis, M., & Parr, R. (2001). Model free least-squares policy iteration. *Technical Report CS-2001-05, Department of Computer Science, Duke University.*

Munos, R. (2003). Experiments in policy iteration with linear approximation. *Submitted to the European Conference on Machine Learning.*

Munos, R., & Moore, A. (2002). Variable resolution discretization in optimal control. *Machine Learning Journal, 49*, 291–323.

Puterman, M. L. (1994). *Markov decision processes, discrete stochastic dynamic programming.* A Wiley-Interscience Publication.

Schoknecht, R. (2002). Optimality of reinforcement learning algorithms with linear function approximation. *Proceedings of the 15th Neural Information Processing Systems conference.*

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. *Bradford Book.*

Tsitsiklis, J., & Van Roy, B. (1996). An analysis of temporal difference learning with function approximation. *Technical report LIDS-P-2322, MIT.*