Value Pursuit Iteration

Amir-massoud Farahmand[*] **Doina Precup** [†]
School of Computer Science, McGill University, Montreal, Canada

Abstract

Value Pursuit Iteration (VPI) is an approximate value iteration algorithm that finds a close to optimal policy for reinforcement learning problems with large state spaces. VPI has two main features: First, it is a nonparametric algorithm that finds a good sparse approximation of the optimal value function given a dictionary of features. The algorithm is almost insensitive to the number of irrelevant features. Second, after each iteration of VPI, the algorithm adds a set of functions based on the currently learned value function to the dictionary. This increases the representation power of the dictionary in a way that is directly relevant to the goal of having a good approximation of the optimal value function. We theoretically study VPI and provide a finite-sample error upper bound for it.

1 Introduction

One often has to use function approximation to represent the near optimal value function of the reinforcement learning (RL) and planning problems with large state spaces. Even though the conventional approach of using a parametric model for the value function has had successes in many applications [1, 2, 3, 4], it has one main weakness: Its success critically depends on whether the chosen function approximation method is suitable for the particular task in hand. Manually designing a suitable function approximator, however, is difficult unless one has considerable domain knowledge about the problem. To address this issue, the problem-dependent choice of function approximator and nonparametric approaches to RL/Planning problems have gained considerable attention in the RL community, e.g., feature generation methods of Petrik [5], Mahadevan and Maggioni [6], Parr et al. [7], Geramifard et al. [8], and nonparametric regularization-based approaches of Jung and Polani [9], Xu et al. [10], Farahmand et al. [11, 12], Taylor and Parr [13] .

One class of approaches that addresses the aforementioned problem is based on the idea of finding a sparse representation of the value function in a large dictionary of features (or *atoms*). In this approach, the designer does not necessarily know a priori whether or not a feature is relevant to the representation of the value function. The feature, therefore, is simply added to the dictionary with the hope that the algorithm itself figures out the necessary subset of features. The usual approach to tackle irrelevant features is to use sparsity-inducing regularizers such as the l_1-norm of the weights in the case of linear function approximators, e.g., Kolter and Ng [14], Johns et al. [15], Ghavamzadeh et al. [16]. Another approach is based on greedily adding atoms to the representation of the target function. Examples of these approaches in the supervised learning setting are Matching Pursuit [17] (or Pure Greedy Algorithm) and Orthogonal Matching Pursuit (OMP) [18, 19] (also known as Orthogonal Greedy Algorithm) (cf. Temlyakov [20] for more information on matching pursuit (or greedy) type of algorithms). These greedy algorithms have successfully been used in the signal processing and statistics/supervised machine learning communities for years, but their application in the RL/Planning problems has just recently attracted some attention. Johns [21] empirically investigated some greedy algorithms, including OMP, for the task of feature selection using dictionary of proto-value functions [6]. A recent paper by Painter-Wakefield and Parr [22] considers two al-

[*] Academic.SoloGen.net.

[†] This research was funded in part by NSERC and ONR.

gorithms (OMP-TD and OMP-BRM; OMP-TD is the same as one of the algorithms by [21]) in the context of policy evaluation and provides some conditions under which OMP-BRM can find the minimally sparse solution. Moreover, they show that OMP-TD cannot recover an s-sparse value function in s iterations.

To address the problem of value function representation in RL when not much a priori knowledge is available, we introduce the **Value Pursuit Iteration (VPI)** algorithm. VPI, which is an Approximate Value Iteration (AVI) algorithm (e.g., [23]), has two main features. The first is that it is a nonparametric algorithm that finds a good sparse approximation of the optimal value function given a set of features (dictionary), by using a modified version of OMP. The second is that after each iteration, the VPI algorithm adds a set of functions based on the currently learned value function to the dictionary. This potentially increases the representation power of the dictionary in a way that is directly relevant to the goal of approximating the optimal value function.

At the core of VPI is the OMP algorithm equipped with a model selection procedure. Using OMP allows VPI to find a sparse representation of the value function in large dictionaries, even countably infinite ones[1]. This property is very desirable for RL/Planning problems for which one usually does not know the right representation of the value function, and so one wishes to add as many features as possible and to let the algorithm automatically detect the best representation. A model selection procedure ensures that OMP is adaptive to the actual difficulty of the problem.

The second main feature of VPI is that it increases the size of the dictionary by adding some basis functions computed from previously learned value functions. To give an intuitive understanding of how this might help, consider the dictionary $\mathcal{B} = \{g_1, g_2, \dots\}$, in which each atom g_i is a real-valued function defined on the state-action space. The goal is to learn the optimal value function by a representation in the form of $Q = \sum_{i \geq 1} w_i g_i$.[2] Suppose that we are lucky and the optimal value function Q^* belongs to the dictionary $\mathcal{B}$, e.g., $g_1 = Q^*$. This is indeed an ideal atom to have in the dictionary since one may have a sparse representation of the optimal value function in the form of $Q^* = \sum_{i \geq 1} w_i g_i$ with $w_1 = 1$ and $w_i = 0$ for $i \geq 2$. Algorithms such as OMP can find this sparse representation quite effectively (details will be specified later). Of course, we are not usually lucky enough to have the optimal value function in our dictionary, but we may still use approximation of the optimal value function. In the exact Value Iteration, $Q_k \to Q^*$ exponentially fast(i.e., $\|Q_k - Q^*\|_\infty \leq \gamma^k \|Q_0 - Q^*\|_\infty$). This ensures that Q_k and $Q_{k+1} = T^* Q_k$ are close enough, so one may use Q_k to explain a large part of Q_{k+1} and use the other atoms of the dictionary to "explain" the residual(it is easy to see that $\|T^* Q_k - Q_k\|_\infty \leq (1+\gamma)\gamma^k \|Q^* - Q_0\|_\infty$). In an AVI procedure, however, the estimated value function sequence $(Q_k)_{k \geq 1}$ does not necessarily converge to Q^*, but one may hope that it gets close to a region around the optimum. In that case, we may very well use the dictionary of $\{Q_1, \dots, Q_k\}$ as the set of candidate atoms to be used in the representation of Q_{k+1}. We show that adding these learned atoms does not hurt and may actually help.

One may also interpret what VPI does as a form of deep representation learning. After the k-th iteration of VPI, the resulting output Q_k is provided as an input to the new iteration of learning. This new input in addition to the the initial dictionary as well as all other estimated value functions provide a rich input representation for the function approximator. As opposed to the conventional deep learning procedures, we use a supervised signal to train each layer. Thus, one may consider VPI as an *extremely deep learning architecture* for the optimal value function learning and representation.

To summarize, the algorithmic contribution of this paper is to introduce the VPI algorithm that finds a sparse representation of the optimal value function in a huge function space and increases the representation capacity of the dictionary problem-dependently. The theoretical contribution of this work is to provide a finite-sample analysis of VPI and to show that the method is sound. We analyze how the errors from earlier iterations affect the function approximation error of the current iteration. This, alongside an analysis of the estimation error at each iteration, leads to an upper bound on the error in approximating $T^* Q_k$ by Q_{k+1} at each iteration of VPI. Finally, we show how these errors are propagated through iterations and affect the performance loss of the resulting policy.

[1] From the statistical viewpoint and ignoring the computational difficulty of working with large dictionaries.

[2] The notation will be defined precisely in Section 2.

2 Definitions

We follow the standard notation and definitions of Markov Decision Processes (MDP) and Reinforcement Learning (RL) (cf. [24]). The definitions can be found in Appendix E. We also need some definitions regarding the function spaces and norms, which are defined later in this section.

For a space Ω with σ-algebra σ_Ω, $\mathcal{M}(\Omega)$ denotes the set of all probability measures over σ_Ω. $B(\Omega)$ denotes the space of bounded measurable functions w.r.t. (with respect to) σ_Ω and $B(\Omega, L)$ denotes the subset of $B(\Omega)$ with bound $0 < L < \infty$.

A *finite-action discounted MDP* is a 5-tuple $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$, where $\mathcal{X}$ is a measurable state space, $\mathcal{A}$ is a finite set of actions, $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$ is the transition probability kernel, $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathbb{R})$ is the reward kernel, and $\gamma \in [0, 1)$ is a discount factor. Let $r(x, a) = \mathbb{E}[\mathcal{R}(\cdot|x, a)]$, and assume that r is uniformly bounded by $R_{\max}$. A measurable mapping $\pi : \mathcal{X} \to \mathcal{A}$ is called a deterministic Markov stationary policy, or just a *policy* for short. A policy π induces the m-step transition probability kernels $(P^\pi)^m : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ and $(P^\pi)^m : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$ for $m \geq 1$.

We use V^π and Q^π to denote the value and action-value function of a policy π. We also use V^* and Q^* for the optimal value and optimal action-value functions, with the corresponding optimal policy π^*. A policy π is *greedy* w.r.t. an action-value function Q, denoted $\pi = \hat{\pi}(\cdot; Q)$, if $\pi(x) = \operatorname{argmax}_{a \in \mathcal{A}} Q(x, a)$ holds for all $x \in \mathcal{X}$ (if there exist multiple maximizers, one of them is chosen in an arbitrary deterministic manner). Define $Q_{\max} = R_{\max}/(1 - \gamma)$. The Bellman optimality operator is denoted by T^*. We use $(PV)(x)$ to denote the expected value of V after the transition according to a probability transition kernel P. Also for a probability measure $\rho \in \mathcal{M}(\mathcal{X})$, the symbol (ρP) represents the distribution over states when the initial state distribution is ρ and we follow P for a single step. A typical choice of P is $(P^\pi)^m$ for $m \geq 1$ (similarly for $\rho \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ and action-value functions).

2.1 Norms and Dictionaries

For a probability measure $\rho \in \mathcal{M}(\mathcal{X})$, and a measurable function $V \in B(\mathcal{X})$, we define the $L_p(\rho)$-norm ($1 \leq p < \infty$) of V as $\|V\|_{p,\rho} \triangleq \left[\int_{\mathcal{X}} |V(x)|^p \, \mathrm{d}\rho(x)\right]^{1/p}$. The $L_\infty(\mathcal{X})$-norm is defined as $\|V\|_\infty \triangleq \sup_{x \in \mathcal{X}} |V(x)|$. Similarly for $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ and $Q \in B(\mathcal{X} \times \mathcal{A})$, we define $\|\cdot\|_{p,\nu}$ as $\|Q\|_{p,\nu}^p \triangleq \int_{\mathcal{X} \times \mathcal{A}} |Q(x, a)|^p d\nu(x, a)$ and $\|Q\|_\infty \triangleq \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |Q(x, a)|$.

Let $z_{1:n}$ denote the $\mathcal{Z}$-valued sequence $(z_1, \ldots, z_n)$. For $\mathcal{D}_n = z_{1:n}$, define the empirical norm of function $f : \mathcal{Z} \to \mathbb{R}$ as $\|f\|_{p,z_{1:n}}^p = \|f\|_{p,\mathcal{D}_n}^p \triangleq \frac{1}{n} \sum_{i=1}^n |f(z_i)|^p$. Based on this definition, one may define $\|V\|_{\mathcal{D}_n}$ (with $\mathcal{Z} = \mathcal{X}$) and $\|Q\|_{\mathcal{D}_n}$ (with $\mathcal{Z} = \mathcal{X} \times \mathcal{A}$). Note that if $\mathcal{D}_n = Z_{1:n}$ is random with $Z_i \sim \nu$, the empirical norm is random as well. For any fixed function f, we have $\mathbb{E}\left[\|f\|_{p,\mathcal{D}_n}\right] = \|f\|_{p,\nu}$. The symbols $\|\cdot\|_\nu$ and $\|\cdot\|_{\mathcal{D}_n}$ refer to an L_2-norm. When we do not want to emphasize the underlying measure, we use $\|\cdot\|$ to denote an L_2-norm.

Consider a Hilbert space $\mathcal{H}$ endowed with an inner product norm $\|\cdot\|$. We call a family of functions $\mathcal{B} = \{g_1, g_2, \ldots, \}$ with atoms $g_i \in \mathcal{H}$ a *dictionary*. The class $\mathcal{L}_1(\mathcal{B}) = \mathcal{L}_1(\mathcal{B}; \|\cdot\|)$ consists of those functions $f \in \mathcal{H}$ that admits an expansion $f = \sum_{g \in \mathcal{B}} c_g g$ with (c_g) being an absolutely summable sequence (these definitions are quoted from Barron et al. [25]; also see [26, 20]). The norm of a function f in this space is defined as $\|f\|_{\mathcal{L}_1(\mathcal{B};\|\cdot\|)} \triangleq \inf\{\sum_{g \in \mathcal{B}} |c_g| : f = \sum_{g \in \mathcal{B}} c_g g\}$. To avoid clutter, when the norm is the empirical norm $\|\cdot\|_{\mathcal{D}_n}$, we may use $\mathcal{L}_1(\mathcal{B}; \mathcal{D}_n)$ instead of $\mathcal{L}_1(\mathcal{B}; \|\cdot\|_{\mathcal{D}_n})$, and when the norm is $\|\cdot\|_\nu$, we may use $\mathcal{L}_1(\mathcal{B}; \nu)$. We denote a ball with radius $r > 0$ w.r.t. the norm of $\mathcal{L}_1(\mathcal{B}; \nu)$ by $B_r(\mathcal{L}_1(\mathcal{B}; \nu))$.

For a dictionary $\mathcal{B}$, we introduce a fixed exhaustion $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \ldots \subset \mathcal{B}$, with the number of atoms $|\mathcal{B}_m|$ being m. If we index our dictionaries as $\mathcal{B}_k$, the symbol $\mathcal{B}_{k,m}$ refers to the m-th element of the exhaustion of $\mathcal{B}_k$. For a real number $\alpha > 0$, the space $\mathcal{L}_{1,\alpha}(\mathcal{B}; \|\cdot\|)$ is defined as the set of all functions f such that for all $m = 1, 2, \ldots$, there exists a function h depending on m such that $\|h\|_{\mathcal{L}_1(\mathcal{B}_m;\|\cdot\|)} \leq C$ and $\|f - h\| \leq C m^{-\alpha}$. The smallest constant C such that these inequalities hold defines a norm for $\mathcal{L}_{1,\alpha}(\mathcal{B}; \|\cdot\|)$. Finally, we define the truncation operator $\beta_L : B(\mathcal{X}) \to B(\mathcal{X})$ for some real number $L > 0$ as follows. For any function $f \in B(\mathcal{X})$, the truncated function of f at

the threshold level L is the function $\beta_L f : B(\mathcal{X}) \to \mathbb{R}$ such that for any $x \in \mathcal{X}$,

$$\beta_L f(x) \triangleq \begin{cases} L & \text{if} \quad f(x) > L, \\ f(x) & \text{if} \quad -L \le f(x) \le L, \\ -L & \text{if} \quad f(x) < -L. \end{cases} \tag{1}$$

We overload β_L to be an operator from $B(\mathcal{X} \times \mathcal{A})$ to $B(\mathcal{X} \times \mathcal{A})$ by applying it component-wise, i.e., $\beta_L Q(x, \cdot) \triangleq [\beta_L Q(x, a_1), \dots, \beta_L Q(x, a_{\mathcal{A}})]^\top$.

3 VPI Algorithm

In this section, we first describe the behaviour of VPI in the ideal situation when the Bellman optimality operator T^* can be applied exactly in order to provide the intuitive understanding of why VPI might work. Afterwards, we describe the algorithm that does not have access to the Bellman optimality operator and only uses a finite sample of transitions.

VPI belongs to the family of AVI algorithms, which start with an initial action-value function Q_0 and at each iteration $k = 0, 1, \dots$, approximately apply the Bellman optimality operator T^* to the most recent estimate Q_k to get a new estimate $Q_{k+1} \approx T^* Q_k$. The size of the error between Q_{k+1} and $T^* Q_k$ is a key factor in determining the performance of an AVI procedure.

Suppose that $T^* Q_k$ can be calculated, but it is not possible to represent it exactly. In this case, one may use an approximant Q_{k+1} to represent $T^* Q_k$. In this paper we would like to represent Q_{k+1} as a linear function of some atoms in a dictionary $\mathcal{B} = \{g_1, g_2, \dots\}$ ($g \in \mathcal{H}(\mathcal{X} \times \mathcal{A})$ and $\|g\| = 1$ for all $g \in \mathcal{B}$), that is $Q_{k+1} = \sum_{g \in \mathcal{B}} c_g g$. Our goal is to find a representation that is as sparse as possible, i.e., uses only a few atoms in $\mathcal{B}$. From statistical viewpoint, the smallest representation among all those that have the same function approximation error is desirable as it leads to smaller estimation error. The goal of finding the sparsest representation, however, is computationally intractable. Nevertheless, it is possible to find a "reasonable" suboptimal sparse approximation using algorithms such as OMP, which is the focus of this paper.

The OMP algorithm works as follows. Let $\tilde{Q}^{(0)} = 0$. For each $i = 1, 2, \dots$, define the residual $r^{(i-1)} = T^* Q_k - \tilde{Q}^{(i-1)}$. Define the new atom to be added to the representation as $g^{(i)} \in \operatorname{Argmax}_{g \in \mathcal{B}} \left| \left\langle r^{(i-1)}, g \right\rangle \right|$, i.e., choose an element of the dictionary that has the maximum correlation with the residual. Here $\langle \cdot, \cdot \rangle$ is the inner product for a Hilbert space $\mathcal{H}(\mathcal{X} \times \mathcal{A})$ to which $T^* Q_k$ and atoms of the dictionary belong. Let $\Pi^{(i)}$ be the orthogonal projection onto $\operatorname{span}(g^{(1)}, \dots, g^{(i)})$, i.e., $\Pi^{(i)} T^* Q_k \triangleq \operatorname{argmin}_{Q \in \operatorname{span}(g^{(1)}, \dots, g^{(i)})} \|Q - T^* Q_k\|$. We then have $\tilde{Q}^{(i)} = \Pi^{(i)} T^* Q_k$. OMP continues iteratively.

To quantify the approximation error at the i-th iteration, we use the $\mathcal{L}_1(\mathcal{B}; \|\cdot\|)$-norm of the target function of the OMP algorithm, which is $T^* Q_k$ in our case (with the norm being the one induced by the inner product used in the OMP procedure). Recall that this class consists of functions that admit an expansion in the form $\sum_{g \in \mathcal{B}} c_g g$ and (c_g) being an absolutely summable sequence. If $T^* Q_k$ belongs to the class of $\mathcal{L}_1(\mathcal{B}; \|\cdot\|)$, it can be shown (e.g., Theorem 2.1 of Barron et al. [25]) that after i iterations of OMP, the returned function $\tilde{Q}^{(i)}$ is such that $\|\tilde{Q}^{(i)} - T^* Q_k\| \le \frac{\|T^* Q_k\|_{\mathcal{L}_1(\mathcal{B}; \|\cdot\|)}}{\sqrt{i+1}}$. The problem with this result is that it requires $T^* Q_k$ to belong to $\mathcal{L}_1(\mathcal{B}; \|\cdot\|)$. This depends on how expressive the dictionary $\mathcal{B}$ is. If it is not expressive enough, we still would like OMP to quickly converge to the best approximation of $T^* Q_k \notin \mathcal{L}_1(\mathcal{B}; \|\cdot\|)$ in $\mathcal{L}_1(\mathcal{B}; \|\cdot\|)$. Fortunately, such a result exists (Theorem 2.3 by Barron et al. [25], quoted as Lemma 4 in Appendix A) and we use it in the proof of our main result.

We now turn to the more interesting case when we do not have access to $T^* Q_k$. Instead we are only given a set of transitions in the form of $\mathcal{D}_n^{(k)} = \{(X_i^{(k)}, A_i^{(k)}, R_i^{(k)}, X_i'^{(k)})\}_{i=1}^n$, where $(X_i^{(k)}, A_i^{(k)})$ are drawn from the sampling distribution $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$, $X_i' \sim P(\cdot | X_i, A_i)$, and $R_i \sim \mathcal{R}(\cdot | X_i, A_i)$. Instead of using $T^* Q_k$, we use the empirical Bellman operator for the dataset $\mathcal{D}_n^{(k)}$. The operator is defined as follows.

Definition 1 (Empirical Bellman Optimality Operator)**.** *Let* $\mathcal{D}_n = \{(X_1, A_1, R_1, X_1'), \dots, (X_n, A_n, R_n, X_n')\}$*, defined similarly as above. Define the ordered multi-*

Algorithm 1 Value Pursuit Iteration$(\mathcal{B}_0, m, \{\sigma_i\}_{i=1}^{m'}, \nu, K)$

Input: Initial dictionary $\mathcal{B}_0$, Number of dictionary atoms used m, Link functions $\{\sigma_i\}_{i=1}^{m'}$, State-action distribution ν, Number of iterations K.
Return: Q_K
$Q_0 \leftarrow 0$.
$\mathcal{B}'_0 \leftarrow \emptyset$.
for $k = 0, 1, \ldots, K-1$ **do**
  Construct a dataset $\mathcal{D}_n^{(k)} = \left\{(X_i^{(k)}, A_i^{(k)}, R_i^{(k)}, X_i'^{(k)})\right\}_{i=1}^{n}$, $(X_i^{(k)}, A_i^{(k)}) \overset{\text{i.i.d.}}{\sim} \nu$
  $\hat{Q}_{k+1}^{(0)} \leftarrow 0$
  // Orthogonal Matching Pursuit loop
  Normalize elements of $\mathcal{B}_{0,m}$ and $\mathcal{B}'_k$ according to $\|\cdot\|_{\mathcal{D}_n^{(k)}}$ and call them $\hat{\mathcal{B}}_k$ and $\hat{\mathcal{B}}'_k$.
  for $i = 1, 2, \ldots, c_1 n$ **do**
    $r^{(i-1)} \leftarrow \hat{T}^* Q_k - \hat{Q}_{k+1}^{(i-1)}$
    $g^{(i)} \leftarrow \text{Argmax}_{g \in \hat{\mathcal{B}}_k \bigcup \hat{\mathcal{B}}'_k} \left| \left\langle r^{(i-1)}, g \right\rangle_{\mathcal{D}_n^{(k)}} \right|$
    $\hat{Q}_{k+1}^{(i)} \leftarrow \Pi^{(i)} \hat{T}^* Q_k$ {$\Pi^{(i)}$: Projection onto $\text{span}(g^{(1)}, \ldots, g^{(i)})$}
  end for
  $i^* \leftarrow \text{argmin}_{i \geq 1} \left\{ \left\| \beta_{Q_{\max}} \hat{Q}_{k+1}^{(i)} - \hat{T}^* Q_k \right\|_{\mathcal{D}_n^{(k)}}^2 + c_2 (Q_{\max}) \frac{i \ln(n)}{n} \right\}$ {Complexity Regularization}
  $Q_{k+1} \leftarrow \hat{Q}_{k+1}^{(i^*)}$
  $\mathcal{B}'_{k+1} \leftarrow \mathcal{B}'_k \bigcup \{\sigma_i(\beta_{Q_{\max}} Q_{k+1}; \mathcal{B}_k \bigcup \mathcal{B}'_k)\}_{i=1}^{m'}$ {Extending the dictionary}
end for

set $S_n = \{(X_1, A_1), \ldots, (X_n, A_n)\}$. *The empirical Bellman optimality operator* $\hat{T}^* : S_n \to \mathbb{R}^n$ *is defined as* $(\hat{T}^* Q)(X_i, A_i) \triangleq R_i + \gamma \max_{a'} Q(X'_i, a')$ *for* $1 \leq i \leq n$.

Since $\mathbb{E}\left[\hat{T}^* Q_k(X_i^{(k)}, A_i^{(k)}) \,\middle|\, Q_k, X_i^{(k)}, A_i^{(k)}\right] = T^* Q_k(X_i^{(k)}, A_i^{(k)})$, we can solve a regression problem and find an estimate for Q_{k+1}, which is close $T^* Q_k$. This regression problem is the core of the family of Fitted Q-Iteration (FQI) algorithms, e.g., [23, 12]. In this paper, the regression function at each iteration is estimated using a modified OMP procedure introduced by Barron et al. [25].

We are now ready to describe the VPI algorithm (Algorithm 1). It gets as input a predefined dictionary $\mathcal{B}_0$. This can be a dictionary of wavelets, proto-value functions, etc. The size of this dictionary can be countably infinite. It also receives an integer m, which specifies how many atoms of $\mathcal{B}_0$ should be used by the algorithm. This defines the effective dictionary $\mathcal{B}_{0,m}$. This value can be set to $m = \lceil n^a \rceil$ for some finite $a > 0$, so it can actually be quite large. This value shows that the effective size of the dictionary can grow even faster than the number of samples (but not exponentially faster).VPI also receives K, the number of iterations, and ν, the sampling distribution. For the simplicity of analysis, we assume that the sampling distribution is fixed, but in practice one may change this sampling distribution after each iteration (e.g., sample new data according to the latest policy). Finally, VPI gets a set of m' link functions $\sigma_i : B(\mathcal{X} \times \mathcal{A}, Q_{\max}) \to B(\mathcal{X} \times \mathcal{A}, Q_{\max})$ for some m' that is smaller than m/K. We describe the role of link functions shortly.

At the k-th iteration of the algorithm, we perform OMP for $c_1 n$ iterations ($c_1 > 0$), similar to what is described above with the difference that instead of using $T^* Q_k$ as the target, we use $\hat{T}^* Q_k$ over empirical samples.[3] This means that we use the empirical inner product $\langle Q_1, Q_2 \rangle_{\mathcal{D}_n^{(k)}} \triangleq \frac{1}{n} \sum_{i=1}^{n} |Q_1(X_i, A_i) \cdot Q_2(X_i, A_i)|$ for $(X_i, A_i) \in \mathcal{D}_n^{(k)}$ and the empirical orthogonal projection.[4] The result would be a sequence $(\hat{Q}_{k+1}^{(i)})_{i \geq 0}$. Next, we perform a model selection procedure to choose the best candidate. This can be done in different ways such as using a separate dataset as a validation

[3] The value of c_1 depends only on $Q_{\max}$ and a. We do not explicitly specify it since the value that is determined by the theory shall be quite conservative. One may instead find it by the trial and error. Moreover, in practice we may stop much earlier than n iterations.

[4] When the number of atoms is larger than the number of samples ($i > n$), one may use the Moore–Penrose pseudoinverse to perform the orthogonal projection.

set. Here we use a complexity regularization technique that penalizes more complex estimates (those that have more atoms in their representation). Note that we use the truncated estimate $\beta_{Q_{\max}}\hat{Q}_{k+1}^{(i)}$ in the model selection procedure (cf. (1)). This is required for the theoretical guarantees. The outcome of this model selection procedure will determine Q_{k+1}.

Finally we use link functions $\{\sigma_i\}_{i=1}^{m'}$ to generate m' new atoms, which are vector-valued $Q_{\max}$-bounded measurable functions from $\mathcal{X} \times \mathcal{A}$ to $\mathbb{R}^{|\mathcal{A}|}$, to be added to the learned dictionary $\mathcal{B}'_k$. The link functions extract "interesting" aspects of Q_{k+1}, potentially by considering the current dictionary $\mathcal{B}_k \bigcup \mathcal{B}'_k$. VPI is quite flexible in how the new atoms are generated and how large m' can be. The theory allows m' to be in the order of n^a ($a > 0$), so one may add many potentially useful atoms without much deterioration in the performance. Regarding the choice of the link functions, the theory requires that at least Q_{k+1} itself is being added to the dictionary, but it leaves other possibilities open. For example, one might apply nonlinearities (e.g., sigmoid functions) to Q_{k+1}. Or one might add atoms localized in parts of the state-action space with high residual errors – a heuristic which has been used previously in basis function construction. This procedure continues for K iterations and the outcome will be Q_K. In the next section, we study the theoretical properties of the greedy policy w.r.t. Q_K, i.e., $\pi_K = \hat{\pi}(\cdot; Q_K)$.

Remark 1 (Comparison of VPI with FQI). Both VPI and FQI are indeed instances of AVI. If we compare VPI with the conventional implementation of FQI that uses a fixed set of linear basis functions, we observe that FQI is the special case of VPI in which all atoms in the dictionary are used in the estimation. As VPI has a model selection step, its chosen estimator is not worse than FQI's (up to a small extra risk) and is possibly much better if the target is sparse in the dictionary. Moreover, extending the dictionary decreases the function approximation error with negligible effect on the model selection error. The same arguments apply to many other FQI versions that use a fixed data-independent set of basis functions and do not perform model selection.

4 Theoretical Analysis

In this section, we first study how the function approximation error propagates in VPI (Section 4.1) and then provide a finite-sample error upper bound as Theorem 3 in Section 4.2. All the proofs are in the appendices.

4.1 Propagation of Function Approximation Error

In this section, we present tools to upper bound the function approximation error at each iteration.

Definition 2 (Concentrability Coefficient of Function Approximation Error Propagation)**.** *(I) Let ν be a distribution over the state-action pairs, $(X, A) \sim \nu$, $\nu_{\mathcal{X}}$ be the marginal distribution of X, and $\pi_b(\cdot|\cdot)$ be the conditional probability of A given X. Further, let P be a transition probability kernel $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(X)$ and $P_{x,a} = P(\cdot|x, a)$. Define the concentrability coefficient of one-step transitions w.r.t. ν by*

$$C_{\nu\to\infty} = \left(\mathbb{E}\left[\sup_{(y,a')\in\mathcal{X}\times\mathcal{A}} \left|\frac{1}{\pi_b(a'|y)}\frac{\mathrm{d}P_{X,A}}{\mathrm{d}\nu_{\mathcal{X}}}(y)\right|\right]\right)^{\frac{1}{2}},$$

where $C_{\nu\to\infty} = \infty$ if $P_{x,a}$ is not absolutely continuous w.r.t. $\nu_{\mathcal{X}}$ for some $(x, a) \in \mathcal{X} \times \mathcal{A}$, or if $\pi_b(a'|y) = 0$ for some $(y, a') \in \mathcal{X} \times \mathcal{A}$. (II) Furthermore, for an optimal policy π^* and an integer $m \geq 0$, let $\nu(P^{\pi^*})^m \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ denote the future state-action distribution obtained after m-steps of following π^*. Define

$$c_\nu(m) \triangleq \left\|\frac{\mathrm{d}\left(\nu(P^{\pi^*})^m\right)}{\mathrm{d}\nu}\right\|_\infty.$$

If the future state-action distribution $\nu(P^{\pi^*})^m$ is not absolutely continuous w.r.t. ν, we let $c_\nu(m) = \infty$.

The constant $C_{\nu\to\infty}$ is large if after transition step, the future states can be highly concentrated at some states where the probability of taking some action a' is small or $\mathrm{d}\nu_{\mathcal{X}}$ is small. Hence, the name

"concentrability of one-step transitions". The definition of $C_{\nu\to\infty}$ is from Chapter 5 of Farahmand [27]. The constant $c_\nu(m)$ shows how much we deviate from ν whenever we follow an optimal policy π^*. It is notable that if ν happens to be the stationary distribution of the optimal policy π^* (e.g., the samples are generated by an optimal expert), $c_\nu(m) = 1$ for all $m \geq 0$.

We now provide the following result that upper bounds the error caused by using Q_k (which is the newly added atom to the dictionary) to approximate T^*Q_k. The proof is provided in Appendix B.

Lemma 1. *Let $(Q_i)_{i=0}^k \subset B(\mathcal{X}\times\mathcal{A}, Q_{max})$ be a Q_{max}-bounded sequence of measurable action-value functions. Define $\varepsilon_i \triangleq T^*Q_i - Q_{i+1}$ $(0 \leq i \leq k-1)$. Then,*

$$\|Q_k - T^*Q_k\|_\nu^2 \leq \frac{(1+\gamma C_{\nu\to\infty})^2}{1-\gamma}\left[\sum_{i=0}^{k-1}\gamma^{k-1-i}c_\nu(k-1-i)\,\|\varepsilon_i\|_\nu^2 + \gamma^k(2Q_{max})^2\right].$$

If there was no error at earlier iterations (i.e., $\|\varepsilon_i\|_\nu = 0$ for $0 \leq i \leq k-1$), the error $\|Q_k - T^*Q_k\|_\nu^2$ would be $O(\gamma^k Q_{\max}^2)$, which is decaying toward zero with a geometrical rate. This is similar to the behaviour of the exact VI, i.e., $\|T^*Q_k - Q_k\|_\infty \leq (1+\gamma)\gamma^k\|Q^* - Q_0\|_\infty$.

The following result is Theorem 5.3 of Farahmand [27]. For the sake of completeness, we provide the proof in Appendix B.

Theorem 2. *Let $(Q_k)_{k=0}^{k-1}$ be a sequence of state-action value functions and define $\varepsilon_i \triangleq T^*Q_i - Q_{i+1}$ $(0 \leq i \leq k)$. Let $\mathcal{F}^{|\mathcal{A}|} : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^{|\mathcal{A}|}$ be a subset of vector-valued measurable functions. Then,*

$$\inf_{Q'\in\mathcal{F}^{|\mathcal{A}|}}\|Q' - T^*Q_k\|_\nu \leq \inf_{Q'\in\mathcal{F}^{|\mathcal{A}|}}\left\|Q' - (T^*)^{(k+1)}Q_0\right\|_\nu + \sum_{i=0}^{k-1}\left(\gamma\, C_{\nu\to\infty}\right)^{k-i}\|\varepsilon_i\|_\nu.$$

This result quantifies the behaviour of the function approximation error and relates it to the function approximation error of approximating $(T^*)^{k+1}Q_0$ (which is a deterministic quantity depending only on the MDP itself, the function space $\mathcal{F}^{|\mathcal{A}|}$, and Q_0) and the errors of earlier iterations. This allows us to provide a tighter upper bound for the function approximation error compared to the so-called *inherent Bellman error* $\sup_{Q\in\mathcal{F}^{|\mathcal{A}|}}\inf_{Q'\in\mathcal{F}^{|\mathcal{A}|}}\|Q' - T^*Q\|_\nu$ introduced by Munos and Szepesvári [28], whenever the errors at previous iterations are small.

4.2 Finite Sample Error Bound for VPI

In this section, we provide an upper bound on the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$. This performance loss indicates the regret of following the policy π_K instead of an optimal policy when the initial state-action is distributed according to ρ. We define the following concentrability coefficients similar to Farahmand et al. [29].

Definition 3 (Expected Concentrability of the Future State-Action Distribution)**.** *Given $\rho, \nu \in \mathcal{M}(\mathcal{X}\times\mathcal{A})$, $m \geq 0$, and an arbitrary sequence of stationary policies $(\pi_m)_{m\geq 1}$, let $\rho P^{\pi_1}P^{\pi_2}\dots P^{\pi_m} \in \mathcal{M}(\mathcal{X}\times\mathcal{A})$ denote the future state-action distribution obtained after m transitions, when the first state-action pair is distributed according to ρ and then we follow the sequence of policies $(\pi_k)_{k=1}^m$. For integers $m_1, m_2 \geq 1$, policy π and the sequence of policies $\pi_1,\dots,\pi_k$ define the concentrability coefficients $c_{VI_1,\rho,\nu}(m_1,m_2;\pi) \triangleq \left(\mathbb{E}\left[\left|\frac{d\left(\rho(P^\pi)^{m_1}(P^{\pi^*})^{m_2}\right)}{d\nu}(X,A)\right|^2\right]\right)^{\frac{1}{2}}$ and $c_{VI_2,\rho,\nu}(m_1;\pi_1,\dots,\pi_k) \triangleq \left(\mathbb{E}\left[\left|\frac{d(\rho(P^{\pi_k})^{m_1}P^{\pi_{k-1}}P^{\pi_{k-2}}\dots P^{\pi_1})}{d\nu}(X,A)\right|^2\right]\right)^{\frac{1}{2}}$, where $(X,A)\sim\nu$. If the future state-action distribution $\rho(P^\pi)^{m_1}(P^{\pi^*})^{m_2}$ (similarly, if $\rho(P^{\pi_k})^{m_1}P^{\pi_{k-1}}P^{\pi_{k-2}}\cdots P^{\pi_1}$) is not absolutely continuous w.r.t. ν, we let $c_{VI_1,\rho,\nu}(m_1,m_2;\pi) = \infty$ (similarly, $c_{VI_2,\rho,\nu}(m_1;\pi_1,\dots,\pi_k) = \infty$).*

Assumption A1 We make the following assumptions:

- For all values of $0 \leq k \leq K-1$, the dataset used by VPI at each iteration is $\mathcal{D}_n^{(k)} = \{(X_i^{(k)}, A_i^{(k)}, R_i^{(k)}, X_i'^{(k)})\}_{i=1}^n$ with independent and identically distributed (i.i.d.)

samples $(X_i^{(k)}, A_i^{(k)}) \sim \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ and $X_i'^{(k)} \sim P(\cdot|X_i^{(k)}, A_i^{(k)})$ and $R_i^{(k)} \sim \mathcal{R}(\cdot, \cdot|X_i^{(k)}, A_i^{(k)})$ for $i = 1, 2, \ldots, n$.

- For $1 \leq k, k' \leq K-1$ and $k \neq k'$, the datasets $\mathcal{D}_n^{(k)}$ and $\mathcal{D}_n^{(k')}$ are independent.
- There exists a constant $Q_{\max} \geq 1$ such that for any $Q \in B(\mathcal{X} \times \mathcal{A}; Q_{\max})$, $|\hat{T}^* Q(X, A)| \leq Q_{\max}$ almost surely (a.s).
- For all $g \in \mathcal{B}_0$, $\|g\|_\infty \leq L < \infty$.
- The number of atoms m used from the dictionary $\mathcal{B}_0$ is $m = \lceil n^a \rceil$ for some finite $a > 0$. The number of link functions m' used at each iteration is at most m/K.
- At iteration k, each of the link functions $\{\sigma_i\}_{i=1}^{m'}$ maps $\beta_{Q_{\max}} Q_{k+1}$ and the dictionary $\mathcal{B}_k \bigcup \mathcal{B}_k'$ to an element of the space of vector-valued $Q_{\max}$-bounded measurable functions $\mathcal{X} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{A}|}$. At least one of the mappings returns $\beta_{Q_{\max}} Q_{k+1}$.

Most of these assumptions are mild and some of them can be relaxed. The i.i.d. assumption can be relaxed using the so called *independent block technique* [30], but it results in much more complicated proofs. We conjecture that the independence of datasets at different iterations might be relaxed as well under certain condition on the Bellman operator (cf. Section 4.2 of [28]). The condition on the number of atoms m and the number of link functions being polynomial in n are indeed very mild.

In order to compactly present our result, we define $a_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}$ for $0 \leq k < K$. Note that the behaviour of $a_k \propto \gamma^{K-k-1}$, so it gives more weight to later iterations. Also define $C_1(k) \triangleq \sum_{i=0}^{k-1} \gamma^{k-i} C_{\nu \to \infty}^{2(k-i)}$ $(k = 1, 2, \ldots)$ and $C_2 \triangleq \frac{(1+\gamma C_{\nu\to\infty})^2}{1-\gamma}$. For $0 \leq s \leq 1$, define

$$C_{\mathrm{VI},\rho,\nu}(K; s) = \left(\frac{1-\gamma}{2}\right)^2 \sup_{\pi_1', \ldots, \pi_K'} \sum_{k=0}^{K-1} a_k^{2(1-s)} \left[\sum_{m \geq 0} \gamma^m \left(c_{\mathrm{VI}_1,\rho,\nu}(m, K-k; \pi_K') + c_{\mathrm{VI}_2,\rho,\nu}(m+1; \pi_{k+1}', \ldots, \pi_K') \right) \right]^2,$$

where in the last definition the supremum is taken over all policies. The following theorem is the main theoretical result of this paper. Its proof is provided in Appendix D using tools developed in Appendices A, B, and C.

Theorem 3. *Consider the sequence $(Q_k)_{k=0}^K$ generated by VPI (Algorithm 1). Let Assumptions A1 hold. For any fixed $0 < \delta < 1$, recursively define the sequence $(b_i)_{i=0}^K$ as follows:*

$$b_0^2 \triangleq c_1 Q_{max}^3 \sqrt{\frac{\log\left(\frac{nK}{\delta}\right)}{n}} + 3 \inf_{Q' \in B_{Q_{max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))} \|Q' - T^* Q_0\|_\nu^2,$$

$$b_k^2 \triangleq c_2 Q_{max}^3 \sqrt{\frac{\log\left(\frac{nK}{\delta}\right)}{n}} + c_3 \min \left\{ \inf_{Q' \in B_{Q_{max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))} \left\|Q' - (T^*)^{k+1} Q_0\right\|_\nu^2 + C_1(k) \sum_{i=0}^{k-1} \gamma^{k-i} b_i^2, \right.$$
$$\left. C_2 \left(\sum_{i=0}^{k-1} \gamma^{k-1-i} c_\nu (k-1-i)\, b_i^2 + \gamma^k (2Q_{max})^2 \right) \right\}, \quad (k \geq 1)$$

for some $c_1, c_2, c_3 > 0$ that are only functions of Q_{max} and L. Then, for any $k = 0, 1, \ldots, K-1$, it holds that $\|Q_{k+1} - T^* Q_k\|_\nu^2 \leq b_k^2$, with probability at least $1 - \frac{k\delta}{K}$. Furthermore, define the discounted sum of errors as $\mathcal{E}(s) \triangleq \sum_{k=0}^{K-1} a_k^{2s} b_k$ (for $s \in [0, 1]$). Choose $\rho \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$. The ρ-weighted performance loss of π_K is upper bounded as

$$\|Q^* - Q^{\pi_K}\|_{1,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} \left[\inf_{s \in [0,1]} C_{VI,\rho,\nu}^{1/2}(K; s) \mathcal{E}^{1/2}(s) + 2\gamma^K Q_{max} \right],$$

with probability at least $1 - \delta$.

The value of b_k is a deterministic upper bound on the error $\|Q_{k+1} - T^*Q_k\|_\nu$ of each iteration of VPI. We would like b_k to be close to zero, because the second part of the theorem implies that $\|Q^* - Q^{\pi_K}\|_{1,\rho}$ would be small too. If we study b_k^2, we observe two main terms: The first term, which behaves as $\sqrt{\frac{\log(nK/\delta)}{n}}$, is the estimation error. The second term describes the function approximation error. For $k \geq 1$, it consists of two terms from which the minimum is selected. The first term inside $\min\{\cdot,\cdot\}$ describes the behaviour of the function approximation error when we only use the predefined dictionary $\mathcal{B}_{0,m}$ to approximate T^*Q_k (see Theorem 2). This means that this term ignores all learned atoms during the process.The second term describes the behaviour of the function approximation error when we only consider Q_k as the approximant of T^*Q_k (see Lemma 1). The error caused by this approximation depends on the error made in earlier iterations. The current analysis only considers the atom Q_k from the learned dictionary, but VPI may actually use other atoms to represent T^*Q_k. This might lead to much smaller function approximation errors. Hence, our analysis shows that in terms of function approximation error, our method is sound and superior to not increasing the size of the dictionary. However, revealing the full power of VPI remains as future work. Just as an example, if $\mathcal{B}_0$ is complete in $L_2(\nu)$, by letting $n, m \to \infty$ both the estimation error and function approximation error goes to zero and the method is consistent and converges to the optimal value function. Finally, notice that the effect of (b_k) on the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$ is quite the same for all AVI procedures (cf. [29]).

5 Conclusion

This work introduced VPI, an approximate value iteration algorithm that aims to find a close to optimal policy using a dictionary of atoms (or features). The VPI algorithm uses a modified Orthogonal Matching Pursuit that is equipped with a model selection procedure. This allows VPI to find a sparse representation of the value function in large, and potentially overcomplete, dictionaries. We theoretically analyzed VPI and provided a finite-sample error upper bound for it. The error bound shows the effect of the number of samples as well as the function approximation properties of the predefined dictionary, and the effect of learned atoms.

This paper is a step forward to better understanding how overcomplete dictionaries and sparsity can effectively be used in the RL/Planning context. A more complete theory describing the effect of adding atoms to the dictionary remains to be established. We are also planning to study VPI's empirical performance, and comparing with other feature construction methods. We note that our main focus was on the statistical properties of the algorithm, not on computational efficiency; optimizing computation speed will be an interesting topic for future investigation.

A Statistical Properties of Orthogonal Matching Pursuit

In this section, we first describe OMP and report a result on its approximation theoretic property. We then focus on the regression setting and show that OMP can be used as a regression procedure. The results of this section closely follow the paper by Barron et al. [25] with the difference that 1) Theorem 6 holds in high probability instead of in expectation, and 2) as far as we know, Lemma 7 is new.

Here we briefly describe the OMP algorithm. Consider a dictionary $\mathcal{B} = \{g_1, g_2, \dots\}$ in some Hilbert space $\mathcal{H}$. Assume that $\|g\| = 1$ for all $g \in \mathcal{B}$. The OMP algorithms approximates a function $f \in \mathcal{H}$ as follows: Let $f_0 = 0$. For each $i = 1, 2, \dots$, define the residual $r_{i-1} = f - f_{i-1}$. Let $g_i = \text{Argmax}_{g \in \mathcal{B}} |\langle r_{i-1}, g \rangle|$. Define Π_i as the orthogonal projection onto $\text{span}(g_i, \dots, g_i)$. We then have $f_i = \Pi_i f$. The procedure continues.

The following result quantifies the function approximation error of the OMP algorithm.

Lemma 4 (Theorem 2.3 by Barron et al. [25])**.** *For all $f \in \mathcal{H}$ and any $h \in \mathcal{L}_1(\mathcal{B}; \|\cdot\|)$, the error of the OMP algorithm satisfies*

$$\|f_i - f\|^2 \leq \|h - f\|^2 + \frac{4\,\|h\|^2_{\mathcal{L}_1(\mathcal{B};\|\cdot\|)}}{i}. \qquad (i = 1, 2, \dots)$$

We now turn to the statistical setting and the problem of regression. Suppose we are given a dataset $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of i.i.d. samples with $(X_i, Y_i) \sim \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ with $\mathcal{X} \subset \mathbb{R}^d$

and $Y \subset \mathbb{R}$. Denote the regression function by $m(x) = \mathbb{E}[Y|X = x]$. We assume that $|Y| \leq L$ almost surely. Moreover, without loss of generality we assume that $L \geq 1$.

We now consider an OMP procedure that uses dictionary $\mathcal{B}$ of size $|\mathcal{B}|$ and provides an estimate $\hat{m}$ of the regression function with guaranteed $L_2(\nu_{\mathcal{X}})$ error upper bound for $\|m - \hat{m}\|_{2,\nu_X}$. The conventional OMP algorithms adds a new element from the dictionary to the representation at each iteration. In the approximation theoretical framework, adding more elements to the representation is desirable, but in the learning scenario it may lead to overfitting. Thus, we use a complexity regularization-based model selection procedure to find the proper size of representation. The algorithm is as follows:

1. Apply OMP using the dictionary $\mathcal{B}$ and the empirical inner product $\langle f, g \rangle_{\mathcal{D}_n}$ to obtain a sequence of estimates $(\hat{m}_k)_{k\geq 0}$.
2. Define
$$k^* \leftarrow \operatorname*{argmin}_{k\geq 0} \{\|Y_i - \beta_L \hat{m}_k\|^2_{\mathcal{D}_n} + \mathrm{Pen}_n(k)\}, \tag{2}$$
in which $\mathrm{Pen}_n(k) \triangleq c_1 \frac{k \log(n|\mathcal{B}|)}{n}$ with some $c_1 > 0$, which is a function of only L.
3. Return $\hat{m} = \beta_L \hat{m}_{k^*}$.

We define some functions spaces that shall be used in the statistical analysis of OMP. For any dictionary $\mathcal{B}$, let $\Lambda \subset \mathcal{B}$ and define $G_\Lambda \triangleq \mathrm{span}\{g : g \in \Lambda\}$. Let $\beta_L G_\Lambda$ be the set of L-truncated functions from G_Λ. Define $\mathcal{F}_k \triangleq \bigcup_{\Lambda \subset D; |\Lambda| \leq k} \beta_L G_\Lambda$, which is the space of all L-truncated functions that can be written as a linear combination of at most k atoms from $\mathcal{B}$. Notice that $\mathcal{F}_k$ is the function space to which $\beta_L \hat{m}_k$ belongs, i.e., $\beta_L \hat{m}_k \in \mathcal{F}_k$. The following lemma, which is borrowed from Barron et al. [25], upper bounds the covering number of $\mathcal{F}_k$.

Lemma 5 (Covering number of $\mathcal{F}_k$ – Lemma 3.3 of Barron et al. [25])**.** *Suppose $\mathcal{X} \subset \mathbb{R}^d$ and $|\mathcal{B}|$, the number of elements in the dictionary, is finite. For any probability measure $\mu \in \mathcal{M}(\mathcal{X})$, for any $0 < \varepsilon < L/4$, we have*

$$\mathcal{N}(\varepsilon, \mathcal{F}_k, \|\cdot\|_{1,\mu}) \leq 3|\mathcal{B}|^k \left(\frac{2\,eL}{\varepsilon} \log\left(\frac{3\,eL}{\varepsilon}\right)\right)^{k+1}.$$

The following theorem is a slight modification of Theorem 3.1 of Barron et al. [25]. The difference is that this current result holds with high probability instead of in expectation.

Theorem 6 (OMP)**.** *Consider the OMP procedure described above with a finite $|\mathcal{B}|$ and k^* selected according to (2). There exist $c_1 > 0$, depending only on L, and constants $c_2, c_3 > 0$ such that for the choice of $\mathrm{Pen}_n(k) = \frac{c_1 k \log(n|\mathcal{B}|)}{n}$, for any $0 < \delta < 1$ and for all $k = 1, 2, 3, \ldots$ and $h \in \mathrm{span}(\mathcal{B})$, the estimator $\hat{m}$ satisfies*

$$\|\hat{m} - m\|^2_{\nu_{\mathcal{X}}} \leq \frac{8\,\|h\|^2_{\mathcal{L}_1(\mathcal{B};\mathcal{D}_n)}}{k} + 3\,\|h - m\|^2_{\nu_{\mathcal{X}}} + \frac{c_2\,k\log(n|\mathcal{B}|) + c_3\,L^4 \log(\frac{1}{\delta})}{n},$$

with probability at least $1 - \delta$.

Proof. In this proof, $\mathcal{D}_n = \{X_i, Y_i\} \sim \nu$ are i.i.d. samples and $(X, Y) \sim \nu$ and is independent from $\mathcal{D}_n$. We decompose the error $\|\hat{m} - m\|^2_{2,\nu_{\mathcal{X}}}$ as

$$\int_{\mathcal{X}} |\hat{m}(x) - m(x)|^2 \mathrm{d}\nu_{\mathcal{X}}(x) = \mathbb{E}\left[|\hat{m}(X) - Y|^2 | \mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right] = T_{1,n} + T_{2,n}$$

with

$$\begin{aligned} T_{1,n} &\triangleq \mathbb{E}\left[|\hat{m}(X) - Y|^2 | \mathcal{D}_n\right] - \mathbb{E}\left[|m(X) - Y|^2\right] \\ &\quad - 2\left(\frac{1}{n}\sum_{i=1}^n |Y_i - \hat{m}(X_i)|^2 - |Y_i - m(X_i)|^2 + \mathrm{Pen}_n(k^*)\right), \\ T_{2,n} &\triangleq 2\left(\frac{1}{n}\sum_{i=1}^n |Y_i - \hat{m}(X_i)|^2 - |Y_i - m(X_i)|^2 + \mathrm{Pen}_n(k^*)\right). \end{aligned}$$

Since $\hat{m} = \beta_L \hat{m}_{k^*}$ is the minimizer of (2) and $|Y| \leq L$ almost surely (so truncation does not increase the error), for all $k = 1, 2, 3, \ldots$, we have

$$\begin{aligned} T_{2,n} &= 2\left(\frac{1}{n}\sum_{i=1}^{n} |Y_i - \beta_L \hat{m}_{k^*}(X_i)|^2 - |Y_i - m(X_i)|^2 + \mathrm{Pen}_n(k)\right) \\ &\leq 2\left(\frac{1}{n}\sum_{i=1}^{n} |Y_i - \beta_L \hat{m}_{k}(X_i)|^2 - |Y_i - m(X_i)|^2 + \mathrm{Pen}_n(k)\right) \\ &\leq 2\left(\frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{m}_{k}(X_i)|^2 - |Y_i - m(X_i)|^2 + \mathrm{Pen}_n(k)\right). \end{aligned}$$

Let $h \in L_2(\nu_{\mathcal{X}})$. Decompose the R.H.S. of the above inequality to $T_{3,n} + T_{4,n}$ defined as

$$\begin{aligned} \frac{1}{2}T_{3,n} &= \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{m}_{k}(X_i)|^2 - |Y_i - h(X_i)|^2, \\ \frac{1}{2}T_{4,n} &= \frac{1}{n}\sum_{i=1}^{n} |Y_i - h(X_i)|^2 - |Y_i - m(X_i)|^2 + \mathrm{Pen}_n(k). \end{aligned}$$

We now upper bound $T_{1,n}$, $T_{3,n}$, and $T_{4,n}$.

Lemma 4 indicates that

$$\frac{1}{2}T_{3,n} \leq \frac{4\|h\|^2_{\mathcal{L}_1(\mathcal{B};\mathcal{D}_n)}}{k}. \tag{3}$$

We now turn to proving a high probability upper bound for $T_{4,n}$. Define $W_i = |h(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2$, so we have $\frac{1}{2}T_{4,n} = \frac{1}{n}\sum_{i=1}^{n} W_i + \mathrm{Pen}_n(k)$. It can be seen that $\mathbb{E}[W_i] = \mathbb{E}\left[|h(X_i) - m(X_i)|^2\right]$. We want to show that $T_{4,n}$ is not much larger than its expectation. We can use Bernstein inequality (e.g., see Lemma A.2 of Györfi et al. [31]) to provide such a guarantee with a high probability.

It is easy to see that $|Z| \leq 4L^2$ a.s. Moreover, $\mathrm{Var}[W_i] \leq \mathbb{E}\left[|W_i|^2\right] = \mathbb{E}\left[\left(h(X_i) - 2Y_i + m(X_i)\right)^2 \left(h(X_i) - m(X_i)\right)^2\right] \leq (4L)^2\mathbb{E}\left[|h(X_i) - m(X_i)|^2\right] = (4L)^2\mathbb{E}[W_i]$, i.e., the variance of W_i is "controlled" by its expectation. This is the key to obtain a fast rate using Bernstein inequality. For $t > 0$, we have the following sequence of inequalities:

$$\begin{aligned} &\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} W_i - \mathbb{E}[W_1] \geq \frac{1}{2}t + \frac{1}{2}\mathbb{E}[W_1]\right\} \leq \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} W_i - \mathbb{E}[W_1] \geq \frac{1}{2}t + \frac{1}{32L^2}\mathrm{Var}[W_1]\right\} \\ &\leq \exp\left(-\frac{n\left(\frac{1}{2}t + \frac{1}{32L^2}\mathrm{Var}[W_1]\right)^2}{2\mathrm{Var}[W_1] + \frac{2}{3}(4L^2)\left(\frac{1}{2}t + \frac{1}{32L^2}\mathrm{Var}[W_1]\right)}\right) \\ &\leq \exp\left(-\frac{n\left(\frac{1}{2}t + \frac{1}{32L^2}\mathrm{Var}[W_1]\right)^2}{64L^2\left(\frac{1}{2}t + \frac{1}{32L^2}\mathrm{Var}[W_1]\right) + \frac{2}{3}(4L^2)\left(\frac{1}{2}t + \frac{1}{32L^2}\mathrm{Var}[W_1]\right)}\right) \\ &= \exp\left(-\frac{n\left(\frac{1}{2}t + \frac{1}{32L^2}\mathrm{Var}[W_1]\right)}{(64 + \frac{8}{3})L^2}\right) \leq \exp\left(-\frac{n\,t}{(128 + \frac{16}{3})L^2}\right). \end{aligned}$$

So for any $0 < \delta_1 < 1$, we have

$$T_{4,n} \leq 3\|h - m\|^2_{\nu_{\mathcal{X}}} + 2\mathrm{Pen}_n(k) + \frac{800L^2}{3n}\ln\left(\frac{1}{\delta_1}\right), \tag{4}$$

with probability at least $1 - \delta_1$.

We use Theorem 10 alongside Lemma 5 to upper bound $T_{1,n}$. Since k^* is random, we use the union bound over all $k \geq 1$. For $t \geq 1/t$, we have

$$\begin{aligned}
\mathbb{P}\left\{T_{1,n} > t\right\} \leq \sum_{k\geq 1} \mathbb{P}\Bigg\{ \exists f \in \mathcal{F}_k : \; & \mathbb{E}\left[|f(X)-Y|^2 | \mathcal{D}_n\right] - \mathbb{E}\left[|m(X)-Y|^2\right] - \\
& \left(\frac{1}{n}\sum_{i=1}^n |f(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right) \geq \\
& \frac{1}{2}\left(t + 2\mathrm{Pen}_n(k) + \mathbb{E}\left[|f(X)-Y|^2|\mathcal{D}_n\right] - \mathbb{E}\left[|m(X)-Y|^2\right]\right) \Bigg\} \\
\leq \sum_{k\geq 1} 14 \sup_{x_{1:n}} \mathcal{N}\left(\frac{t}{40L}, \mathcal{F}_k, \|\cdot\|_{1,x_{1:n}}\right) & \exp\left(-\frac{(\frac{1}{2})^2(1-\frac{1}{2})(\frac{t}{2}+2\mathrm{Pen}_n(k))n}{214(1+\frac{1}{2})L^4}\right) \\
\leq 14\exp\left(-\frac{c_1 n t}{2}\right) & \sum_{k\geq 1} 3|\mathcal{B}|^k \left(2eLn\log(3eLn)\right)^{k+1} \exp\left(-2c_1 \mathrm{Pen}_n(k) n\right),
\end{aligned}$$

in which $c_1 = 1/(568L^4)$. The first inequality is the application of the union bound on k and reorganizing terms in $T_{1,n}$. The second inequality is the application of Theorem 10 with the choice of $\varepsilon = 1/2$, $\beta = t/2$ and $\alpha = t/2 + 2\mathrm{Pen}_n(k)$. In the third inequality, we used $t \geq 1/n$ alongside Lemma 5, and separated terms that are a function of k and $\exp(-c_1 nt/2)$.

With the choice of $\mathrm{Pen}_n(k) = \frac{c_2 k \log(n|\mathcal{B}|)}{n}$ (with c_2 being a function of only L), we have $-2c_1\mathrm{Pen}_n(k)n + k\log(3|\mathcal{B}|) + (k+1)\log(2eLn\log(3eLn)) \leq -2\log k$, which entails that the above summation is smaller than the constant $\pi^2/6$. As a result, $\mathbb{P}\left\{T_{1,n} > t\right\} \leq \frac{7\pi^2}{3}\exp(-\frac{c_1 nt}{2})$. Thus, there exists a constant $c_3 > 0$ such that for any fixed $0 < \delta_2 < 1/2$, it holds that

$$T_{1,n} \leq \frac{2\log(\frac{7\pi^2}{3\delta_2})}{c_1\, n} + \frac{1}{n} \leq \frac{c_3 L^4 \log(\frac{1}{\delta_2})}{n}, \tag{5}$$

with probability at least $1-\delta_2$. Let $\delta_1 = \delta_2 = \delta/2$, and invoke (3), (4), and (5) to obtain the desired result. □

The following lemma upper bounds the $\mathcal{L}_1$-norm of a function w.r.t. the empirical measure $\mathcal{D}_n \sim \nu$ by its $\mathcal{L}_1$-norm w.r.t. ν.

Lemma 7. *Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ with $X_i \overset{i.i.d.}{\sim} \nu$. Consider a finite dictionary $\mathcal{B} = \{g_1, \ldots, g_{|\mathcal{B}|}\}$ of normalized atoms ($\|g\|_\nu = 1$ for all $g \in \mathcal{B}$). Assume that the atoms are bounded, i.e., $\|g\|_\infty \leq L < \infty$ for all $g \in \mathcal{B}$. Define the empirical dictionary $\hat{\mathcal{B}} = \left\{\frac{g_1}{\|g_1\|_{\mathcal{D}_n}}, \ldots, \frac{g_{|\mathcal{B}|}}{\|g_{|\mathcal{B}|}\|_{\mathcal{D}_n}}\right\}$. Consider a fixed function $f : \mathcal{X} \to \mathbb{R}$ that belongs to $\mathcal{L}_1(\mathcal{B};\nu)$. For any fixed $0 < \delta < 1$, it holds that*

$$\|f\|_{\mathcal{L}_1(\hat{\mathcal{B}};\mathcal{D}_n)} \leq \left(1 + L\sqrt{\frac{8\log(|\mathcal{B}|/\delta)}{3n}}\right)\|f\|_{\mathcal{L}_1(\mathcal{B};\nu)},$$

with probability at least $1-\delta$.

Proof. Because $f \in \mathcal{L}_1(\mathcal{B};\nu)$, we can represent it as $f = \sum_{i=1}^{|\mathcal{B}|} c_i g_i$. Likewise, since $\hat{\mathcal{B}} = \{\hat{g}_1, \ldots, \hat{g}_{|\mathcal{B}|}\}$ with $\hat{g}_i = \frac{g_i}{\|g_i\|_{\mathcal{D}_n}}$, we have $f = \sum_{i=1}^{|\mathcal{B}|} \hat{c}_i \hat{g}_i$ with the choice of $\hat{c}_i = c_i \|g_i\|_{\mathcal{D}_n}$ (note that $|\hat{\mathcal{B}}| = |\mathcal{B}|$). We now uniformly upper bound

$$\left(\frac{\hat{c}_i}{c_i}\right)^2 = \frac{\|g_i\|^2_{\mathcal{D}_n}}{\|g_i\|^2_\nu\,[=1]} = \frac{\frac{1}{n}\sum_{j=1}^n g_i^2(X_i)}{\mathbb{E}\left[g_i^2(X)\right]}.$$

Use the union bound and Bernstein inequality (e.g., Lemma A.2 of Györfi et al. [31]) to get that for $t > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left\{\max_{1\le i\le|\mathcal{B}|}\frac{\frac{1}{n}\sum_{j=1}^{n}g_i^2(X_i)}{\mathbb{E}\left[g_i^2(X)\right]}>2+t\right\}\\
\le&|\mathcal{B}|\max_{1\le i\le|\mathcal{B}|}\exp\left(-\frac{n\left(\mathbb{E}\left[g_i^2(X)\right](1+t)\right)^2}{2\mathrm{Var}\left[g_i^2(X)\right]+\frac{2}{3}L^2\left(\mathbb{E}\left[g_i^2(X)\right](1+t)\right)}\right)\\
\le&|\mathcal{B}|\max_{1\le i\le|\mathcal{B}|}\exp\left(-\frac{n\left(\mathbb{E}\left[g_i^2(X)\right](1+t)\right)^2}{2L^2\mathbb{E}\left[g_i^2(X)\right]+\frac{2}{3}L^2\left(\mathbb{E}\left[g_i^2(X)\right](1+t)\right)}\right)\\
\le&|\mathcal{B}|\max_{1\le i\le|\mathcal{B}|}\exp\left(-\frac{n\left(\mathbb{E}\left[g_i^2(X)\right](1+t)\right)^2}{(2+\frac{2}{3})L^2\left(\mathbb{E}\left[g_i^2(X)\right](1+t)\right)}\right)\le|\mathcal{B}|\max_{1\le i\le|\mathcal{B}|}\exp\left(-\frac{3n(1+t)}{8L^2}\right).
\end{aligned}$$

Here we used $\|g_i\|_\infty \le L$ in the first inequality, $\mathrm{Var}\left[g_i^2(X)\right] \le L^2$ in the second inequality, and $1+t>1$ in the third inequality. Finally, we benefitted from the fact that the atoms $g_i \in \mathcal{B}$ are normalized according to $\|\cdot\|_\nu$, i.e., $\mathbb{E}\left[g_i^2(X)\right]=1$.

Thus, for any fixed $0<\delta<1$, we get $\max_{1\le i\le|\mathcal{B}|}(\frac{\hat{c}_i}{c_i})^2 \le 2+t \le 1+\frac{8L^2\log(|\mathcal{B}|/\delta)}{3n}$, with probability at least $1-\delta$. On the event that this inequality holds, we have

$$\|f\|_{\mathcal{L}_1(\hat{\mathcal{B}};\mathcal{D}_n)}=\inf\left\{\sum_{i=1}^{|\mathcal{B}|}|\tilde{c}_i|\,:\,f=\sum_{i=1}^{|\mathcal{B}|}\tilde{c}_i\hat{g}_i\right\}\le\sum_{i=1}^{|\mathcal{B}|}|\hat{c}_i|\le\left(1+L\sqrt{\frac{8\log(|\mathcal{B}|/\delta)}{3n}}\right)\sum_{i=1}^{|\mathcal{B}|}|c_i|.$$

Take infimum over all possible admissible $(c_i)_{i=1}^{\mathcal{B}}$ to get the desired result. □

B Propagation of Function Approximation Error: Proofs for Section 4.1

We first present a lemma which shows that the Bellman optimality operator is Lipschitz when viewed as an operator of the Banach space of action-value functions equipped with $\|\cdot\|_\nu$. This result is Lemma 5.11 of Farahmand [27]. Afterwards, we present the proof of Lemma 1 and Theorem 2. The former result is new to this paper, while the latter was proven as Theorem 5.3 of Farahmand [27].

Lemma 8. *For any given $Q_1, Q_2 \in \mathcal{F}^{|\mathcal{A}|}$, we have $\|T^*Q_1 - T^*Q_2\|_\nu \le \gamma\, C_{\nu\to\infty}\, \|Q_1 - Q_2\|_\nu$.*

Proof. Jensen's inequality, followed by the application of the elementary inequality $|\max_\theta f(\theta) - \max_\theta g(\theta)|^2 \le \max_\theta |f(\theta)-g(\theta)|^2$ gives

$$\begin{aligned}
\|T^*Q_1 - T^*Q_2\|_{2,\nu}^2 &= \gamma^2\int_{\mathcal{X}\times\mathcal{A}} d\nu(x,a)\left|\int_{\mathcal{X}} dP_{x,a}(y)\left(\max_{a'\in\mathcal{A}}Q_1(y,a')-\max_{a'\in\mathcal{A}}Q_2(y,a')\right)\right|^2\\
&\le \gamma^2\int_{\mathcal{X}\times\mathcal{A}} d\nu(x,a)\int_{\mathcal{X}} dP_{x,a}(y)\left|\max_{a'\in\mathcal{A}}Q_1(y,a')-\max_{a'\in\mathcal{A}}Q_2(y,a')\right|^2\\
&\le \gamma^2\int_{\mathcal{X}\times\mathcal{A}} d\nu(x,a)\int_{\mathcal{X}} dP_{x,a}(y)\max_{a'\in\mathcal{A}}\left|Q_1(y,a')-Q_2(y,a')\right|^2.
\end{aligned}$$

Inequality $\max_{a' \in \mathcal{A}} |Q(y,a')|^2 \leq \max_{a'' \in \mathcal{A}} [\frac{1}{\pi_b(a''|y)}] \sum_{a' \in \mathcal{A}} \pi_b(a'|y) |Q(y,a')|^2$ together with a change of measure argument gives

$$
\begin{aligned}
&\|T^*Q_1 - T^*Q_2\|_{2,\nu}^2 \\
&\leq \gamma^2 \int_{\mathcal{X}\times\mathcal{A}} d\nu(x,a) \int_{\mathcal{X}} \sum_{a' \in \mathcal{A}} dP_{x,a}(y) \max_{a'' \in \mathcal{A}} \left\{ \frac{1}{\pi_b(a''|y)} \right\} \pi_b(a'|y) \left| Q_1(y,a') - Q_2(y,a') \right|^2 \\
&\leq \gamma^2 \int_{\mathcal{X}\times\mathcal{A}} d\nu(x,a) \int_{\mathcal{X}} \sum_{a' \in \mathcal{A}} \sup_{(z,a'') \in \mathcal{X}\times\mathcal{A}} \left[\frac{1}{\pi_b(a''|z)} \frac{dP_{x,a}}{d\nu_{\mathcal{X}}}(z) \right] d\nu_{\mathcal{X}}(y) \pi_b(a'|y) \left| Q_1(y,a') - Q_2(y,a') \right|^2 \\
&= \gamma^2 \left[\int_{\mathcal{X}\times\mathcal{A}} d\nu(x,a) \sup_{(z,a'') \in \mathcal{X}\times\mathcal{A}} \left[\frac{1}{\pi_b(a''|z)} \frac{dP_{x,a}}{d\nu_{\mathcal{X}}}(z) \right] \right] \left[\int_{\mathcal{X}\times\mathcal{A}} d\nu(y,a') \left| Q_1(y,a') - Q_2(y,a') \right|^2 \right] \\
&= \gamma^2 C_{\nu\to\infty}^2 \|Q_1 - Q_2\|_\nu^2 .
\end{aligned}
$$

where in the second to last equation we exploited that $\pi_b \otimes \nu_{\mathcal{X}} = \nu$. □

Proof of Lemma 1. By the triangle inequality, Lemma 8 and the fact that $T^*Q^* = Q^*$, we have

$$
\begin{aligned}
\|Q_k - T^*Q_k\|_\nu &= \|Q_k - T^*Q^* + T^*Q^* - T^*Q_k\|_\nu \leq \|Q_k - T^*Q^*\|_\nu + \|T^*Q^* - T^*Q_k\|_\nu \\
&\leq \|Q_k - Q^*\|_\nu + (\gamma C_{\nu\to\infty}) \|Q^* - Q_k\|_\nu = (1 + \gamma C_{\nu\to\infty}) \|Q_k - Q^*\|_\nu . \quad (6)
\end{aligned}
$$

It is shown by Munos [32, Equation 4.2] that

$$
Q^* - Q_k \leq \sum_{i=0}^{k-1} \gamma^{k-1-i} (P^{\pi^*})^{k-1-i} \varepsilon_i + \gamma^k (P^{\pi^*})^k (Q^* - Q_0).
$$

We now calculate the $L_2(\nu)$-norm of the LHS. Define $N = \sum_{i=0}^k \gamma^i$, and notice that the sequence $(\frac{\gamma^i}{N})_{i=0}^k$ is a probability distribution. Square both sides, use the convexity of $x \mapsto |x|^2$ to apply Jensen's inequality twice, once considering the sequence $(\frac{\gamma^i}{N})_{i=0}^k$ and once considering the stochastic operators $((P^{\pi^*})^i)_{i=0}^{k-1}$, to get

$$
\begin{aligned}
|Q^* - Q_k|^2 &\leq N^2 \left| \sum_{i=0}^{k-1} \frac{\gamma^{k-1-i}}{N} (P^{\pi^*})^{k-1-i} \varepsilon_i + \frac{\gamma^k}{N} (P^{\pi^*})^k (Q^* - Q_0) \right|^2 \\
&\leq N \left[\sum_{i=0}^{k-1} \gamma^{k-1-i} (P^{\pi^*})^{k-1-i} |\varepsilon_i|^2 + \gamma^k (P^{\pi^*})^k |Q^* - Q_0|^2 \right].
\end{aligned}
$$

We apply ν to both sides. By Radon-Nikodym theorem, $\nu (P^{\pi^*})^{k-1-i} |\varepsilon_i|^2 = \int \nu(dx) (P^{\pi^*})^{k-1-i} (dy|x) \, |\varepsilon_i(y)|^2 = \int \frac{\mathrm{d}(\nu (P^{\pi^*})^{k-1-i})}{\mathrm{d}\nu}(y) \nu(dy) |\varepsilon_i(y)|^2 \leq \int \| \frac{\mathrm{d}(\nu (P^{\pi^*})^{k-1-i})}{\mathrm{d}\nu} \|_\infty \nu(\mathrm{d}y) |\varepsilon_i(y)|^2 = c_\nu(k-1-i) \|\varepsilon_i\|_\nu^2$ (cf. Definition 2). This, in addition to the fact that $\|Q^* - Q_0\|_\infty \leq 2Q_{\max}$, leads to

$$
\|Q^* - Q_k\|_\nu^2 = \nu |Q^* - Q_k|^2 \leq N \left[\sum_{i=0}^{k-1} \gamma^{k-1-i} c_\nu(k-1-i) \nu |\varepsilon_i|^2 + \gamma^k |2Q_{\max}|^2 \right]. \quad (7)
$$

Combining inequalities (6) and (7) and noticing that $N \leq \frac{1}{1-\gamma}$ finish the proof. □

Proof of Theorem 2. Let $Q_0, \ldots, Q_{K-1}$ be action-value functions, $\varepsilon_k = T^*Q_k - Q_{k+1}$, $b_k = \|\varepsilon_k\|_\nu$. Our goal is to bound $\inf_{Q' \in \mathcal{F}^{|\mathcal{A}|}} \|Q' - T^*Q_k\|_\nu$. For this, pick any $Q' \in \mathcal{F}^{|\mathcal{A}|}$. Then, by the triangle inequality,

$$
\|Q' - T^*Q_k\|_\nu \leq \|Q' - (T^*)^{k+1} Q_0\|_\nu + \|(T^*)^{k+1} Q_0 - T^*Q_k\|_\nu,
$$

therefore, it remains to upper bound $\|(T^*)^{k+1}Q_0 - T^*Q_k\|_\nu$. Since by Lemma 8, T^* is $L \triangleq \gamma C_{\nu\to\infty}$-Lipschitz w.r.t. $\|\cdot\|_\nu$, we have $\|(T^*)^{k+1}Q_0 - T^*Q_k\|_\nu \leq L \|(T^*)^k Q_0 - Q_k\|_\nu$. Using the

definition of ε_k, $\|(T^*)^k Q_0 - Q_k\|_\nu = \|(T^*)^k Q_0 - (T^* Q_{k-1} - \varepsilon_{k-1})\|_\nu \le \|(T^*)^k Q_0 - T^* Q_{k-1}\| + \|\varepsilon_{k-1}\|_\nu \le L\|(T^*)^{k-1} Q_0 - Q_{k-1}\| + \|\varepsilon_{k-1}\|_\nu$. Finishing the recursion gives

$$\|(T^*)^k Q_0 - Q_k\|_\nu \le \|\varepsilon_{k-1}\|_\nu + L\|\varepsilon_{k-2}\|_\nu + \ldots + L^{k-1}\|\varepsilon_0\|_\nu.$$

Combining the inequalities obtained so far, we get

$$\|Q' - T^* Q_k\|_\nu \le \|Q' - (T^*)^k Q_0\|_\nu + \sum_{i=0}^{k-1} L^{k-i}\|\varepsilon_i\|_\nu,$$

from which the desired statement follows immediately. □

C Error Propagation for AVI Algorithms

Let $Q_0, Q_1, \ldots, Q_K \subset B(\mathcal{X} \times \mathcal{A}, Q_{\max})$ be a sequence of action-value functions, perhaps generated by some approximate value iteration procedure that approximates $T^* Q_k$ by Q_{k+1}. Let the error at iteration k be

$$\varepsilon_k = T^* Q_k - Q_{k+1}. \tag{8}$$

Further, let π_K be the policy greedy w.r.t. Q_K and $p \ge 1$.

In this section, we use a slight modification of a result by Farahmand et al. [29] to relate the performance loss $\|Q^* - Q^{\pi_K}\|_{p,\rho}$ to the ν-weighted L_{2p}-norms of the error sequence $(\varepsilon_k)_{k=0}^{K-1}$.[5] This performance loss indicates the regret of following policy π_K instead of an optimal policy when the initial state-action is distributed according to ρ.

To relate these two measures that are entangled through the MDP, we use the concentrability coefficients defined as Definition 3. The concentrability coefficients are used in a change of measure argument. Due to the dynamics of MDP and AVI, this change depends not only on ν and ρ, but also on the transition kernels P^π and P^{π^*}, see e.g., Munos [32], Farahmand et al. [29].

In order to compactly present our results, we define $a_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}$ for $0 \le k < K$. Also for $0 \le s \le 1$, define the discounted sum of errors as $\mathcal{E}(\varepsilon_0, \ldots, \varepsilon_{K-1}; s) = \sum_{k=0}^{K-1} a_k^{2s} \|\varepsilon_k\|_{2p,\nu}^{2p}$ and

$$C_{\mathrm{VI},\rho,\nu}(K; s) = \left(\frac{1-\gamma}{2}\right)^2 \sup_{\pi'_1,\ldots,\pi'_K} \sum_{k=0}^{K-1} a_k^{2(1-s)} \left[\sum_{m \ge 0} \gamma^m \left(c_{\mathrm{VI}_1,\rho,\nu}(m, K-k; \pi'_K) + c_{\mathrm{VI}_2,\rho,\nu}(m+1; \pi'_{k+1}, \ldots, \pi'_K)\right)\right]^2,$$

where in the last definition the supremum is taken over all policies. We now state the following error propagation result.

Theorem 9 (Error Propagation for AVI – Farahmand et al. [29])**.** *Let $p \ge 1$ be a real number, K be a positive integer, and $Q_{max} \le \frac{R_{max}}{1-\gamma}$. Then, for any sequence $(Q_k)_{k=0}^K \subset B(\mathcal{X} \times \mathcal{A}, Q_{max})$, and the corresponding sequence $(\varepsilon_k)_{k=0}^{K-1}$ defined in (8), we have*

$$\|Q^* - Q^{\pi_K}\|_{p,\rho} \le \frac{2\gamma}{(1-\gamma)^2} \left[\inf_{s \in [0,1]} C_{VI,\rho,\nu}^{\frac{1}{2p}}(K; s) \mathcal{E}^{\frac{1}{2p}}(\varepsilon_0, \ldots, \varepsilon_{K-1}; r) + 2\gamma^{\frac{K}{p}} Q_{max}\right].$$

The significance of this result and its comparison to the previous work such as Munos [32] is provided by Farahmand et al. [29] in detail.

D Proof of Theorem 3

In the proof, $c_1, c_2, \ldots$ are constants whose values may change from line to line – unless specified otherwise.

[5] The modification is that as opposed to [29] who define ε_k as $T^* V_k - V_{k+1}$ and provide an upper bound on $\|V^* - V^{\pi_k}\|_{p,\rho}$, here the errors are defined according to the action-value functions.

Proof of Theorem 3. Fix $0 < \delta < 1$ and let $\delta_0 = \delta_1 = \cdots = \delta_{K-1} = \delta/K$. As before, we denote $\varepsilon_i = T^*Q_i - Q_{i+1}$ for $0 \le i \le K-1$.

For any $k = 0, 1, 2, \ldots$ and with our choice of $\mathrm{Pen}_n(k)$, Theorem 6 states that there exists constant $c_1 > 0$ for all $s = 1, 2, 3, \ldots$ and any $Q' \in \mathrm{span}(\hat{\mathcal{B}}_{0,m} \cup \hat{\mathcal{B}}'_k)$ such that

$$\|\varepsilon_k\|_\nu^2 = \|Q_{k+1} - T^*Q_k\|_\nu^2 \le \frac{8\,\|Q'\|^2_{\mathcal{L}_1(\hat{\mathcal{B}}_{0,m}\cup\hat{\mathcal{B}}'_k;\mathcal{D}_n^{(k)})}}{s} + 3\,\|Q' - T^*Q_k\|_\nu^2 + \frac{c_1 Q_{\max}^4 s \log(\frac{|\hat{\mathcal{B}}_{0,m}\cup\hat{\mathcal{B}}'_k|n}{\delta_k/2})}{n}, \tag{9}$$

with probability at least $1 - \delta_k/2$. We denote the event that the inequality holds by $\mathcal{E}_k^{(1)}$ (for each k).

Note that $\|Q'\|_{\mathcal{L}_1(\hat{\mathcal{B}}_{0,m}\cup\hat{\mathcal{B}}'_k;\mathcal{D}_n^{(k)})}$ is random, so we upper bound it by a deterministic quantity using Lemma 7. By assumption, all $g \in \mathcal{B}_0$ are bounded by L. Moreover, for any Q' we have $\|Q'\|_{\mathcal{L}_1(\hat{\mathcal{B}}_{0,m}\cup\hat{\mathcal{B}}'_k;\mathcal{D}_n^{(k)})} \le \|Q'\|_{\mathcal{L}_1(\hat{\mathcal{B}}_{0,m};\mathcal{D}_n^{(k)})}$, i.e., increasing the size of the dictionary does not increase a function's $\mathcal{L}_1$ norm. Therefore,

$$\|Q'\|^2_{\mathcal{L}_1(\hat{\mathcal{B}}_{0,m}\cup\hat{\mathcal{B}}'_k;\mathcal{D}_n^{(k)})} \le 2\left(1 + L^2\frac{8\log\left(\frac{|\hat{\mathcal{B}}_{0,m}|}{\delta_k/2}\right)}{3n}\right)\|Q'\|^2_{\mathcal{L}_1(\hat{\mathcal{B}}_{0,m};\nu)}, \tag{10}$$

with probability at least $1 - \delta_k/2$. We denote the event that this inequality holds by $\mathcal{E}_k^{(2)}$ (for each k).

We now turn to choose a Q' for the right hand side (9). The goal is to make sure the first two terms behave reasonably. The case of $k = 0$ is simpler than $k \ge 1$ because for $k = 0$ the dictionary $\mathcal{B}_{0,m}$ is fixed (ignoring the normalization by $\|\cdot\|_{\mathcal{D}_n^{(0)}}$), but in later iterations we have to deal with the random dictionary $\mathcal{B}_{0,m} \cup \mathcal{B}'_k$. So we start with $k = 0$.

For $k = 0$, we choose Q' to belong to the ball $B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))$ and use inequalities (9) and (10) to get that on the event $\mathcal{E}_0^{(1)} \cup \mathcal{E}_0^{(2)}$ we have

$$\|Q_1 - T^*Q_0\|_\nu^2 \le \frac{16Q_{\max}^2}{s}\left(1 + \frac{8L^2\log(\frac{2m}{\delta_0})}{3n}\right) + 3\inf_{Q'\in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))}\|Q' - T^*Q_0\|_\nu^2 + \frac{c_2 Q_{\max}^4 s\log(\frac{n}{\delta_0})}{n},$$

for some $c_2 > 0$. Here we used the fact that since $m = \lceil n^a \rceil$ ($a > 0$), the size of the dictionary $|\hat{\mathcal{B}}_{0,m} \cup \hat{\mathcal{B}}'_k| = |\hat{\mathcal{B}}_{0,m}|$ is equal to $\lceil n^a \rceil$ to simplify the RHS. By $s = \frac{c_3}{Q_{\max}}\sqrt{\frac{n}{\log(n/\delta_0)}}$, we get that on the event $\mathcal{E}_0^{(1)} \cup \mathcal{E}_0^{(2)}$, it holds that

$$\|\varepsilon_0\|_\nu^2 = \|Q_1 - T^*Q_0\|_\nu^2 \le b_0^2 \triangleq c_4 Q_{\max}^3\sqrt{\frac{\log(n/\delta_0)}{n}} + 3\inf_{Q'\in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))}\|Q' - T^*Q_0\|_\nu^2 \tag{11}$$

for some constant $c_4 > 0$. Note that the term $\frac{128Q_{\max}^3L^2}{c_3}\left(\frac{\log(n/\delta_0)}{n}\right)^{3/2}$ has been dominated by the slower terms.

For $k \ge 1$, we choose Q' to belong to the ball $B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m} \cup \mathcal{B}'_k;\nu))$, and then upper bound the function approximation error $\|Q' - T^*Q_k\|_\nu^2$ when Q' is confined to the specified ball. In cases when Q_k is a good approximation to T^*Q_k, one may choose $Q' = Q_k$. Otherwise, one may choose a function from the predefined dictionary $\mathcal{B}_{0,m}$. Thus, we have

$$\inf_{Q'\in B_{Q_{\max}}\left(\mathcal{L}_1(\mathcal{B}_{0,m}\cup\mathcal{B}'_k;\nu)\right)}\|Q' - T^*Q_k\|^2 \le \min\Big\{\inf_{Q'\in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))}\|Q' - T^*Q_k\|^2, \inf_{Q'\in B_{Q_{\max}}\left(\mathcal{L}_1(\mathcal{B}'_k;\nu)\right)}\|Q' - T^*Q_k\|^2\Big\}. \tag{12}$$

We use Theorem 2 to upper bound $\inf_{Q' \in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))} \|Q' - T^*Q_k\|^2$ as

$$\inf_{Q' \in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))} \|Q' - T^*Q_k\|^2 \leq \left[\inf_{Q' \in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))} \left\|Q' - (T^*)^{k+1}Q_0\right\|_\nu + \sum_{i=0}^{k-1} (\gamma C_{\nu\to\infty})^{k-i} \|T^*Q_i - Q_{i+1}\|_\nu \right]^2$$
$$\leq 2 \inf_{Q' \in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))} \left\|Q' - (T^*)^{k+1}Q_0\right\|_\nu + \left(\sum_{i=0}^{k-1} \gamma^{k-i} C_{\nu\to\infty}^{2(k-i)}\right) \left(\sum_{i=0}^{k-1} \gamma^{k-i} \|\varepsilon_i\|_\nu^2\right), \tag{13}$$

where we used the Cauchy-Schwarz inequality in the last step.

To upper bound the second term inside the $\min\{\cdot,\cdot\}$ in (12), note that if we choose $Q' = Q_k = \frac{Q_k}{\|Q_k\|_\nu} \|Q_k\|_\nu$, since $\frac{Q_k}{\|Q_k\|_\nu}$ belongs to the dictionary $\mathcal{B}'_k$ (which is normalized according to $\|\cdot\|_\nu$ and not $\|\cdot\|_{\mathcal{D}_n^{(k)}}$), according to the definition of the $\mathcal{L}_1(\mathcal{B}'_k;\nu)$-norm, we have $\|Q'\|_{\mathcal{L}_1(\mathcal{B}'_k;\nu)} \leq \|Q_k\|_\nu$. Consequently, we upper bound $\|Q_k\|_\nu$ by its supremum norm to get $\|Q'\|_{\mathcal{L}_1(\mathcal{B}'_k;\nu)} \leq Q_{\max}$. Thus, Q' belongs to $B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}'_k;\nu))$, as desired. Because of the truncation, all elements of the sequence $(Q_i)_{i=0}^k$ are bounded by $Q_{\max}$, so we can use Lemma 1 to get

$$\inf_{Q' \in B_{Q_{\max}}\left(\mathcal{L}_1(\mathcal{B}'_k;\nu)\right)} \|Q' - T^*Q_k\|^2 \leq \|Q_k - T^*Q_k\|_\nu^2 \leq$$
$$\frac{(1+\gamma C_{\nu\to\infty})^2}{1-\gamma} \left[\sum_{i=0}^{k-1} \gamma^{k-1-i} c_\nu(k-1-i) \|\varepsilon_i\|_\nu^2 + \gamma^k (2Q_{\max})^2\right]. \tag{14}$$

Substitute (13) and (14) in (12) to provide an upper bound on the function approximation error $\inf_{Q' \in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))} \|Q' - T^*Q_k\|^2$, in which Q' is restricted to have $\mathcal{L}_1(\mathcal{B}_{m,0} \cup \mathcal{B}'_k;\nu)$-norm of at most $Q_{\max}$. By Lemma 7, and similar to inequality (10), we know that the empirical $\mathcal{L}_1(\mathcal{B}_{m,0} \cup \mathcal{B}'_k; \mathcal{D}_n^{(k)})$ is not much larger (only by a factor of $1 + L\sqrt{\frac{\log(2m/\delta_k)}{3n}}$) than the one w.r.t. ν either, which is less than $Q_{\max}$. We can now plug all these results into (9) and set $s_k = \frac{c_3}{Q_{\max}} \sqrt{\frac{n}{\log(n/\delta_k)}}$ to obtain that on the event $\bigcup_{k=0,\dots,K-1} \left(\mathcal{E}_k^{(1)} \cup \mathcal{E}_k^{(2)}\right)$, which holds with probability at least $1-\delta$, for any $k \geq 1$, we have

$$\|Q_{k+1} - T^*Q_k\|_\nu^2 \leq b_k^2 \triangleq c_4 Q_{\max}^3 \sqrt{\frac{\log\left(\frac{nK}{\delta}\right)}{n}}$$
$$+ c_5 \min\left\{ \inf_{Q' \in B_{Q_{\max}}(\mathcal{L}_1(\mathcal{B}_{0,m};\nu))} \left\|Q' - (T^*)^{k+1}Q_0\right\|_\nu^2 + C_1(k) \sum_{i=0}^{k-1} \gamma^{k-i} b_i^2, \right.$$
$$\left. C_2 \left(\sum_{i=0}^{k-1} \gamma^{k-1-i} c_\nu(k-1-i)\, b_i^2 + \gamma^k (2Q_{\max})^2\right) \right\},$$

with $C_1(k) \triangleq \sum_{i=0}^{k-1} \gamma^{k-i} C_{\nu\to\infty}^{2(k-i)}$ and $C_2 \triangleq \frac{(1+\gamma C_{\nu\to\infty})^2}{1-\gamma}$.

The second part of the theorem is the direct application of Theorem 9 with the choice of $p = 1$. □

E Auxiliary Results

We use the following relative deviation inequality (or the modulus of continuity of the empirical process) in the proof of Theorem 6. This result is Theorem 11.4 of Györfi et al. [31], which is based on Theorem 3 by Lee et al. [33]. In this result, we have the same regression setup as in Appendix A.

Theorem 10 (Relative Deviation Inequality – Theorem 11.4 of Györfi et al. [31]). *Assume $|Y| \le L$ a.s. and $L \ge 1$. Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ with $\|f\|_\infty \le L$. For each $n \ge 1$, we have*

$$\mathbb{P}\Bigg\{\exists f \in \mathcal{F} :\ \mathbb{E}\left[|f(X)-Y|^2\right] - \mathbb{E}\left[|m(X)-Y|^2\right] - \frac{1}{n}\sum_{i=1}^{n}\left[|f(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right] \\ \ge \varepsilon\left(\alpha+\beta+\mathbb{E}\left[|f(X)-Y|^2\right] - \mathbb{E}\left[|m(X)-Y|^2\right]\right)\Bigg\} \\ \le 14 \sup_{x_{1:n}} \mathcal{N}\left(\frac{\beta\varepsilon}{20L}, \mathcal{F}, \|\cdot\|_{1,x_{1:n}}\right) \exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{214(1+\varepsilon)L^4}\right),$$

where $\alpha, \beta > 0$ and $0 < \varepsilon \le \frac{1}{2}$.

F Markov Decision Processes

In this section, we summarize the definitions required to describe MDPs and RL. The reader is referred to Bertsekas and Tsitsiklis [34], Sutton and Barto [35], Buşoniu et al. [36], Szepesvári [24] for the background information on MDP and RL.

For a space Ω, with σ-algebra σ_Ω, $\mathcal{M}(\Omega)$ denotes the set of all probability measures over σ_Ω. $B(\Omega)$ denotes the space of bounded measurable functions w.r.t. σ_Ω and $B(\Omega, L)$ denotes the subset of $B(\Omega)$ with bound $0 < L < \infty$.

A *finite-action discounted MDP* is a 5-tuple $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$, where $\mathcal{X}$ is a measurable state space, $\mathcal{A}$ is a finite set of actions, $P : \mathcal{X}\times\mathcal{A} \to \mathcal{M}(\mathcal{X})$ is the transition probability kernel, $\mathcal{R} : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0,1)$ is a discount factor. Let $r(x,a) = \mathbb{E}[\mathcal{R}(\cdot|x,a)]$, and assume that r is uniformly bounded by $R_{\max}$. A measurable mapping $\pi : \mathcal{X} \to \mathcal{A}$ is called a deterministic Markov stationary policy, or just a *policy* for short. Following a policy π means that at each time step, $A_t = \pi(X_t)$.

A policy π induces two transition probability kernels $P^\pi : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ and $P^\pi : \mathcal{X}\times\mathcal{A} \to \mathcal{M}(\mathcal{X}\times\mathcal{A})$. For a measurable subset S of $\mathcal{X}$ and a measurable subset S' of $\mathcal{X}\times\mathcal{A}$, we define $(P^\pi)(S|x) \triangleq \int P(dy|x,\pi(x))\mathbb{I}_{\{y\in S\}}$ and $(P^\pi)(S'|x,a) \triangleq \int P(dy|x,a)\mathbb{I}_{\{(y,\pi(y))\in S'\}}$. The m-step transition probability kernel $(P^\pi)^m : \mathcal{X}\times\mathcal{A} \to \mathcal{M}(\mathcal{X}\times\mathcal{A})$ for $m = 2, 3, \cdots$ are inductively defined as $(P^\pi)^m(S'|x,a) \triangleq \int_{\mathcal{X}} P(dy|x,a)(P^\pi)^{m-1}(B|y,\pi(y))$ (similarly for $(P^\pi)^m : \mathcal{X} \to \mathcal{M}(\mathcal{X})$).

Given probability transition kernels $P : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ and $P : \mathcal{X}\times\mathcal{A} \to \mathcal{M}(\mathcal{X}\times\mathcal{A})$, define the right-linear operators $P\cdot : B(\mathcal{X}) \to B(\mathcal{X})$ and $P\cdot : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X}\times\mathcal{A})$ by $(PV)(x) \triangleq \int_{\mathcal{X}} P(dy|x)V(y)$ and $(PQ)(x,a) \triangleq \int_{\mathcal{X}\times\mathcal{A}} P(dy,da'|x,a)Q(y,a')$. In words, $(PV)(x)$ is the expected value of V after the transition P. For a probability measure $\rho \in \mathcal{M}(\mathcal{X})$ and a measurable subset S of $\mathcal{X}$, define the left-linear operator $\cdot P : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ by $(\rho P)(S) = \int \rho(dx)P(dy|x)\mathbb{I}_{\{y\in A\}}$. In words, ρP represents the distribution over states when the initial state distribution is ρ and we follow P for a single step. Similarly, for a probability measure $\rho \in \mathcal{M}(\mathcal{X}\times\mathcal{A})$ and a measurable subset B of $\mathcal{X}\times\mathcal{A}$, define the left-linear operator $\cdot P : \mathcal{M}(\mathcal{X}\times\mathcal{A}) \to \mathcal{M}(\mathcal{X}\times\mathcal{A})$ by $(\rho P)(B) = \int \rho(dx,da)P(dy,da'|x,a)\mathbb{I}_{\{(y,a')\in B\}}$. A typical choice of P is $(P^\pi)^m$ for $m \ge 1$.

The value function V^π and the action-value function Q^π of a policy π are defined as follows: Let $(R_t; t \ge 1)$ be the sequence of rewards when the Markov chain is started from state X_1 (or state-action (X_1, A_1) for Q^π) drawn from a positive probability distribution over $\mathcal{X}$ ($\mathcal{X}\times\mathcal{A}$) and the agent follows the policy π. Then $V^\pi(x) \triangleq \mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_t \,\middle|\, X_1 = x\right]$ and $Q^\pi(x,a) \triangleq \mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_t \,\middle|\, X_1 = x, A_1 = a\right]$. The value of V^π and Q^π are uniformly bounded by $Q_{\max} = R_{\max}/(1-\gamma)$, independent of the choice of π.

The *optimal value* and *optimal action-value* functions are defined as $V^*(x) = \sup_\pi V^\pi(x)$ for all $x \in \mathcal{X}$ and $Q^*(x,a) = \sup_\pi Q^\pi(x,a)$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$. A policy π^* is *optimal* if $V^{\pi^*} = V^*$. A policy π is *greedy* w.r.t. an action-value function Q, denoted $\pi = \hat{\pi}(\cdot;Q)$, if

$\pi(x) = \text{argmax}_{a \in \mathcal{A}} Q(x,a)$ holds for all $x \in \mathcal{X}$ (if there exist multiple maximizers, one of them is chosen in an arbitrary deterministic manner). Note that a greedy policy w.r.t. the optimal action-value function Q^* is an optimal policy.

For a fixed policy π, the Bellman operators $T^\pi : B(\mathcal{X}) \to B(\mathcal{X})$ and $T^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ (for the action-value functions) are defined as $(T^\pi V)(x) \triangleq r(x, \pi(x)) + \gamma \int_{\mathcal{X}} V(y) P(dy|x, \pi(x))$ and $(T^\pi Q)(x,a) \triangleq r(x,a) + \gamma \int_{\mathcal{X}} Q(y, \pi(y)) P(dy|x,a)$. The fixed point of the Bellman operator is the (action-)value function of the policy π, i.e., $T^\pi Q^\pi = Q^\pi$ and $T^\pi V^\pi = V^\pi$. Similarly, the Bellman optimality operators $T^* : B(\mathcal{X}) \to B(\mathcal{X})$ and $T^* : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ (for the action-value functions) are defined as $(T^* V)(x) \triangleq \max_a \left\{ r(x,a) + \gamma \int_{\mathbb{R} \times \mathcal{X}} V(y) P(dr, dy|x,a) \right\}$ and $(T^* Q)(x,a) \triangleq r(x,a) + \gamma \int_{\mathbb{R} \times \mathcal{X}} \max_{a'} Q(y,a') P(dr, dy|x,a)$. Again, these operators enjoy a fixed-point property similar to that of the Bellman operators: $T^* Q^* = Q^*$ and $T^* V^* = V^*$.

References

[1] Pieter Abeel, Adam Coates, Morgan Quigley, and Andrew Y. Ng. An application of reinforcement learning to aerobatic helicopter flight. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1–8. MIT Press, Cambridge, MA, 2007.

[2] Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 673–680. ACM, 2006.

[3] Joelle Pineau, Marc G. Bellemare, A. John Rush, Adrian Ghizaru, and Susan A. Murphy. Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence*, 88, supplement 2:S52-S60, 2007.

[4] David Silver, Richard S. Sutton, and Martin Müller. Reinforcement learning of local shape in the game of go. In Manuela M. Veloso, editor, *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1053–1058, 2007.

[5] Marek Petrik. An analysis of Laplacian methods for value function approximation in MDPs. In *International Joint Conference on Artificial Intelligence (ICAJ)*, pages 2574–2579, 2007.

[6] Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in markov decision processes. *Journal of Machine Learning Research*, 8:2169–2231, 2007. 1

[7] Ronald Parr, Christopher Painter-Wakefield, Lihong Li, and Michael Littman. Analyzing feature generation for value-function approximation. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 737 – 744, New York, NY, USA, 2007. ACM. 1

[8] Alborz Geramifard, Finale Doshi, Joshua Redding, Nicholas Roy, and Jonathan How. Online discovery of feature dependencies. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 881–888, New York, NY, USA, June 2011. ACM.

[9] Tobias Jung and Daniel Polani. Least squares SVM for least squares TD learning. In *In Proc. 17th European Conference on Artificial Intelligence*, pages 499–503, 2006.

[10] Xin Xu, Dewen Hu, and Xicheng Lu. Kernel-based least squares policy iteration for reinforcement learning. *IEEE Trans. on Neural Networks*, 18:973–992, 2007.

[11] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS - 21)*, pages 441–448. MIT Press, 2009. 1

[12] Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted Q-iteration for planning in continuous-space Markovian Decision Problems. In *Proceedings of American Control Conference (ACC)*, pages 725–730, June 2009. 1, 5

[13] Gavin Taylor and Ronald Parr. Kernelized value function approximation for reinforcement learning. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1017–1024, New York, NY, USA, 2009. ACM. 1

[14] J. Zico Kolter and Andrew Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 521–528. ACM, 2009. 1

[15] Jeff Johns, Christopher Painter-Wakefield, and Ronald Parr. Linear complementarity for regularized policy evaluation and improvement. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS - 23)*, pages 1009–1017. 2010. 1

[16] Mohammad Ghavamzadeh, Alessandro Lazaric, Rémi Munos, and Matthew Hoffman. Finite-sample analysis of lasso-TD. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1177–1184, New York, NY, USA, June 2011. ACM. ISBN 978-1-4503-0619-5. 1

[17] Stéphane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[18] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993. 1

[19] Geoffrey M. Davis, Stéphane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Journal of Constructive Approximation*, 13:57–98, 1997. 1

[20] Vladimir N. Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3:33–107, 2008.

[21] Jeff Johns. *Basis Construction and Utilization for Markov Decision Processes using Graphs*. PhD thesis, University of Massachusetts Amherst, 2010. 1, 2

[22] Christopher Painter-Wakefield and Ronald Parr. Greedy algorithms for sparse reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML) (Accepted)*, 2012. 1

[23] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005. 2, 5

[24] Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan Claypool Publishers, 2010. 3, 18

[25] Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, and Ronald A. Devore. Approximation and learning by greedy algorithms. *The Annals of Statistics*, 36(1):64–94, 2008. 3, 4, 5, 9, 10

[26] Ronald A. Devore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[27] Amir-massoud Farahmand. *Regularization in Reinforcement Learning*. PhD thesis, University of Alberta, 2011. 7, 13

[28] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008. 7, 8

[29] Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS - 23)*, pages 568–576. 2010. 7, 9, 15

[30] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, January 1994. 8

[31] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Verlag, New York, 2002.

[32] Rémi Munos. Performance bounds in L_p norm for approximate value iteration. *SIAM Journal on Control and Optimization*, pages 541–561, 2007.

[33] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, November 1998.

[34] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)*. Athena Scientific, 1996.

[35] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, 1998.

[36] Lucian Buşoniu, Robert Babuška, Bart De Schutter, and Damien Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, 2010.