Journal of Machine Learning Research 17 (2016) 1-66 Submitted 1/13; Revised 12/15; Published 9/16

Regularized Policy Iteration with Nonparametric Function Spaces

Amir-massoud Farahmand FARAHMAND@MERL.COM
Mitsubishi Electric Research Laboratories (MERL)
201 Broadway, 8th Floor
Cambridge, MA 02139, USA

Mohammad Ghavamzadeh GHAVAMZA@ADOBE.COM
Adobe Research
321 Park Avenue
San Jose, CA 95110, USA

Csaba Szepesvári SZEPESVA@UALBERTA.CA
Department of Computing Science
University of Alberta
Edmonton, AB, T6G 2E8, Canada

Shie Mannor SHIE@EE.TECHNION.AC.IL
Department of Electrical Engineering
The Technion
Haifa 32000, Israel

Editor: Peter Auer

Abstract

We study two regularization-based approximate policy iteration algorithms, namely REG-LSPI and REG-BRM, to solve reinforcement learning and planning problems in discounted Markov Decision Processes with large state and finite action spaces. The core of these algorithms are the regularized extensions of the Least-Squares Temporal Difference (LSTD) learning and Bellman Residual Minimization (BRM), which are used in the algorithms' policy evaluation steps. Regularization provides a convenient way to control the complexity of the function space to which the estimated value function belongs and as a result enables us to work with rich nonparametric function spaces. We derive efficient implementations of our methods when the function space is a reproducing kernel Hilbert space. We analyze the statistical properties of REG-LSPI and provide an upper bound on the policy evaluation error and the performance loss of the policy returned by this method. Our bound shows the dependence of the loss on the number of samples, the capacity of the function space, and some intrinsic properties of the underlying Markov Decision Process. The dependence of the policy evaluation bound on the number of samples is minimax optimal. This is the first work that provides such a strong guarantee for a nonparametric approximate policy iteration algorithm.[1]

Keywords: reinforcement learning, approximate policy iteration, regularization, nonparametric method, finite-sample analysis

1. This work is an extension of the NIPS 2008 conference paper by Farahmand et al. (2009b).

1. Introduction

We study the approximate policy iteration (API) approach to find a close to optimal policy in a Markov Decision Process (MDP), either in a reinforcement learning (RL) or in a planning scenario. The basis of API, which is explained in Section 3, is the policy iteration algorithm that iteratively evaluates a policy (i.e., finding the value function of the policy—the *policy evaluation* step) and then improves it (i.e., computing the *greedy* policy with respect to (w.r.t.) the recently obtained value function—the *policy improvement* step). When the state space is large (e.g., a subset of $\mathbb{R}^d$ or a finite state space that has too many states to be exactly represented), the policy evaluation step cannot be performed exactly, and as a result the use of function approximation is inevitable. The appropriate choice of the function approximation method, however, is far from trivial. The best choice is problem-dependent and it also depends on the number of samples in the input data.

In this paper we propose a *nonparametric regularization*-based approach to API. This approach provides a flexible and easy way to implement the policy evaluation step of API. The advantage of nonparametric methods over parametric methods is that they are flexible: Whereas a parametric model, which has a fixed and finite parameterization, limits the range of functions that can be represented, irrespective of the number of samples, the nonparametric models avoid such undue restrictions by increasing the power of the function approximation as necessary. Moreover, the regularization-based approach to nonparametrics is elegant and powerful: It has a simple algorithmic form and the estimator achieves minimax optimal rates in a number of scenarios. Further discussion of and specific results about nonparametric methods, particularly in the supervised learning scenario, can be found in the books by Györfi et al. (2002) and Wasserman (2007).

The nonparametric approaches to solve RL/Planning problems have received some attention in the RL community. For instance, Petrik (2007); Mahadevan and Maggioni (2007); Parr et al. (2007); Mahadevan and Liu (2010); Geramifard et al. (2011); Farahmand and Precup (2012); Böhmer et al. (2013) and Milani Fard et al. (2013) suggest methods to generate data-dependent basis functions, to be used in general linear models. Ormoneit and Sen (2002) use smoothing kernel-based estimate of the model and then use value iteration to find the value function. Barreto et al. (2011, 2012) benefit from "stochastic factorization trick" to provide computationally efficient ways to scale up the approach of Ormoneit and Sen (2002). In the context of approximate value iteration, Ernst et al. (2005) consider growing ensembles of trees to approximate the value function. In addition, there have been some works where regularization methods have been applied to the RL/Planning problems, e.g., Engel et al. (2005); Jung and Polani (2006); Loth et al. (2007); Farahmand et al. (2009a,b); Taylor and Parr (2009); Kolter and Ng (2009); Johns et al. (2010); Ghavamzadeh et al. (2011); Farahmand (2011b); Ávila Pires and Szepesvári (2012); Hoffman et al. (2012); Geist and Scherrer (2012). Nevertheless, most of these papers are algorithmic results and do not analyze the statistical properties of these methods (the exceptions are Farahmand et al. 2009a,b; Farahmand 2011b; Ghavamzadeh et al. 2011; Ávila Pires and Szepesvári 2012). We compare these methods with ours in more detail in Sections 5.3.1 and 6.

It is worth mentioning that one might use a regularized estimator alongside a feature generation approach to control the complexity of function space induced by the features. An approach alternative to regularization for controlling the complexity of a function space is to

use greedy algorithms, such as Matching Pursuit (Mallat and Zhang, 1993) and Orthogonal Matching Pursuit (Pati et al., 1993), to select features from a large set of features. Greedy algorithms have recently been developed for the value function estimation by Johns (2010); Painter-Wakefield and Parr (2012); Farahmand and Precup (2012); Geramifard et al. (2013). We do not discuss these methods any further.

1.1 Contributions

The algorithmic contribution of this work is to introduce two regularization-based nonparametric approximate policy iteration algorithms, namely *Regularized Least-Squares Policy Improvement (REG-LSPI)* and *Regularized Bellman Residual Minimization (REG-BRM)*. These are flexible methods that, upon the proper selection of their parameters, are sample efficient. Each of REG-BRM and REG-LSPI is formulated as two coupled regularized optimization problems (Section 4). As we argue in Section 4.1, having a regularized objective in both optimization problems is necessary for rich nonparametric function spaces. Despite the unusual coupled formulation of the underlying optimization problems, we prove that the solutions can be computed in a closed-form when the estimated action-value function belongs to the family of *reproducing kernel Hilbert spaces (RKHS)* (Section 4.2).

The theoretical contribution of this work (Section 5) is to analyze the statistical properties of REG-LSPI and to provide upper bounds on the policy evaluation error and the performance difference between the optimal policy and the policy returned by this method (Theorem 14). The result demonstrates the dependence of the bounds on the number of samples, the capacity of the function space to which the estimated action-value function belongs, and some intrinsic properties of the MDP. It turns out that the dependence of the policy evaluation error bound on the number of samples is minimax optimal. This paper, alongside its conference (Farahmand et al., 2009b) and the dissertation (Farahmand, 2011b) versions, is the first work that analyzes a nonparametric regularized API algorithm and provides such a strong guarantee for it.

2. Background and Notation

In the first part of this section, we provide a brief summary of some of the concepts and definitions from the theory of MDPs and RL (Section 2.1). For more information, the reader is referred to Bertsekas and Shreve (1978); Bertsekas and Tsitsiklis (1996); Sutton and Barto (1998); Szepesvári (2010). In addition to this background on MDPs, we introduce the notations we use to denote function spaces and their corresponding norms (Section 2.2) as well as the considered learning problem (Section 2.3).

2.1 Markov Decision Processes

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

Definition 1 *A* finite-action discounted MDP *is a 4-tuple* $(\mathcal{X}, \mathcal{A}, P, \gamma)$*, where* $\mathcal{X}$ *is a measurable state space,* $\mathcal{A}$ *is a finite set of actions,* $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathbb{R} \times \mathcal{X})$ *is a mapping with domain* $\mathcal{X} \times \mathcal{A}$*, and* $0 \leq \gamma < 1$ *is a discount factor. Mapping* P *evaluated*

at $(x,a) \in \mathcal{X} \times \mathcal{A}$ gives a distribution over $\mathbb{R} \times \mathcal{X}$, which we shall denote by $P(\cdot,\cdot|x,a)$. We denote the marginals of P by the overloaded symbol $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$ defined as $P(\cdot|x,a) = P_{x,a}(\cdot) = \int_{\mathbb{R}} P(dr,\cdot|x,a)$ (transition probability kernel) and $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathbb{R})$ defined as $\mathcal{R}(\cdot|x,a) = \int_{\mathcal{X}} P(\cdot,dy|x,a)$ (reward distribution).

An MDP together with an initial distribution P_1 of states encode the laws governing the temporal evolution of a discrete-time stochastic process controlled by an agent as follows: The controlled process starts at time $t = 1$ with random initial state $X_1 \sim P_1$ (here and in what follows $X \sim Q$ denotes that the random variable X is drawn from distribution Q). At stage t, action $A_t \in \mathcal{A}$ is selected by the agent controlling the process. In response, the pair (R_t, X_{t+1}) is drawn from $P(\cdot,\cdot|X_t, A_t)$, i.e., $(R_t, X_{t+1}) \sim P(\cdot,\cdot|X_t, A_t)$, where, R_t is the reward that the agent receives at time t and X_{t+1} is the state at time $t+1$. The process then repeats with the agent selecting action A_{t+1}, etc.

In general, the agent can use all past states, actions, and rewards in deciding about its current action. However, for our purposes it will suffice to consider action-selection procedures, or policies, that select an action deterministically and time-invariantly solely based on the current state:

Definition 2 (Deterministic Markov Stationary Policy) *A measurable mapping $\pi : \mathcal{X} \to \mathcal{A}$ is called a* deterministic Markov stationary policy*, or just* policy *in short. Following a policy π in an MDP means that at each time step t it holds that $A_t = \pi(X_t)$.*

Policy π induces the transition probability kernels $P^\pi : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$ defined as follows: For a measurable subset C of $\mathcal{X} \times \mathcal{A}$, let $(P^\pi)(C|x,a) \triangleq \int P(dy|x,a)\mathbb{I}_{\{(y,\pi(y))\in C\}}$. The m-step transition probability kernels $(P^\pi)^m : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$ for $m = 2, 3, \cdots$ are defined inductively by $(P^\pi)^m(C|x,a) \triangleq \int_{\mathcal{X}} P(dy|x,a)(P^\pi)^{m-1}(C|y,\pi(y))$. Also given a probability transition kernel $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$, we define the right-linear operator $P\cdot : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ by $(PQ)(x,a) \triangleq \int_{\mathcal{X}\times\mathcal{A}} P(dy,da'|x,a)Q(y,a')$. For a probability measure $\rho \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ and a measurable subset C of $\mathcal{X} \times \mathcal{A}$, we define the left-linear operators $\cdot P : \mathcal{M}(\mathcal{X} \times \mathcal{A}) \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$ by $(\rho P)(C) = \int \rho(dx,da)P(dy,da'|x,a)\mathbb{I}_{\{(y,a')\in C\}}$.

To study MDPs, two auxiliary functions are of central importance: the *value* and the *action-value functions* of a policy π.

Definition 3 (Value Functions) *For a policy π, the value function V^π and the action-value function Q^π are defined as follows: Let $(R_t; t \geq 1)$ be the sequence of rewards when the Markov chain is started from a state X_1 (or state-action (X_1, A_1) for the action-value function) drawn from a positive probability distribution over $\mathcal{X}$ (or $\mathcal{X} \times \mathcal{A}$) and the agent follows 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 $Q^\pi(x,a) \triangleq \mathbb{E}\left[\sum_{t=1}^\infty \gamma^{t-1}R_t \,\middle|\, X_1 = x, A_1 = a\right]$.*

It is easy to see that for any policy π, if the magnitude of the immediate expected reward $r(x,a) = \int r\, P(dr,dy|x,a)$ is uniformly bounded by $R_{\max}$, then the functions V^π and Q^π are bounded by $V_{\max} = Q_{\max} = R_{\max}/(1-\gamma)$, independent of the choice of π.

For a discounted MDP, we define the *optimal value* and *optimal action-value* functions by $V^*(x) = \sup_\pi V^\pi(x)$ for all states $x \in \mathcal{X}$ and $Q^*(x,a) = \sup_\pi Q^\pi(x,a)$ for all state-actions $(x,a) \in \mathcal{X} \times \mathcal{A}$. We say that a policy π^* is *optimal* if it achieves the best values in every

state, i.e., if $V^{\pi^*} = V^*$. We say that a policy π is *greedy* w.r.t. an action-value function Q if $\pi(x) = \operatorname{argmax}_{a \in \mathcal{A}} Q(x,a)$ for all $x \in \mathcal{X}$. We define function $\hat{\pi}(x; Q) \triangleq \operatorname{argmax}_{a \in \mathcal{A}} Q(x,a)$ (for all $x \in \mathcal{X}$) that returns a greedy policy of an action-value function Q (If there exist multiple maximizers, a maximizer is chosen in an arbitrary deterministic manner). Greedy policies are important because a greedy policy w.r.t. the optimal action-value function Q^* is an optimal policy. Hence, knowing Q^* is sufficient for behaving optimally (cf. Proposition 4.3 of Bertsekas and Shreve 1978).[2]

Definition 4 (Bellman Operators) *For a policy π, the Bellman operators $T^\pi : B(\mathcal{X}) \to B(\mathcal{X})$ (for value functions) and $T^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ (for action-value functions) are defined as*

$$(T^\pi V)(x) \triangleq r(x, \pi(x)) + \gamma \int_{\mathcal{X}} V(y) P(dy|x, \pi(x)),$$
$$(T^\pi Q)(x,a) \triangleq r(x,a) + \gamma \int_{\mathcal{X}} Q(y, \pi(y)) P(dy|x,a).$$

To avoid unnecessary clutter, we use the same symbol to denote both operators. However, this should not introduce any ambiguity: Given some expression involving T^π one can always determine which operator T^π means by looking at the type of function T^π is applied to. It is known that 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 $T^\pi V^\pi = V^\pi$, see e.g., Proposition 4.2(b) of Bertsekas and Shreve (1978). We will also need to define the so-called Bellman *optimality* operators:

Definition 5 (Bellman Optimality Operators) *The Bellman optimality operators $T^* : B(\mathcal{X}) \to B(\mathcal{X})$ (for value functions) and $T^* : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ (for action-value functions) are defined as*

$$(T^* V)(x) \triangleq \max_a \left\{ r(x,a) + \gamma \int_{\mathcal{X}} V(y) P(dy|x,a) \right\},$$
$$(T^* Q)(x,a) \triangleq r(x,a) + \gamma \int_{\mathcal{X}} \max_{a'} Q(y,a') P(dy|x,a).$$

Again, we use the same symbol to denote both operators; the previous comment that no ambiguity should arise because of this still applies. The Bellman optimality operators enjoy a fixed-point property similar to that of the Bellman operators. In particular, $T^* V^* = V^*$ and $T^* Q^* = Q^*$, see e.g., Proposition 4.2(a) of Bertsekas and Shreve (1978). The Bellman optimality operator thus provides a vehicle to compute the optimal action-value function and therefore to compute an optimal policy.

2. Measurability issues are dealt with in Section 9.5 of the same book. In the case of finitely many actions, no additional condition is needed besides the obvious measurability assumptions on the immediate reward function and the transition kernel (Bertsekas and Shreve, 1978, Corollary 9.17.1), which we will assume from now on.

2.2 Norms and Function Spaces

In what follows we use $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ to denote a subset of measurable functions. The exact specification of this set will be clear from the context. Further, we let $\mathcal{F}^{|\mathcal{A}|} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{A}|}$ to be a subset of vector-valued measurable functions with the identification of

$$\mathcal{F}^{|\mathcal{A}|} = \left\{ (Q_1, \dots, Q_{|\mathcal{A}|}) \,:\, Q_i \in \mathcal{F},\ i = 1, \dots, |\mathcal{A}| \right\}.$$

We shall use $\|Q\|_{p,\nu}$ to denote the $L_p(\nu)$-norm ($1 \le p < \infty$) of a measurable function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, i.e., $\|Q\|_{p,\nu}^p \triangleq \int_{\mathcal{X}\times\mathcal{A}} |Q(x,a)|^p d\nu(x,a)$.

Let $z_{1:n}$ denote the $\mathcal{Z}$-valued sequence $(z_1, \dots, z_n)$. For $\mathcal{D}_n = z_{1:n}$, define the empirical norm of function $f : \mathcal{Z} \to \mathbb{R}$ as

$$\|f\|_{p,\mathcal{D}_n}^p = \|f\|_{p,z_{1:n}}^p \triangleq \frac{1}{n} \sum_{i=1}^{n} |f(z_i)|^p. \tag{1}$$

When there is no chance of confusion about $\mathcal{D}_n$, we may denote the empirical norm by $\|f\|_{p,n}^p$. Based on this definition, one may define $\|Q\|_{p,\mathcal{D}_n}$ with the choice of $\mathcal{Z} = \mathcal{X} \times \mathcal{A}$. Note that if $\mathcal{D}_n = (Z_i)_{i=1}^n$ is random with $Z_i \sim \nu$, the empirical norm is random too, and for any fixed function f, we have $\mathbb{E}\left[\|f\|_{p,\mathcal{D}_n}\right] = \|f\|_{p,\nu}$. When $p = 2$, we simply use $\|\cdot\|_\nu$ and $\|\cdot\|_{\mathcal{D}_n}$.

2.3 Offline Learning Problem and Empirical Bellman Operators

We consider the *offline learning* scenario when we are only given a batch of data[3]

$$\mathcal{D}_n = \{(X_1, A_1, R_1, X_1'), \dots, (X_n, A_n, R_n, X_n')\}, \tag{2}$$

with $X_i \sim \nu_{\mathcal{X}}$, $A_i \sim \pi_b(\cdot|X_i)$, and $(R_i, X_i') \sim P(\cdot, \cdot|X_i, A_i)$ for $i = 1, \dots, n$. Here $\nu_{\mathcal{X}} \in \mathcal{M}(\mathcal{X})$ is a fixed distribution over the states and π_b is the data generating behavior policy, which is a stochastic stationary Markov policy, i.e., given any state $x \in \mathcal{X}$, it assigns a probability distribution over $\mathcal{A}$. We shall also denote the common distribution underlying (X_i, A_i) by $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$.

Samples X_i and X_{i+1} may be sampled independently (we call this the *"Planning scenario"*), or may be coupled through $X_i' = X_{i+1}$ (*"RL scenario"*). In the latter case the data comes from a single trajectory. Under either of these scenarios, we say that the data $\mathcal{D}_n$ meets the *standard offline sampling assumption*. We analyze the Planning scenario, where the states are independent, but one may also analyze dependent processes by considering mixing processes and using tools such as the independent blocks technique (Yu, 1994; Doukhan, 1994), as has been done by Antos et al. (2008b); Farahmand and Szepesvári (2012).

The data set $\mathcal{D}_n$ allows us to define the so-called *empirical Bellman operators*, which can be thought of as empirical approximations to the true Bellman operators.

3. In what follows, when $\{\cdot\}$ is used in connection to a data set, we treat the set as an ordered multiset, where the ordering is given by the time indices of the data points.

Definition 6 (Empirical Bellman Operators) *Let $\mathcal{D}_n$ be a data set as above. Define the ordered multiset $S_n = \{(X_1, A_1), \ldots, (X_n, A_n)\}$. For a given fixed policy π, the empirical Bellman operator $\hat{T}^\pi : \mathbb{R}^{S_n} \to \mathbb{R}^n$ is defined as*

$$(\hat{T}^\pi Q)(X_i, A_i) \triangleq R_i + \gamma Q(X_i', \pi(X_i')), \quad 1 \le i \le n.$$

Similarly, the empirical Bellman optimality operator $\hat{T}^* : \mathbb{R}^{S_n} \to \mathbb{R}^n$ is defined as

$$(\hat{T}^* Q)(X_i, A_i) \triangleq R_i + \gamma \max_{a'} Q(X_i', a'), \quad 1 \le i \le n.$$

In words, the empirical Bellman operators get an n-element list S_n and return an n-dimensional real-valued vector of the single-sample estimate of the Bellman operators applied to the action-value function Q at the selected points. It is easy to see that the empirical Bellman operators provide an unbiased estimate of the Bellman operators in the following sense: For any fixed bounded measurable deterministic function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, policy π and $1 \le i \le n$, it holds that $\mathbb{E}\left[\hat{T}^\pi Q(X_i, A_i) | X_i, A_i\right] = T^\pi Q(X_i, A_i)$ and $\mathbb{E}\left[\hat{T}^* Q(X_i, A_i) | X_i, A_i\right] = T^* Q(X_i, A_i)$.

3. Approximate Policy Iteration

The policy iteration algorithm computes a sequence of policies such that the new policy in the iteration is greedy w.r.t. the action-value function of the previous policy. This procedure requires one to compute the action-value function of the most recent policy (policy evaluation step) followed by the computation of the greedy policy (policy improvement step). In API, the exact, but infeasible, policy evaluation step is replaced by an approximate one. Thus, the skeleton of API methods is as follows: At the k^{th} iteration and given a policy π_k, the API algorithm approximately evaluates π_k to find a Q_k. The action-value function Q_k is typically chosen to be such that $Q_k \approx T^{\pi_k} Q_k$, i.e., it is an approximate fixed point of T^{π_k}. The API algorithm then calculates the greedy policy w.r.t. the most recent action-value function to obtain a new policy π_{k+1}, i.e., $\pi_{k+1} = \hat{\pi}(\cdot; Q_k)$. The API algorithm continues by repeating this process again and generating a sequence of policies and their corresponding approximate action-value functions $Q_0 \to \pi_1 \to Q_1 \to \pi_2 \to \cdots$.[4]

The success of an API algorithm hinges on the way the approximate policy evaluation step is implemented. Approximate policy evaluation is non-trivial for at least two reasons. First, policy evaluation is an inverse problem,[5] so the underlying learning problem is unlike a standard supervised learning problem in which the data take the form of input-output pairs. The second problem is the off-policy sampling problem: The distribution of (X_i, A_i) in the data samples (possibly generated by a behavior policy) is typically different from the distribution that would be induced if we followed the to-be-evaluated policy (i.e., target policy). This causes a problem since the methods must be able to handle this mismatch of

4. In an actual API implementation, one does not need to compute π_{k+1} for all states, which in fact is infeasible for large state spaces. Instead, one uses Q_k to compute π_{k+1} at some select states, as required in the approximate policy evaluation step.
5. Given an operator $\mathcal{L} : \mathcal{F} \to \mathcal{F}$, the inverse problem is the problem of solving $g = \mathcal{L} f$ for f when g is known. In the policy evaluation problem, $\mathcal{L} = \mathbf{I} - \gamma P^\pi$, $g(\cdot) = r(\cdot, \pi(\cdot))$, and $f = Q^\pi$.

distributions.[6] In the rest of this section, we review generic LSTD and BRM methods for approximate policy evaluation. We introduce our regularized version of LSTD and BRM in Section 4.

3.1 Bellman Residual Minimization

The idea of BRM goes back at least to the work of Schweitzer and Seidmann (1985). It was later used in the RL community by Williams and Baird (1994) and Baird (1995). The basic idea of BRM comes from noticing that the action-value function is the unique fixed point of the Bellman operator: $Q^\pi = T^\pi Q^\pi$ (or similarly $V^\pi = T^\pi V^\pi$ for the value function). Whenever we replace Q^π by an action-value function Q different from Q^π, the fixed-point equation would not hold anymore, and we have a non-zero residual function $Q - T^\pi Q$. This quantity is called the *Bellman residual* of Q. The same is true for the Bellman optimality operator T^*.

The BRM algorithm minimizes the norm of the Bellman residual of Q, which is called the *Bellman error*. It can be shown that if $\|Q - T^*Q\|$ is small, then the value function of the greedy policy w.r.t. Q, that is $V^{\hat{\pi}(\cdot;Q)}$, is also in some sense close to the optimal value function V^*, see e.g., Williams and Baird (1994); Munos (2003); Antos et al. (2008b); Farahmand et al. (2010), and Theorem 13 of this work. The BRM algorithm is defined as the procedure minimizing the following loss function:

$$L_{BRM}(Q;\pi) \triangleq \|Q - T^\pi Q\|_\nu^2 ,$$

where ν is the distribution of state-actions in the input data. Using the empirical L_2-norm defined in (1) with samples $\mathcal{D}_n$ defined in (2), and by replacing $(T^\pi Q)(X_t, A_t)$ with the empirical Bellman operator (Definition 6), the empirical estimate of $L_{BRM}(Q;\pi)$ can be written as

$$\hat{L}_{BRM}(Q;\pi,n) \triangleq \left\|Q - \hat{T}^\pi Q\right\|_{\mathcal{D}_n}^2 = \frac{1}{n}\sum_{t=1}^{n}\left[Q(X_t,A_t) - \Big(R_t + \gamma Q\big(X_t', \pi(X_t')\big)\Big)\right]^2 . \quad (3)$$

Nevertheless, it is well-known that $\hat{L}_{BRM}$ is *not* an unbiased estimate of L_{BRM} when the MDP is not deterministic (Lagoudakis and Parr, 2003; Antos et al., 2008b). To address this issue, Antos et al. (2008b) propose the modified BRM loss that is a new empirical loss function with an extra *de-biasing* term. The idea of the modified BRM is to cancel the unwanted variance by introducing an auxiliary function h and a new loss function

$$L_{BRM}(Q,h;\pi) = L_{BRM}(Q;\pi) - \|h - T^\pi Q\|_\nu^2 , \quad (4)$$

and approximating the action-value function Q^π by solving

$$Q_{BRM} = \operatorname*{argmin}_{Q\in\mathcal{F}^{|\mathcal{A}|}} \sup_{h\in\mathcal{F}^{|\mathcal{A}|}} L_{BRM}(Q,h;\pi), \quad (5)$$

6. A number of works in the domain adaptation literature consider this scenario under the name of covariate shift problem, see e.g., Ben-David et al. 2006; Mansour et al. 2009; Ben-David et al. 2010; Cortes et al. 2015.

where the supremum comes from the negative sign of $\|h - T^\pi Q\|_\nu^2$. They have shown that optimizing the new loss function still makes sense and the empirical version of this loss is unbiased.

The min-max optimization problem (5) is equivalent to the following coupled (nested) optimization problems:

$$h(\cdot; Q) = \operatorname*{argmin}_{h' \in \mathcal{F}^{|\mathcal{A}|}} \left\| h' - T^\pi Q \right\|_\nu^2 ,$$
$$Q_{BRM} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \Big[\|Q - T^\pi Q\|_\nu^2 - \|h(\cdot; Q) - T^\pi Q\|_\nu^2 \Big]. \tag{6}$$

In practice, the norm $\|\cdot\|_\nu$ is replaced by the empirical norm $\|\cdot\|_{\mathcal{D}_n}$ and $T^\pi Q$ is replaced by its sample-based approximation $\hat{T}^\pi Q$, i.e.,

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left\| h - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2 , \tag{7}$$
$$\hat{Q}_{BRM} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \Big[\left\| Q - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2 - \left\| \hat{h}_n(\cdot; Q) - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2 \Big]. \tag{8}$$

From now on, whenever we refer to the BRM algorithm, we are referring to this modified BRM.

3.2 Least-Squares Temporal Difference Learning

The Least-Squares Temporal Difference learning (LSTD) algorithm for policy evaluation was first proposed by Bradtke and Barto (1996), and later used in an API procedure by Lagoudakis and Parr (2003) and was called Least-Squares Policy Iteration (LSPI).

The original formulation of LSTD finds a solution to the fixed-point equation $Q = \Pi_\nu T^\pi Q$, where Π_ν is the simplified notation for ν-weighted projection operator onto the space of admissible functions $\mathcal{F}^{|\mathcal{A}|}$, i.e., $\Pi_\nu \triangleq \Pi_{\mathcal{F}_\nu^{|\mathcal{A}|}} : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ is defined by $\Pi_{\mathcal{F}_\nu^{|\mathcal{A}|}} Q = \operatorname{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \|h - Q\|_\nu^2$ for $Q \in B(\mathcal{X} \times \mathcal{A})$. We, however, use a different optimization-based formulation. The reason is that whenever ν is not the stationary distribution induced by π, the operator $(\Pi_\nu T^\pi)$ does not necessarily have a fixed point, but the optimization problem is always well-defined.

We define the LSTD solution as the minimizer of the L_2-norm between Q and $\Pi_\nu T^\pi Q$:

$$L_{LSTD}(Q; \pi) \triangleq \|Q - \Pi_\nu T^\pi Q\|_\nu^2 . \tag{9}$$

The minimizer of $L_{LSTD}(Q; \pi)$ is well-defined, and whenever ν is the stationary distribution of π (i.e., on-policy sampling), the solution to this optimization problem is the same as the solution to $Q = \Pi_\nu T^\pi Q$. The LSTD solution can therefore be written as the solution to the following set of coupled optimization problems:

$$h(\cdot; Q) = \operatorname*{argmin}_{h' \in \mathcal{F}^{|\mathcal{A}|}} \left\| h' - T^\pi Q \right\|_\nu^2 ,$$
$$Q_{LSTD} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \|Q - h(\cdot; Q)\|_\nu^2 , \tag{10}$$

Algorithm 1 Regularized Policy Iteration(K,$\hat{Q}^{(-1)}$,$\mathcal{F}^{|\mathcal{A}|}$,$J$,$\{(\lambda_{Q,n}^{(k)}, \lambda_{h,n}^{(k)})\}_{k=0}^{K-1}$)

// K: Number of iterations
// $\hat{Q}^{(-1)}$: Initial action-value function
// $\mathcal{F}^{|\mathcal{A}|}$: The action-value function space
// J: The regularizer
// $\{(\lambda_{Q,n}^{(k)}, \lambda_{h,n}^{(k)})\}_{k=0}^{K}$: The regularization coefficients
for $k = 0$ to $K-1$ **do**
  $\pi_k(\cdot) \leftarrow \hat{\pi}(\cdot; \hat{Q}^{(k-1)})$
  Generate training samples $\mathcal{D}_n^{(k)}$
  $\hat{Q}^{(k)} \leftarrow$ REG-LSTD/BRM$(\pi_k, \mathcal{D}_n^{(k)}; \mathcal{F}^{|\mathcal{A}|}, J, \lambda_{Q,n}^{(k)}, \lambda_{h,n}^{(k)})$
end for
return $\hat{Q}^{(K-1)}$ and $\pi_K(\cdot) = \hat{\pi}(\cdot; \hat{Q}^{(K-1)})$

where the first equation finds the projection of $T^\pi Q$ onto $\mathcal{F}^{|\mathcal{A}|}$, and the second one minimizes the distance of Q and the projection. The corresponding empirical version based on data set $\mathcal{D}_n$ is

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left\| h - \hat{T}^\pi Q \right\|_{\mathcal{D}_n}^2, \tag{11}$$

$$\hat{Q}_{LSTD} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left\| Q - \hat{h}_n(\cdot; Q) \right\|_{\mathcal{D}_n}^2. \tag{12}$$

For general spaces $\mathcal{F}^{|\mathcal{A}|}$, these optimization problems can be difficult to solve, but when $\mathcal{F}^{|\mathcal{A}|}$ is a linear subspace of $B(\mathcal{X} \times \mathcal{A})$, the minimization problem becomes computationally feasible.

Comparison of BRM and LSTD is noteworthy. The population version of LSTD loss minimizes the distance between Q and $\Pi_\nu T^\pi Q$, which is $\|Q - \Pi_\nu T^\pi Q\|_\nu^2$. Meanwhile, BRM minimizes another distance function that is the distance between $T^\pi Q$ and $\Pi_\nu T^\pi Q$ subtracted from the distance between Q and $T^\pi Q$, i.e., $\|Q - T^\pi Q\|_\nu^2 - \|\hat{h}_n(\cdot; Q) - T^\pi Q\|_\nu^2$. See Figure 1a for a pictorial presentation of these distances. When $\mathcal{F}^{|\mathcal{A}|}$ is linear, because of the Pythagorean theorem, the solution to the modified BRM (6) coincides with the LSTD solution (10) (Antos et al., 2008b).

4. Regularized Policy Iteration Algorithms

In this section we introduce two *Regularized Policy Iteration* algorithms, which are instances of the generic API algorithms. These algorithms are built on the regularized extensions of BRM (Section 3.1) and LSTD (Section 3.2) for the task of approximate policy evaluation.

The pseudo-code of the Regularized Policy Iteration algorithms is shown in Algorithm 1. The algorithm receives K (the number of API iterations), an initial action-value function $\hat{Q}^{(-1)}$, the function space $\mathcal{F}^{|\mathcal{A}|}$, the regularizer $J : \mathcal{F}^{|\mathcal{A}|} \to \mathbb{R}$, and a set of regularization coefficients $\{(\lambda_{Q,n}^{(k)}, \lambda_{h,n}^{(k)})\}_{k=0}^{K-1}$. Each iteration starts with a step of policy improvement, i.e.,

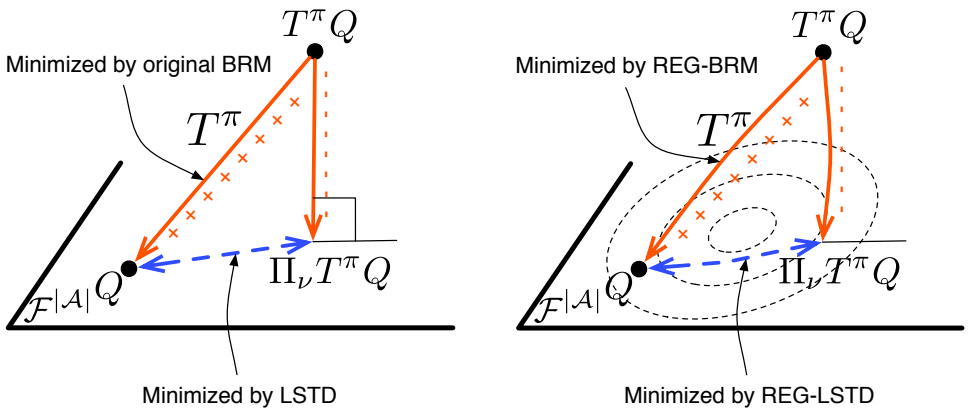

Figure 1: (a) This figure shows the loss functions minimized by the original BRM, the modified BRM, and the LSTD methods. The function space $\mathcal{F}^{|\mathcal{A}|}$ is represented by the plane. The Bellman operator T^π maps an action-value function $Q \in \mathcal{F}^{|\mathcal{A}|}$ to a function $T^\pi Q$. The function $T^\pi Q - \Pi_\nu T^\pi Q$ is orthogonal to $\mathcal{F}^{|\mathcal{A}|}$. The original BRM loss function is $\|Q - T^\pi Q\|_\nu^2$ (solid line), the modified BRM loss is $\|Q - T^\pi Q\|_\nu^2 - \|T^\pi Q - \Pi_\nu T^\pi Q\|_\nu^2$ (the difference of two solid line segments; note the $+$ and $-$ symbols), and the LSTD loss is $\|Q - \Pi_\nu T^\pi Q\|_\nu^2$ (dashed line). LSTD and the modified BRM are equivalent for linear function spaces. (b) REG-LSTD and REG-BRM minimize regularized objective functions. Regularization makes the function $T^\pi Q - \Pi_\nu T^\pi Q$ to be non-orthogonal to $\mathcal{F}^{|\mathcal{A}|}$. The dashed ellipsoids represent the level-sets defined by the regularization functional J.

$\pi_k \leftarrow \hat{\pi}(\cdot; \hat{Q}^{(k-1)}) = \operatorname{argmax}_{a' \in \mathcal{A}} \hat{Q}^{(k-1)}(\cdot, a')$. For the first iteration ($k = 0$), one may ignore this step and provide an initial policy π_0 instead of $\hat{Q}^{(-1)}$. Afterwards, we have a data generating step: At each iteration $k = 0, \ldots, K-1$, the agent follows the data generating policy π_{b_k} to obtain $\mathcal{D}_n^{(k)} = \{(X_t^{(k)}, A_t^{(k)}, R_t^{(k)}, X_t'^{(k)})\}_{1 \le t \le n}$. For the k^{th} iteration of the algorithm, we use training samples $\mathcal{D}_n^{(k)}$ to evaluate policy π_k. In practice, one might want to change π_{b_k} at each iteration in such a way that the agent ultimately achieves a better performance. The relation between the performance and the choice of data samples, however, is complicated. For simplicity of analysis, in the rest of this work we assume that a fixed behavior policy is used in all iterations, i.e., $\pi_{b_k} = \pi_b$.[7] This leads to K independent data sets $\mathcal{D}_n^{(0)}, \ldots, \mathcal{D}_n^{(K-1)}$. From now on, to avoid clutter, we use symbols $\mathcal{D}_n, X_t, \ldots$ instead of $\mathcal{D}_n^{(k)}, X_t^{(k)}, \ldots$ with the understanding that each $\mathcal{D}_n$ in various iterations is referring to an independent set of data samples, which should be clear from the context.

The approximate policy evaluation step is performed by REG-LSTD/BRM, which will be discussed shortly. REG-LSTD/BRM receives policy π_k, the training samples $\mathcal{D}_n^{(k)}$, the function space $\mathcal{F}^{|\mathcal{A}|}$, the regularizer J, and the regularization coefficients $(\lambda_{Q,n}^{(k)}, \lambda_{h,n}^{(k)})$, and

7. So we are in the so-called *off-policy sampling* scenario.

returns an estimate of the action-value function of policy π_k. This procedure repeats for K iterations.

REG-BRM approximately evaluates policy π_k by solving the following coupled optimization problems:

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| h - \hat{T}^{\pi_k} Q \right\|_{\mathcal{D}_n}^2 + \lambda_{h,n}^{(k)} J^2(h) \right], \tag{13}$$

$$\hat{Q}^{(k)} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| Q - \hat{T}^{\pi_k} Q \right\|_{\mathcal{D}_n}^2 - \left\| \hat{h}_n(\cdot; Q) - \hat{T}^{\pi_k} Q \right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n}^{(k)} J^2(Q) \right], \tag{14}$$

where $J : \mathcal{F}^{|\mathcal{A}|} \to \mathbb{R}$ is the regularization functional (or simply regularizer or penalizer), and $\lambda_{h,n}^{(k)}, \lambda_{Q,n}^{(k)} > 0$ are regularization coefficients. The regularizer can be any pseudo-norm defined on $\mathcal{F}^{|\mathcal{A}|}$; and $\mathcal{D}_n$ is defined as (2).[8] The regularizer is often chosen such that the functions that we believe are more "complex" have larger values of J. The notion of complexity, however, is subjective and depends on the choice of $\mathcal{F}^{|\mathcal{A}|}$ and J. Finally note that we call $J(Q)$ the smoothness of Q, even though it might not coincide with the conventional derivative-based notions of smoothness.

An example of the case that J has a derivative-based interpretation is when the function space $\mathcal{F}^{|\mathcal{A}|}$ is a Sobolev space and the regularizer J is defined as its corresponding norm. In this case, we are penalizing the weak-derivatives of the estimate (Györfi et al., 2002; van de Geer, 2000). One can generalize the notion of smoothness beyond the usual derivative-based ones (cf. Chapter 1 of Triebel 2006) and define function spaces such as the family of Besov spaces (Devore, 1998). The RKHS norm for shift-invariant and radial kernels can also be interpreted as a penalizer of higher-frequency terms of the function (i.e., a low-pass filter Evgeniou et al. 1999), so they effectively encourage "smoother" functions. The choice of kernel determines the frequency response of the filter. One may also use other data-dependent regularizers such as manifold regularization (Belkin et al., 2006) and Sample-based Approximate Regularization (Bachman et al., 2014). As a final example, for the functions in the form of $Q(x,a) = \sum_{i \geq 1} \phi_i(x,a) w_i$, if we choose a sparsity-inducing regularizer such as $J(Q) \triangleq \sum_{i \geq 1} |w_i|$ as the measure of smoothness, then a function that has a sparse representation in the dictionary $\{\phi_i\}_{i \geq 1}$ is, by definition, a smooth function—even though there is not necessarily any connection to the derivative-based smoothness.

REG-LSTD approximately evaluates the policy π_k by solving the following coupled optimization problems:

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| h - \hat{T}^{\pi_k} Q \right\|_{\mathcal{D}_n}^2 + \lambda_{h,n}^{(k)} J^2(h) \right], \tag{15}$$

$$\hat{Q}^{(k)} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| Q - \hat{h}_n(\cdot; Q) \right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n}^{(k)} J^2(Q) \right]. \tag{16}$$

Note that the difference between (7)-(8) ((11)-(12)) and (13)-(14) ((15)-(16)) is the addition of the regularizers $J^2(h)$ and $J^2(Q)$.

Unlike the non-regularized case described in Section 3, the solutions of REG-BRM and REG-LSTD are not the same. As a result of the *regularized* projection, (13) and

8. A pseudo-norm J satisfies all properties of a norm except that $J(Q) = 0$ does not imply that $Q = 0$.

(15), the function $\hat{h}_n(\cdot; Q) - \hat{T}^{\pi_k} Q$ is not orthogonal to the function space $\mathcal{F}^{|\mathcal{A}|}$—even if $\mathcal{F}^{|\mathcal{A}|}$ is a linear space. Therefore, the Pythagorean theorem is not applicable anymore: $\|Q - \hat{h}_n(\cdot; Q)\|^2 \neq \|Q - \hat{T}^{\pi_k} Q\|^2 - \|\hat{h}_n(\cdot; Q) - \hat{T}^{\pi_k} Q\|^2$ (See Figure 1b).

One may ask why we have regularization terms in both optimization problems, as opposed to only in the projection term (15) (similar to the Lasso-TD algorithm Kolter and Ng 2009; Ghavamzadeh et al. 2011) or only in (16) (similar to Geist and Scherrer 2012; Ávila Pires and Szepesvári 2012). We discuss this question in Section 4.1. Briefly speaking, for large function spaces such as the Sobolev spaces or the RKHS with universal kernels, if we remove the regularization term in (15), the coupled optimization problems reduces to (unmodified) BRM, which is biased as discussed earlier; whereas if the regularization term in (16) is removed, the solution can be arbitrary bad due to overfitting.

Finally note that the choice of the function space $\mathcal{F}^{|\mathcal{A}|}$, the regularizer J, and the regularization coefficients $\lambda_{Q,n}^{(k)}$ and $\lambda_{h,n}^{(k)}$ all affect the sample efficiency of the algorithms. If one knew $J(Q^\pi)$, the regularization coefficients could be chosen optimally. Nonetheless, the value of $J(Q^\pi)$ is often not known, so one has to use a model selection procedure to set the best function space and the regularization coefficients. The situation is similar to the problem of model selection in supervised learning (though the solutions are different). After developing some tools necessary for discussing this issue in Section 5, we return to the problem of choosing the regularization coefficients after Theorem 11 as well as in Section 6.

Remark 7 *To the best of our knowledge, Antos et al. (2008b) were the first who explicitly considered LSTD as the optimizer of the loss function* (9)*. Their discussion was mainly to prove the equivalence of modified BRM* (5) *and LSTD when $\mathcal{F}^{|\mathcal{A}|}$ is a linear function space. In their work, the loss function is not used to derive any new algorithm. Farahmand et al. (2009b) used this loss function to develop the regularized variant of LSTD* (15)-(16)*. This loss function was later called* mean-square projected Bellman error *by Sutton et al. (2009), and was used to derive the GTD2 and TDC algorithms.*

4.1 Why Two Regularizers?

We discuss why using regularizers in both optimization problems (15) and (16) of REG-LSTD is necessary for large function spaces such as the Sobolev spaces and the RKHS with universal kernels. Here we show that for large function spaces, depending on which regularization term we remove, either the coupled optimization problems reduces to the regularized variant of the unmodified BRM, which has a bias, or the solution can be arbitrary bad.

Let us focus on REG-LSTD for a given policy π. Assume that the function space $\mathcal{F}^{|\mathcal{A}|}$ is rich enough in the sense that it is dense in the space of continuous functions w.r.t. the supremum norm. This is satisfied by many large function spaces such as RKHS with universal kernels (Definition 4.52 of Steinwart and Christmann 2008) and the Sobolev spaces on compact domains. We consider what would happen if instead of the current formulation of REG-LSTD (15)-(16), we only used a regularizer either in the first or second optimization problem. We study each case separately. For notational simplicity, we omit the dependence on the iteration number k.

Case 1. In this case, we only regularize the empirical error $\|Q - \hat{h}_n(\cdot; Q)\|^2_{\mathcal{D}_n}$, but we do not regularize the projection, i.e.,

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left\| h - \hat{T}^\pi Q \right\|^2_{\mathcal{D}_n},$$
$$\hat{Q} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| Q - \hat{h}_n(\cdot; Q) \right\|^2_{\mathcal{D}_n} + \lambda_{Q,n} J^2(Q) \right]. \tag{17}$$

When the function space $\mathcal{F}^{|\mathcal{A}|}$ is rich enough, there exists a function $\hat{h}_n \in \mathcal{F}^{|\mathcal{A}|}$ that fits perfectly well to its target values at data points $\{(X_i, A_i)\}_{i=1}^n$, that is, $\hat{h}_n((X_i, A_i); Q) = (\hat{T}^\pi Q)(X_i, A_i)$ for $i = 1, \ldots, n$.[9] Such a function is indeed the minimizer of the loss $\|Q - \hat{h}_n(\cdot; Q)\|^2_{\mathcal{D}_n}$. The second optimization problem (17) becomes

$$\hat{Q} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| Q - \hat{T}^\pi Q \right\|^2_{\mathcal{D}_n} + \lambda_{Q,n} J^2(Q) \right].$$

This is the regularized version of the original (i.e., unmodified) formulation of the BRM algorithm. As discussed in Section 3.1, the unmodified BRM algorithm is biased when the MDP is not deterministic. Adding a regularizer does not solve the biasedness problem of the unmodified BRM loss. So without regularizing the first optimization problem, the function $\hat{h}_n$ overfits to the noise and as a result the whole algorithm becomes incorrect.

Case 2. In this case, we only regularize the empirical projection $\|h - \hat{T}^\pi Q\|^2_{\mathcal{D}_n}$, but we do not regularize $\|Q - \hat{h}_n(\cdot; Q)\|^2_{\mathcal{D}_n}$, i.e.,

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| h - \hat{T}^\pi Q \right\|^2_{\mathcal{D}_n} + \lambda_{h,n} J^2(h) \right],$$
$$\hat{Q} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \left\| Q - \hat{h}_n(\cdot; Q) \right\|^2_{\mathcal{D}_n}. \tag{18}$$

For a fixed Q, the first optimization problem is the standard regularized regression estimator with the regression function $\mathbb{E}\left[(\hat{T}^\pi Q)(X, A) | X = x, A = a\right] = (T^\pi Q)(x, a)$. Therefore, if the function space $\mathcal{F}^{|\mathcal{A}|}$ is rich enough and we set the regularization coefficient $\lambda_{h,n}$ properly, $\|h - T^\pi Q\|_\nu$ and $\|h - T^\pi Q\|_{\mathcal{D}_n}$ go to zero as the sample size grows (the rate of convergence depends on the complexity of the target function; cf. Lemma 15 and Theorem 16). So we can expect $\hat{h}_n(\cdot; Q)$ to get closer to $T^\pi Q$ as the sample size grows.

For simplicity of discussion, suppose that we are in the ideal situation where for any Q, we have $\hat{h}_n((x, a); Q) = (T^\pi Q)(x, a)$ for all $(x, a) \in \{(X_i, A_i)\}_{i=1}^n \cup \{(X'_i, \pi(X'_i))\}_{i=1}^n$, that

9. To be more precise: First, for an $\varepsilon > 0$, we construct a continuous function $\bar{h}_\varepsilon(z) = \sum_{Z_i \in \{(X_i, A_i)\}_{i=1}^n} \max\left\{1 - \frac{\|z - Z_i\|}{\varepsilon}, 0\right\} (\hat{T}^\pi Q)(Z_i)$. We then use the denseness of the function space $\mathcal{F}^{|\mathcal{A}|}$ in the supremum norm to argue that there exists $h_\varepsilon \in \mathcal{F}^{|\mathcal{A}|}$ such that $\left\| h_\varepsilon - \bar{h}_\varepsilon \right\|_\infty$ is arbitrarily close to zero. So when $\varepsilon \to 0$, the value of function h_ε is arbitrarily close to $T^\pi Q$ at data points. We then choose $\hat{h}_n(\cdot; Q) = h_\varepsilon$. This construction is similar to Theorem 2 of Nadler et al. (2009). See also the argument in Case 2 for more detail.

is, we precisely know $T^\pi Q$ at all data points.[10] Substituting this $\hat{h}_n((x,a);Q)$ in the second optimization problem (17), we get that we are solving the following optimization problem:

$$\hat{Q} = \operatorname*{argmin}_{Q \in \mathcal{F}^{|\mathcal{A}|}} \|Q - T^\pi Q\|^2_{\mathcal{D}_n} . \tag{19}$$

This is the Bellman error minimization problem. We do not have the biasedness problem here as we have $T^\pi Q$ instead of $\hat{T}^\pi Q$ in the loss. Nonetheless, we face another problem: Minimizing this empirical risk minimization without controlling the complexity of the function space might lead to an overfitted solution, very similar to the same phenomenon in supervised learning.

To see it more precisely, we first construct a continuous function

$$\bar{Q}_\varepsilon(z) = \sum_{Z_i \in \{(X_i,A_i)\}_{i=1}^n \cup \{(X'_i,\pi(X'_i))\}_{i=1}^n} \max\left\{1 - \frac{\|z - Z_i\|}{\varepsilon}, 0\right\} Q^\pi(Z_i),$$

which for small enough $\varepsilon > 0$ has the property that $\left\|\bar{Q}_\varepsilon - T^\pi \bar{Q}_\varepsilon\right\|^2_{\mathcal{D}_n}$ is zero, i.e., it is a minimizer of the empirical loss. Due to the denseness of $\mathcal{F}^{|\mathcal{A}|}$, we can find a $Q_\varepsilon \in \mathcal{F}^{|\mathcal{A}|}$ that is arbitrarily close to the continuous function $\bar{Q}_\varepsilon$. Therefore, for small enough ε, the function Q_ε is a minimizer of (19), i.e., the value of $\|Q_\varepsilon - T^\pi Q_\varepsilon\|^2_{\mathcal{D}_n}$ is zero. But Q_ε is not a good approximation of Q^π because Q_ε consists of spikes in the ε-neighbourhood of data points and zero elsewhere. In other words, Q_ε does not generalize well beyond the data points when ε is chosen to be small.

Of course the solution is to control the complexity of $\mathcal{F}^{|\mathcal{A}|}$ so that spiky functions such as Q_ε are not selected as the solution of the optimization problem. When we regularize both optimization problems, as we do in this work, none of these problems happen.

This argument applies to rich function spaces that can approximate any reasonably complex functions (e.g., continuous functions) arbitrarily well. If the function space $\mathcal{F}^{|\mathcal{A}|}$ is much more limited, for example if it is a parametric function space, we *may* not need to regularize both optimization problems. An example of such an approach for parametric spaces has been analyzed by Ávila Pires and Szepesvári (2012).

4.2 Closed-Form Solutions

In this section we provide a closed-form solution for (13)-(14) and (15)-(16) for two cases: 1) When $\mathcal{F}^{|\mathcal{A}|}$ is a finite dimensional linear space and $J^2(\cdot)$ is defined as the weighted squared sum of parameters describing the function (a setup similar to the ridge regression Hoerl and Kennard 1970) and 2) $\mathcal{F}^{|\mathcal{A}|}$ is an RKHS and $J(\cdot)$ is the corresponding inner-product norm, i.e., $J^2(\cdot) = \|\cdot\|^2_{\mathcal{H}}$. Here we use a generic π and $\mathcal{D}_n$ instead of π_k and $\mathcal{D}_n^{(k)}$ at the k^{th} iteration.

10. This is an ideal situation because 1) $\|h - \hat{T}^\pi Q\|_\nu$ is equal to zero only asymptotically and not in finite samples regime, and 2) even if $\|h - \hat{T}^\pi Q\|_\nu = 0$, it does not imply that $\hat{h}_n(x,a;Q) = (T^\pi Q)(x,a)$ almost surely on $\mathcal{X} \times \mathcal{A}$. Nonetheless, these simplifications are only in favour of the algorithm considered in this case, so for simplicity of discussion we assume that they hold.

4.2.1 A Parametric Formulation for REG-BRM and REG-LSTD

In this section we consider the case when h and Q are both given as linear combinations of some basis functions:

$$h(\cdot) = \boldsymbol{\phi}(\cdot)^\top \boldsymbol{u}, \quad Q(\cdot) = \boldsymbol{\phi}(\cdot)^\top \boldsymbol{w}, \tag{20}$$

where $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^p$ are parameter vectors and $\boldsymbol{\phi}(\cdot) \in \mathbb{R}^p$ is a vector of p linearly independent basis functions defined over the space of state-action pairs.[11] These basis functions might be predefined (e.g., Fourier (Konidaris et al., 2011) or wavelets) or constructed data-dependently by one of already mentioned feature generation methods. We further assume that the regularization terms take the form

$$J^2(h) = \boldsymbol{u}^\top \boldsymbol{\Psi} \boldsymbol{u}\,,$$
$$J^2(Q) = \boldsymbol{w}^\top \boldsymbol{\Psi} \boldsymbol{w}\,.$$

for some user-defined choice of positive definite matrix $\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$. A simple and common choice would be $\boldsymbol{\Psi} = \mathbf{I}$. Define $\boldsymbol{\Phi}, \boldsymbol{\Phi}' \in \mathbb{R}^{n \times p}$ and $\boldsymbol{r} \in \mathbb{R}^n$ as follows:

$$\boldsymbol{\Phi} = \Big(\boldsymbol{\phi}(Z_1), \dots, \boldsymbol{\phi}(Z_n)\Big)^\top, \; \boldsymbol{\Phi}' = \Big(\boldsymbol{\phi}(Z_1'), \dots, \boldsymbol{\phi}(Z_n')\Big)^\top, \; \boldsymbol{r} = \Big(R_1, \dots, R_n\Big)^\top, \tag{21}$$

with $Z_i = (X_i, A_i)$ and $Z_i' = (X_i', \pi(X_i'))$.
The solution to REG-BRM is given by the following proposition.

Proposition 8 (Closed-form solution for REG-BRM) *Under the setting of this section, the approximate action-value function returned by REG-BRM is* $\hat{Q}(\cdot) = \phi(\cdot)^\top \boldsymbol{w}^*$, *where*

$$\boldsymbol{w}^* = \Big[\boldsymbol{B}^\top \boldsymbol{B} - \gamma^2 \boldsymbol{C}^\top \boldsymbol{C} + n\lambda_{Q,n} \boldsymbol{\Psi}\Big]^{-1} \Big(\boldsymbol{B}^\top + \gamma \boldsymbol{C}^\top (\boldsymbol{\Phi}\boldsymbol{A} - \mathbf{I})\Big)\, \boldsymbol{r},$$

with $\boldsymbol{A} = \big(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + n\lambda_{h,n} \boldsymbol{\Psi}\big)^{-1} \boldsymbol{\Phi}^\top$, $\boldsymbol{B} = \boldsymbol{\Phi} - \gamma \boldsymbol{\Phi}'$, $\boldsymbol{C} = (\boldsymbol{\Phi} A - \mathbf{I})\boldsymbol{\Phi}'$.

Proof Using (20) and (21), we can rewrite (13)-(14) as

$$\boldsymbol{u}^*(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^p} \left\{ \frac{1}{n} \big[\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big] + \lambda_{h,n} \boldsymbol{u}^\top \boldsymbol{\Psi} \boldsymbol{u} \right\}, \tag{22}$$

$$\begin{aligned} \boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^p} \Bigg\{ & \frac{1}{n} \big[\boldsymbol{\Phi}\boldsymbol{w} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{w} - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big] - \\ & \frac{1}{n} \big[\boldsymbol{\Phi}\boldsymbol{u}^*(\boldsymbol{w}) - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{u}^*(\boldsymbol{w}) - (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w})\big] + \lambda_{Q,n} \boldsymbol{w}^\top \boldsymbol{\Psi} \boldsymbol{w} \Bigg\}. \end{aligned} \tag{23}$$

Taking the derivative of (22) w.r.t. $\boldsymbol{u}$ and equating it to zero, we obtain $\boldsymbol{u}^*$ as a function of $\boldsymbol{w}$:

$$\boldsymbol{u}^*(\boldsymbol{w}) = \Big(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + n\lambda_{h,n} \boldsymbol{\Psi}\Big)^{-1} \boldsymbol{\Phi}^\top (\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w}) = \boldsymbol{A}(\boldsymbol{r} + \gamma \boldsymbol{\Phi}' \boldsymbol{w}). \tag{24}$$

Plug $\boldsymbol{u}^*(\boldsymbol{w})$ from (24) into (23), take the derivative w.r.t. $\boldsymbol{w}$ and equate it to zero to obtain the parameter vector $\boldsymbol{w}^*$ as announced above. ■

The solution returned by REG-LSTD is given in the following proposition.

11. At the cost of using generalized inverses, everything in this section extends to the case when the basis functions are not linearly independent.

Proposition 9 (Closed-form solution for REG-LSTD) *Under the setting of this section, the approximate action-value function returned by REG-LSTD is* $\hat{Q}(\cdot) = \phi(\cdot)^\top \boldsymbol{w}^*$, *where*

$$\boldsymbol{w}^* = \Big[\boldsymbol{E}^\top \boldsymbol{E} + n\lambda_{Q,n}\boldsymbol{\Psi}\Big]^{-1} \boldsymbol{E}^\top \boldsymbol{A}\boldsymbol{r},$$

with $\boldsymbol{A} = \big(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + n\lambda_{h,n}\boldsymbol{\Psi}\big)^{-1}\boldsymbol{\Phi}^\top$ *and* $\boldsymbol{E} = (\boldsymbol{\Phi} - \gamma\boldsymbol{A}\boldsymbol{\Phi}')$.

Proof Using (20) and (21), we can rewrite (15)-(16) as

$$\boldsymbol{u}^*(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^p} \left\{ \frac{1}{n}\big[\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{r} + \gamma\boldsymbol{\Phi}'\boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{u} - (\boldsymbol{r} + \gamma\boldsymbol{\Phi}'\boldsymbol{w})\big] + \lambda_{h,n}\boldsymbol{u}^\top\boldsymbol{\Psi}\boldsymbol{u} \right\}, \tag{25}$$

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^p} \left\{ \big[\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}\boldsymbol{u}^*(\boldsymbol{w})\big]^\top \big[\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}\boldsymbol{u}^*(\boldsymbol{w})\big] + \lambda_{Q,n}\boldsymbol{w}^\top\boldsymbol{\Psi}\boldsymbol{w} \right\}. \tag{26}$$

Similar to the parametric REG-BRM, we solve (25) and obtain $\boldsymbol{u}^*(\boldsymbol{w})$ which is the same as (24). If we plug this $\boldsymbol{u}^*(\boldsymbol{w})$ into (26), take derivate w.r.t. $\boldsymbol{w}$, and find the minimizer, the parameter vector $\boldsymbol{w}^*$ will be as announced. ■

4.2.2 RKHS Formulation for REG-BRM and REG-LSTD

The class of reproducing kernel Hilbert spaces provides a flexible and powerful family of function spaces to choose $\mathcal{F}^{|\mathcal{A}|}$ from. An RKHS $\mathcal{H} : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$ is defined by a positive definite kernel $\mathrm{K} : (\mathcal{X}\times\mathcal{A})\times(\mathcal{X}\times\mathcal{A}) \to \mathbb{R}$. With such a choice, we can use the corresponding squared RKHS norm $\|\cdot\|^2_{\mathcal{H}}$ as the regularizer $J^2(\cdot)$. REG-BRM with an RKHS function space $\mathcal{F}^{|\mathcal{A}|} = \mathcal{H}$ would be

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h\in\mathcal{F}^{|\mathcal{A}|}[=\mathcal{H}]} \left[\left\| h - \hat{T}^\pi Q \right\|^2_{\mathcal{D}_n} + \lambda_{h,n}\|h\|^2_{\mathcal{H}} \right], \tag{27}$$

$$\hat{Q} = \operatorname*{argmin}_{Q\in\mathcal{F}^{|\mathcal{A}|}[=\mathcal{H}]} \left[\left\| Q - \hat{T}^\pi Q \right\|^2_{\mathcal{D}_n} - \left\| \hat{h}_n(\cdot; Q) - \hat{T}^\pi Q \right\|^2_{\mathcal{D}_n} + \lambda_{Q,n}\|Q\|^2_{\mathcal{H}} \right], \tag{28}$$

and the coupled optimization problems for REG-LSTD are

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h\in\mathcal{F}^{|\mathcal{A}|}[=\mathcal{H}]} \left[\left\| h - \hat{T}^\pi Q \right\|^2_{\mathcal{D}_n} + \lambda_{h,n}\|h\|^2_{\mathcal{H}} \right], \tag{29}$$

$$\hat{Q} = \operatorname*{argmin}_{Q\in\mathcal{F}^{|\mathcal{A}|}[=\mathcal{H}]} \left[\left\| Q - \hat{h}_n(\cdot; Q) \right\|^2_{\mathcal{D}_n} + \lambda_{Q,n}\|Q\|^2_{\mathcal{H}} \right]. \tag{30}$$

We can solve these coupled optimization problems by the application of the generalized representer theorem for RKHS (Schölkopf et al., 2001). The result, which is stated in the next theorem, shows that the infinite dimensional optimization problem defined on $\mathcal{F}^{|\mathcal{A}|} = \mathcal{H}$ boils down to a finite dimensional problem with the dimension twice the number of data points.

Theorem 10 *Let* $\tilde{Z}$ *be a vector defined as* $\tilde{Z} = (Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n)^\top$. *Then the optimizer* $\hat{Q} \in \mathcal{H}$ *of* (27)-(28) *can be written as* $\hat{Q}(\cdot) = \sum_{i=1}^{2n} \tilde{\alpha}_i \mathrm{K}(\tilde{Z}_i, \cdot)$ *for some values of*

$\tilde{\alpha} \in \mathbb{R}^{2n}$. *The same holds for the solution to* (29)-(30). *Further, the coefficient vectors can be obtained in the following form:*

$$\textbf{\textit{REG-BRM:}} \qquad \tilde{\boldsymbol{\alpha}}_{BRM} = (\boldsymbol{C}\boldsymbol{K}_Q + n\lambda_{Q,n}\mathbf{I})^{-1}(\boldsymbol{D}^\top + \gamma \boldsymbol{C}_2^\top \boldsymbol{B}^\top \boldsymbol{B})\boldsymbol{r}\,,$$

$$\textbf{\textit{REG-LSTD:}} \qquad \tilde{\boldsymbol{\alpha}}_{LSTD} = (\boldsymbol{F}^\top \boldsymbol{F}\boldsymbol{K}_Q + n\lambda_{Q,n}\mathbf{I})^{-1}\boldsymbol{F}^\top \boldsymbol{E}\boldsymbol{r}\,,$$

where $\boldsymbol{r} = (R_1, \dots, R_n)^\top$ *and the matrices* $\boldsymbol{K}_Q, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{C}_2, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F}$ *are defined as follows:* $\boldsymbol{K}_h \in \mathbb{R}^{n\times n}$ *is defined as* $[\boldsymbol{K}_h]_{ij} = \kappa(Z_i, Z_j)$, $1 \le i, j \le n$, *and* $\boldsymbol{K}_Q \in \mathbb{R}^{2n\times 2n}$ *is defined as* $[\boldsymbol{K}_Q]_{ij} = \kappa(\tilde{Z}_i, \tilde{Z}_j)$, $1 \le i, j \le 2n$. *Let* $\boldsymbol{C}_1 = \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix}$ *and* $\boldsymbol{C}_2 = \begin{pmatrix} \mathbf{0}_{n\times n} & \mathbf{I}_{n\times n} \end{pmatrix}$. *Denote* $\boldsymbol{D} = \boldsymbol{C}_1 - \gamma \boldsymbol{C}_2$, $\boldsymbol{E} = \boldsymbol{K}_h(\boldsymbol{K}_h + n\lambda_{h,n}\mathbf{I})^{-1}$, $\boldsymbol{F} = \boldsymbol{C}_1 - \gamma \boldsymbol{E}\boldsymbol{C}_2$, $\boldsymbol{B} = \boldsymbol{K}_h(\boldsymbol{K}_h + n\lambda_{h,n}\mathbf{I})^{-1} - \mathbf{I}$, *and* $\boldsymbol{C} = \boldsymbol{D}^\top \boldsymbol{D} - \gamma^2(\boldsymbol{B}\boldsymbol{C}_2)^\top(\boldsymbol{B}\boldsymbol{C}_2)$.

Proof See Appendix A. ■

5. Theoretical Analysis

In this section, we analyze the statistical properties of REG-LSPI and provide a finite-sample upper bound on the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$. Here, π_K is the policy greedy w.r.t. $\hat{Q}^{(K-1)}$ and ρ is the performance evaluation measure. The distribution ρ is chosen by the user and is often different from the sampling distribution ν.

Our study has two main parts. First, we analyze the policy evaluation error of REG-LSTD in Section 5.1. We suppose that given any policy π, we obtain $\hat{Q}$ by solving (15)-(16) with π_k in these equations being replaced by π. Theorem 11 provides an upper bound on the Bellman error $\|\hat{Q} - T^\pi \hat{Q}\|_\nu$. We discuss the optimality of this upper bound for policy evaluation for some general classes of function spaces. We show that the result is not only optimal in its convergence rate, but also in its dependence on $J(Q^\pi)$. After that in Section 5.2, we show how the Bellman errors of the policy evaluation procedure propagate through the API procedure (Theorem 13). The main result of this paper, which is an upper bound on the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$, is stated as Theorem 14 in Section 5.3, followed by its discussion. We compare this work's statistical guarantee with some other papers' in Section 5.3.1.

To analyze the statistical performance of the REG-LSPI procedure, we make the following assumptions. We discuss their implications and the possible relaxations after stating each of them.

Assumption A1 (MDP Regularity) The set of states $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. The random immediate rewards $R_t \sim \mathcal{R}(\cdot|X_t, A_t)$ $(t = 1, 2, \dots)$ as well as the expected immediate rewards $r(x, a)$ are uniformly bounded by $R_{\max}$, i.e., $|R_t| \le R_{\max}$ $(t = 1, 2, \dots)$ and $\|r\|_\infty \le R_{\max}$.

Even though the algorithms were presented for a general measurable state space $\mathcal{X}$, the theoretical results are stated for the problems whose state space is a compact subset of $\mathbb{R}^d$. Generalizing Assumption A1 to other state spaces should be possible under certain regularity conditions. One example could be any Polish space, i.e., separable completely

metrizable topological space. Nevertheless, we do not investigate such generalizations here. The boundedness of the rewards is a reasonable assumption that can be replaced by a more relaxed condition such as its sub-Gaussianity (Vershynin, 2012; van de Geer, 2000). This relaxation, however, increases the technicality of the proofs without adding much to the intuition. We remark on the compactness assumption after stating Assumption A4.

Assumption A2 (Sampling) At iteration k of REG-LSPI (for $k = 0, \dots, K-1$), n fresh independent and identically distributed (i.i.d.) samples are drawn from distribution $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$, i.e., $\mathcal{D}_n^{(k)} = \left\{ \left(Z_t^{(k)}, R_t^{(k)}, X_t'^{(k)} \right) \right\}_{t=1}^n$ with $Z_t^{(k)} = (X_t^{(k)}, A_t^{(k)}) \overset{\text{i.i.d.}}{\sim} \nu$ and $X_t'^{(k)} \sim P(\cdot | X_t^{(k)}, A_t^{(k)})$.

The i.i.d. requirement of Assumption A2 is primarily used to simplify the proofs. With much extra effort, these results can be extended to the case when the data samples belong to a single trajectory generated by a fixed policy. In the single trajectory scenario, samples are not independent anymore, but under certain conditions on the Markov process, the process (X_t, A_t) gradually "forgets" its past. One way to quantify this forgetting is through mixing processes. For these processes, tools such as the *independent blocks* technique (Yu, 1994; Doukhan, 1994) or information theoretical inequalities (Samson, 2000) can be used to carry on the analysis—as have been done by Antos et al. (2008b) in the API context, by Farahmand and Szepesvári (2012) for analyzing the regularized regression problem, and by Farahmand and Szepesvári (2011) in the context of model selection for RL problems.

It is worthwhile to emphasize that we do not require that the distribution ν to be known. The sampling distribution is also generally different from the distribution induced by the target policy π_k. For example, it might be generated by drawing state samples from a given $\nu_{\mathcal{X}}$ and choosing actions according to a behavior policy π_b, which is different from the policy being evaluated. So we are in the off-policy sampling setting. Moreover, changing ν at each iteration based on the previous iterations is a possibility with potential practical benefits, which has theoretical justifications in the context of imitation learning (Ross et al., 2011). For simplicity of the analysis, however, we assume that ν is fixed in all iterations. Finally, we note that the proofs work fine if we reuse the same data sets in all iterations. We comment on it later after the proof of Theorem 11 in Appendix B.

Assumption A3 (Regularizer) Define two regularization functionals $J : B(\mathcal{X}) \to \mathbb{R}$ and $J : B(\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$ that are pseudo-norms on $\mathcal{F}$ and $\mathcal{F}^{|\mathcal{A}|}$, respectively.[12] For all $Q \in \mathcal{F}^{|\mathcal{A}|}$ and $a \in \mathcal{A}$, we have $J(Q(\cdot, a)) \le J(Q)$.

The regularizer $J(Q)$ measures the complexity of an action-value function Q. The functions that are more complex have larger values of $J(Q)$. We also need to define a related regularizer for value functions $Q(\cdot, a)$ $(a \in \mathcal{A})$. The latter regularizer is not explicitly used in the algorithm, and is only used in the analysis. This assumption imposes some mild restrictions on these regularization functionals. The condition that the regularizers be pseudo-norms is satisfied by many commonly-used regularizers such as the Sobolev norms,

12. Note that here we are slightly abusing the notations as the same symbol is used for the regularizer over both $B(\mathcal{X})$ and $B(\mathcal{X} \times \mathcal{A})$. However, this should not cause any confusion since in any specific expression the identity of the regularizer should always be clear from the context.

the RKHS norms, and the l_2-regularizer defined in Section 4.2.1 with a positive semi-definite choice of matrix $\mathbf{\Psi}$. Moreover, the condition $J(Q(\cdot, a)) \leq J(Q)$ essentially states that the complexity of Q should upper bound the complexity of $Q(\cdot, a)$ for all $a \in \mathcal{A}$. If the regularizer $J : B(\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$ is derived from a regularizer $J' : B(\mathcal{X}) \to \mathbb{R}$ through $J(Q) = \|(J'(Q(\cdot, a))_{a \in \mathcal{A}}\|_p$ for some $p \in [1, \infty]$, then J will satisfy the second part of the assumption. From a computational perspective, a natural choice for RKHS is to choose $p = 2$ and to define $J^2(Q) = \sum_{a \in \mathcal{A}} \|Q(\cdot, a)\|^2_{\mathcal{H}}$ for $\mathcal{H}$ being the RKHS defined on $\mathcal{X}$.

Assumption A4 (Capacity of Function Space) For $R > 0$, let $\mathcal{F}_R = \{f \in \mathcal{F} : J(f) \leq R\}$. There exist constants $C > 0$ and $0 < \alpha < 1$ such that for any $u, R > 0$ the following metric entropy condition is satisfied:

$$\log \mathcal{N}_\infty(u, \mathcal{F}_R) \leq C \left(\frac{R}{u}\right)^{2\alpha}.$$

This assumption characterizes the capacity of the ball with radius R in $\mathcal{F}$. The value of α is an essential quantity in our upper bounds. The metric entropy is precisely defined in Appendix G, but roughly speaking it is the logarithm of the minimum number of balls with radius u that are required to completely cover a ball with radius R in $\mathcal{F}$. This is a measure of complexity of a function space as it is more difficult to estimate a function when the metric entropy grows fast when u decreases. As a simple example, when the function space is finite, we effectively need to have good estimate of $|\mathcal{F}|$ functions in order not to choose the wrong one. In this case, $\mathcal{N}_\infty(u, \mathcal{F}_R)$ can be replaced by $|\mathcal{F}|$, so $\alpha = 0$ and $C = \log |\mathcal{F}|$. When the state space $\mathcal{X}$ is finite and all functions are bounded by $Q_{\max}$, we have $\log \mathcal{N}_\infty(u, \mathcal{F}_R) \leq \log \mathcal{N}_\infty(u, \mathcal{F}) = |\mathcal{X}| \log(\frac{2Q_{\max}}{u})$. This shows that the metric entropy for problems with finite state spaces grows much slower than what we consider here. Assumption A4 is suitable for large function spaces and is indeed satisfied for the Sobolev spaces and various RKHS. Refer to van de Geer (2000); Zhou (2002, 2003); Steinwart and Christmann (2008) for many examples.

An alternative assumption would be to have a similar metric entropy for the balls in $\mathcal{F}^{|\mathcal{A}|}$ (instead of $\mathcal{F}$). This would slightly change a few steps of the proofs, but leave the results essentially the same. Moreover, it makes the requirement that $J(Q(\cdot, a)) \leq J(Q)$ in Assumption A3 unnecessary. Nevertheless, as results on the capacity of $\mathcal{F}$ is more common in the statistical learning theory literature, we stick to the combination of Assumptions A3 and A4.

The metric entropy here is defined w.r.t. the supremum norm. All proofs, except that of Lemma 23, only require the same bound to hold when the supremum norm is replaced by the more relaxed empirical L_2-norm, i.e., those results require that there exist constants $C > 0$ and $0 < \alpha < 1$ such that for any $u, R > 0$ and all $x_1, \ldots, x_n \in \mathcal{X}$, we have $\log \mathcal{N}_2(u, \mathcal{F}_R, x_{1:n}) \leq C \left(\frac{R}{u}\right)^{2\alpha}$. Of course, the metric entropy w.r.t. the supremum norm implies the one with the empirical norm. It is an interesting question to relax the supremum norm assumption in Lemma 23.

We can now remark on the requirement that $\mathcal{X}$ is compact (Assumption A1). We stated that requirement mainly because most of the metric entropy results in the literature are for compact spaces (one exception is Theorem 7.34 of Steinwart and Christmann (2008), which relaxes the compactness requirement by adding some assumptions on the tail of $\nu_{\mathcal{X}}$ on $\mathcal{X}$).

So we could remove the compactness requirement from Assumption A1 and implicitly let Assumption A4 satisfy it, but we preferred to be explicit about it at the cost of a bit of redundancy in our set of assumptions.

Assumption A5 (Function Space Boundedness) The subset $\mathcal{F}^{|\mathcal{A}|} \subset B(\mathcal{X} \times \mathcal{A}; Q_{\max})$ is a separable and complete Carathéodory set with $R_{\max} \leq Q_{\max} < \infty$.

Assumption A5 requires all the functions in $\mathcal{F}^{|\mathcal{A}|}$ to be bounded so that the solutions of optimization problems (15)-(16) stay bounded. If they are not, they should be truncated, and thus, the truncation argument should be used in the analysis, see e.g., the proof of Theorem 21.1 of Györfi et al. (2002). The truncation argument does not change the final result, but complicates the proof at several places, so we stick to the above assumption to avoid unnecessary clutter. Moreover, in order to avoid the measurability issues resulting from taking supremum over an uncountable function space $\mathcal{F}^{|\mathcal{A}|}$, we require the space to be a separable and complete Carathéodory set (cf. Section 7.3 of Steinwart and Christmann 2008).

Assumption A6 (Function Approximation Property) The action-value function of any policy π belongs to $\mathcal{F}^{|\mathcal{A}|}$, i.e., $Q^\pi \in \mathcal{F}^{|\mathcal{A}|}$.

This "no function approximation error" assumption is standard in analyzing regularization-based nonparametric methods. This assumption is realistic and is satisfied for rich function spaces such as RKHS defined by universal kernels, e.g., Gaussian or exponential kernels (Section 4.6 of Steinwart and Christmann 2008). On the other hand, if the space is not large enough, we might have function approximation error. The behavior of the function approximation error for certain classes of "small" RKHS has been discussed by Smale and Zhou (2003); Steinwart and Christmann (2008). We stick to this assumption to simplify many key steps in the proofs.

Assumption A7 (Expansion of Smoothness) For all $Q \in \mathcal{F}^{|\mathcal{A}|}$, there exist constants $0 \leq L_R, L_P < \infty$, depending only on the MDP and $\mathcal{F}^{|\mathcal{A}|}$, such that for policy π,

$$J(T^\pi Q) \leq L_R + \gamma L_P J(Q).$$

We require that the complexity of $T^\pi Q$ to be comparable to the complexity of Q itself. In other words, we require that if Q is smooth according to the regularizer J of a function space $\mathcal{F}^{|\mathcal{A}|}$, it stays smooth after the application of the Bellman operator. We believe that this is a reasonable assumption for many classes of MDPs with "sufficient" stochasticity and when $\mathcal{F}^{|\mathcal{A}|}$ is rich enough. The intuition is that if the Bellman operator has a "smoothing" effect, the norm of $T^\pi Q$ does not blow up and the function can still be represented well within $\mathcal{F}^{|\mathcal{A}|}$. Proposition 25 in Appendix F presents the conditions that for the so-called *convolutional* MDPs, Assumption A7 is satisfied. Briefly speaking, the conditions are 1) the transition probability kernel should have a finite gain (in the control-theoretic sense) in its frequency response, and 2) the reward function should be smooth according to the regularizer J. Of course, this is only an example of the class of problems for which this assumption holds.

5.1 Policy Evaluation Error

In this section, we focus on the k^{th} iteration of REG-LSPI. To simplify the notation, we use $\mathcal{D}_n = \{(Z_t, R_t, X'_t)\}_{t=1}^n$ to refer to $\mathcal{D}_n^{(k)}$. The policy π_k depends on data used in the earlier iterations, but since we use independent set of samples $\mathcal{D}_n^{(k)}$ for the k^{th} iteration and π_k is independent of $\mathcal{D}_n^{(k)}$, we can safely ignore the randomness of π_k by working on the probability space obtained by conditioning on $\mathcal{D}_n^{(0)}, \dots, \mathcal{D}_n^{(k-1)}$, i.e., the probability space used in the k^{th} iteration is $(\Omega, \sigma_\Omega, \mathbb{P}_k)$ with $\mathbb{P}_k = \mathbb{P}\left\{\cdot \middle| \mathcal{D}_n^{(0)}, \dots, \mathcal{D}_n^{(k-1)}\right\}$. In order to avoid clutter, we do not use the conditional probability symbol. In the rest of this section, π refers to a $\sigma(\mathcal{D}_n^{(0)}, \dots, \mathcal{D}_n^{(k-1)})$-measurable policy and is independent of $\mathcal{D}_n$; $\hat{Q}$ and $\hat{h}_n(Q) = \hat{h}_n(\cdot; Q)$ refer to the solution to (15)-(16) when π, $\lambda_{h,n}$, and $\lambda_{Q,n}$ replace π_k, $\lambda_{h,n}^{(k)}$, and $\lambda_{Q,n}^{(k)}$ in that set of equations, respectively.

The following theorem is the main result of this section and provides an upper bound on the statistical behavior of the policy evaluation procedure REG-LSTD.

Theorem 11 (Policy Evaluation) *For any fixed policy π, let $\hat{Q}$ be the solution to the optimization problem (15)-(16) with the choice of*

$$\lambda_{h,n} = \lambda_{Q,n} = \left[\frac{1}{n\, J^2(Q^\pi)}\right]^{\frac{1}{1+\alpha}}.$$

If Assumptions A1–A7 hold, there exists $c(\delta) > 0$ such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have

$$\left\|\hat{Q} - T^\pi \hat{Q}\right\|_\nu^2 \le c(\delta)\, n^{-\frac{1}{1+\alpha}},$$

with probability at least $1 - \delta$. Here $c(\delta)$ is equal to

$$c(\delta) = c_1 \left(1 + (\gamma L_P)^2\right) J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln(1/\delta) + c_2 \left(L_R^{\frac{2\alpha}{1+\alpha}} + \frac{L_R^2}{[J(Q^\pi)]^{\frac{2}{1+\alpha}}}\right),$$

for some constants $c_1, c_2 > 0$.

Theorem 11, which is proven in Appendix B, indicates how the number of samples and the difficulty of the problem as characterized by $J(Q^\pi)$, L_P, and L_R influence the policy evaluation error.[13]

This upper bound provides some insights about the behavior of the REG-LSTD algorithm. To begin with, it shows that under the specified conditions, REG-LSTD is a consistent algorithm: As the number of samples increases, the Bellman error decreases and asymptotically converges to zero. This is due to the use of a nonparametric function space and the proper control of its complexity through regularization. A parametric function space, e.g., a linear function approximator with a fixed number of features, does not generally have a similar guarantee unless the value function happens to belong to the span of the features. Achieving consistency for parametric function spaces requires careful choice

13. Without loss of generality and for simplicity we assumed that $J(Q^\pi) > 0$.

of features, and might be difficult. On the other hand, a rich enough nonparametric function space, for example one defined by a universal kernel (cf. Assumption A6), ensures the consistency of the policy evaluation algorithm.

This theorem, however, is much more powerful than a consistency result as it provides a finite-sample upper bound guarantee for the error, too. If the parameters of the REG-LSTD algorithm are selected properly, one may achieve the sample complexity upper bound of $O(n^{-1/(1+\alpha)})$. For the case of the Sobolev space $\mathbb{W}^k(\mathcal{X})$ with $\mathcal{X}$ being an open Euclidean ball in $\mathbb{R}^d$ and $k > d/2$, one may choose $\alpha = d/2k$ to obtain the error upper bound of $O(n^{-d/(2k+d)})$.[14]

To study the upper bound a bit closer, let us focus on the special case of $\gamma = 0$. For this choice of the discount factor, $T^\pi Q$ is equal to r^π and $(\hat{T}^\pi Q)(X_i, A_i)$ is equal to R_i. One can see that the policy evaluation problem becomes a regression problem with the regression function r^π. The guarantee of this theorem would be then on $\|\hat{Q} - T^\pi \hat{Q}\|_\nu^2 = \|\hat{Q} - r^\pi\|_\nu^2$, which is the usual squared error in the regression literature. Hence we reduced a regression problem to a policy evaluation problem. Because of this reduction, any lower bound on the regression would also be a lower bound on the policy evaluation problem.

It is well-known that the convergence rate of $n^{-d/(2k+d)}$ is asymptotically minimax optimal for the regression estimation for target functions belonging to the Sobolev space $\mathbb{W}^k(\mathcal{X})$ as well as some other smoothness classes with the k order of smoothness, cf. e.g., Nussbaum (1999) for the results for the Sobolev spaces, Stone (1982) for a closely related Hölder space $C^{p,\alpha}$, which with the choice of $k = p + \alpha$ ($k \in \mathbb{N}$ and $0 < \alpha \leq 1$) has the same rate, and Tsybakov (2009) for several results on minimax optimality of nonparametric estimators. More generally, the rate of $O(n^{-1/(1+\alpha)})$ is optimal too: For a regression function belonging to a function space $\mathcal{F}$ with a packing entropy in the same form as in the upper bound of Assumption A4, the rate $\Omega(n^{-1/(1+\alpha)})$ is its minimax lower bound (Yang and Barron, 1999), making the upper bound optimal. Comparing these lower bounds with the upper bound $O(n^{-1/(1+\alpha)})$ (or $O(n^{-d/(2k+d)})$ for the Sobolev space) of this theorem indicates that REG-LSTD algorithm has the optimal error rate as a function of the number of samples n, which is a remarkable result.

Furthermore, to understand the fine behavior of the upper bound, beyond the dependence of the rate on n and α, we focus on the multiplicative term $c(\delta)$. Again we consider the special case of regression estimation as it is the only case we have some known lower bounds. With the choice of $\gamma = 0$, we have $Q^\pi = r^\pi$, so $J(Q^\pi) = J(r^\pi)$. Moreover, since $T^\pi Q = r^\pi + 0\mathcal{P}^\pi Q = r^\pi$, we can choose $L_R = J(r^\pi)$ in Assumption A7. As a result $c(\delta) = c_1 \, J^{\frac{2\alpha}{1+\alpha}}(r^\pi) \ln(1/\delta)$ for a constant $c_1 > 0$. We are interested in studying the dependence of the upper bound on $J(r^\pi)$. We study its behavior when the function space is the Sobolev space $\mathbb{W}^k([0,1])$ and $J(\cdot)$ is the corresponding Sobolev space norm. We choose $\alpha = 1/2k$ to get $J^{\frac{2}{2k+1}}(r^\pi)$ dependence of $c(\delta)$. On the other hand, for the regression estimation problem within the subset $\mathcal{F}_1^{|\mathcal{A}|} = \left\{ Q(\cdot, a) \in \mathbb{W}^k([0,1]) \, : \, J(Q) \leq J(r^\pi), \forall a \in \mathcal{A} \right\}$ of this Sobolev space, the fine behavior of the asymptotic minimax rate is determined by

14. For examples of the metric entropy results for the Sobolev spaces, refer to Section A.5.6 alongside Lemma 6.21 of Steinwart and Christmann (2008), or Theorem 2.4 of van de Geer (2000) for $\mathcal{X} = [0,1]$ or Lemma 20.6 of Györfi et al. (2002) for $\mathcal{X} = [0,1]^d$. Also in this paper we use the notation $\mathbb{W}^k(\mathcal{X})$ to refer to $\mathbb{W}^{k,2}(\mathcal{X})$, the Sobolev space defined based on the L_2-norm of the weak derivatives.

the so-called Pinsker constant, whose dependence on J is in fact $J^{\frac{2}{2k+1}}(r^\pi)$, cf. e.g., Nussbaum (1999, 1985); Golubev and Nussbaum (1990), or Section 3.1 of Tsybakov (2009).[15] Therefore, not only the exponent of the rate is optimal for this function space, but also its multiplicative dependence on the smoothness $J(r^\pi)$ is optimal.

For function spaces other than this choice of Sobolev space (i.e., the general case of α), we are not aware of any refined lower bound that indicates the optimality of $J^{\frac{2\alpha}{1+\alpha}}(r^\pi)$. We note that some available upper bounds for regression with comparable assumptions on the metric entropy have the same dependence on $J(r^\pi)$, e.g., Steinwart et al. (2009)[16] or Farahmand and Szepesvári (2012), whose result is for the regression setting with exponential β-mixing input, but can also be shown for i.i.d. data. We conjecture that under our assumptions this dependence is optimal.

One may note that the proper selection of the regularization coefficients to achieve the optimal rate requires the knowledge of an unknown quantity $J(Q^\pi)$. This, however, is not a major concern as a proper model selection procedure finds parameters that result in a performance which is almost the same as the optimal performance. We comment on this issue in more detail in Section 6.

The proof of this theorem requires several auxiliary results, which are presented in the appendices, but the main idea behind the proof is as follows. Since $\|\hat{Q} - T^\pi \hat{Q}\|_\nu^2 \leq 2\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\|_\nu^2 + 2\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi \hat{Q}\|_\nu^2$, we may upper bound the Bellman error by upper bounding each term in the right-hand side (RHS). One can see that for a fixed Q, the optimization problem (15) essentially solves a regularized least-squares regression problem, which leads to small value of $\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi \hat{Q}\|_\nu$, when there are enough samples and under proper conditions. The relation of the optimization problem (16) with $\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\|_\nu$ is evident too. The difficulty, however, is that these two optimization problems are coupled: $\hat{h}_n(\cdot;\hat{Q})$ is a function of $\hat{Q}$ which itself is a function of $\hat{h}_n(\cdot;\hat{Q})$. Thus, Q appearing in (15) is not fixed, but is a random function $\hat{Q}$. The same is true for the other optimization problem as well. The coupling of the optimization problems makes the analysis more complicated than the usual supervised learning type of analysis. The dependencies between all the results that lead to the proof of Theorem 14 is depicted in Figure 2 in Appendix B.

In order to obtain fast convergence rates, we use concepts and techniques from the empirical process theory such as the peeling device, the chaining technique, and the modulus of continuity of the empirical process, cf. e.g., van de Geer (2000). By focusing on the behavior of the empirical process over local subsets of the function space, these techniques allow us to study the deviations of the process in a more refined way compared to a global approach that studies the supremum of the empirical process in the whole function space. These techniques are crucial to obtain a fast rate for large function spaces. We discuss them in more detail as we proceed in the proofs.

15. The Pinsker constant determines the effect of the noise variance too. We do not present such information in our bounds. Also note that most aforementioned results, except Golubev and Nussbaum (1990), consider a normal noise model, which is different from our bounded noise.

16. This is obtained by using Corollary 3 of Steinwart et al. (2009) after substituting $A_2(\lambda)$ by its upper bound $\lambda \|f\|_{\mathcal{H}}^2$, which is valid whenever $f^* \in \mathcal{H}$, as is in our case. This result can be used after one converts the metric entropy condition to the condition on the decay rate of eigenvalues of a certain integral operator.

We mentioned earlier that one can actually reuse a single data set in all iterations. To keep the presentation more clear, we keep the current setup. The reason behind this can be explained better after the proof of Theorem 11. But note that from the convergence-rate point of view, the difference between reusing data or not is insignificant. If we have a batch of data with size n and we divide it into K chunks and only use one chunk per iteration of API, the rate would be $O((\frac{n}{K})^{-\frac{1}{1+\alpha}})$. For finite K, or slowly growing K, this is essentially the same as $O(n^{-\frac{1}{1+\alpha}})$.

5.2 Error Propagation in API

Consider an API algorithm that generates the sequence $\hat{Q}^{(0)} \to \pi_1 \to \hat{Q}^{(1)} \to \pi_2 \to \cdots \to \hat{Q}^{(K-1)} \to \pi_K$, where π_k is the greedy policy w.r.t. $\hat{Q}^{(k-1)}$ and $\hat{Q}^{(k)}$ is the approximate action-value function for policy π_k. For the sequence $(\hat{Q}^{(k)})_{k=0}^{K-1}$, denote the Bellman Residual (BR) of the k^{th} action-value function by

$$\varepsilon_k^{\text{BR}} = \hat{Q}^{(k)} - T^{\pi_k}\hat{Q}^{(k)}. \tag{31}$$

The goal of this section is to study the effect of the ν-weighted L_2-norm of the Bellman residual sequence $(\varepsilon_k^{\text{BR}})_{k=0}^{K-1}$ on the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$ of the resulting policy π_K. Because of the dynamical nature of the MDP, the performance loss $\|Q^* - Q^{\pi_K}\|_{p,\rho}$ depends on the difference between the sampling distribution ν and the future state-action distribution in the form of $\rho P^{\pi_1} P^{\pi_2} \cdots$. The precise form of this dependence is formalized in Theorem 13, which is a slight modification of a result by Farahmand et al. (2010).[17]

Before stating the results, we define the following *concentrability* coefficients that are used in a change of measure argument, see e.g., Munos (2007); Antos et al. (2008b); Farahmand et al. (2010).

Definition 12 (Expected Concentrability of Future State-Action Distributions) *Given the distributions $\rho, \nu \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$, an integer $m \geq 0$, and an arbitrary sequence of stationary policies $(\pi_m)_{m \geq 1}$, let $\rho P^{\pi_1} P^{\pi_2} \cdots P^{\pi_m} \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$ denote the future state-action distribution obtained when the first state-action is distributed according to ρ and then we follow the sequence of policies $(\pi_k)_{k=1}^m$. Define the following concentrability coefficients:*

$$c_{PI_1,\rho,\nu}(m_1, m_2; \pi) \triangleq \left(\mathbb{E}\left[\left| \frac{d\left(\rho (P^{\pi^*})^{m_1} (P^{\pi})^{m_2}\right)}{d\nu}(X, A) \right|^2 \right] \right)^{\frac{1}{2}},$$

with $(X, A) \sim \nu$. If the future state-action distribution $\rho(P^{\pi^*})^{m_1}(P^{\pi})^{m_2}$ is not absolutely continuous w.r.t. ν, then we take $c_{PI_1,\rho,\nu}(m_1, m_2; \pi) = \infty$.

In order to compactly present our results, we define the following notation:

$$a_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}. \qquad (0 \leq k < K) \tag{32}$$

17. The difference of these two results is in the way the norm of functions from the space $\mathcal{F}^{|\mathcal{A}|}$ is defined, which in turn corresponds to whether the distributions ν and ρ are defined over the state space $\mathcal{X}$, as Farahmand et al. (2010) defined, or over the state-action space $\mathcal{X} \times \mathcal{A}$, as we define here. These differences do not change the general form of the proof. See Theorem 3.2 in Chapter 3 of Farahmand (2011b) for the proof of the current result.

Theorem 13 (Error Propagation for API—Theorem 3 of Farahmand et al. 2010) *Let $p \geq 1$ be a real number, K be a positive integer, and $Q_{max} \leq \frac{R_{max}}{1-\gamma}$. Then for any sequence $(\hat{Q}^{(k)})_{k=0}^{K-1} \subset B(\mathcal{X} \times \mathcal{A}, Q_{max})$ and the corresponding sequence $(\varepsilon_k^{BR})_{k=0}^{K-1}$ defined in (31), we have*

$$\|Q^* - Q^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} \left[\inf_{r \in [0,1]} C_{PI,\rho,\nu}^{\frac{1}{2p}}(K;r) \mathcal{E}^{\frac{1}{2p}}(\varepsilon_0^{BR}, \dots, \varepsilon_{K-1}^{BR}; r) + \gamma^{\frac{K}{p}-1} R_{max} \right],$$

where $\mathcal{E}(\varepsilon_0^{BR}, \dots, \varepsilon_{K-1}^{BR}; r) = \sum_{k=0}^{K-1} a_k^{2r} \left\|\varepsilon_k^{BR}\right\|_{2p,\nu}^{2p}$ and

$$C_{PI,\rho,\nu}(K;r) = \left(\frac{1-\gamma}{2}\right)^2 \sup_{\pi'_0,\dots,\pi'_K} \sum_{k=0}^{K-1} a_k^{2(1-r)} \left[\sum_{m \geq 0} \gamma^m \Big(c_{PI_1,\rho,\nu}(K-k-1, m+1; \pi'_{k+1}) + c_{PI_1,\rho,\nu}(K-k, m; \pi'_k) \Big) \right]^2.$$

For better understanding of the intuition behind the error propagation results in general, refer to Munos (2007); Antos et al. (2008b); Farahmand et al. (2010). The significance of this particular theorem and the ways it improves previous similar error propagation results such as that of Antos et al. (2008b) (for API) and Munos (2007) (for AVI) is thoroughly discussed by Farahmand et al. (2010). We briefly comment on it in Section 5.3.

5.3 Performance Loss of REG-LSPI

In this section, we use the error propagation result (Theorem 13 in Section 5.2) together with the upper bound on the policy evaluation error (Theorem 11 in Section 5.1) to derive an upper bound on the performance loss $\|Q^* - Q^{\pi_K}\|_{1,\rho}$ of REG-LSPI. This is the main theoretical result of this work. Before stating the theorem, let us denote $\hat{\Pi}(\mathcal{F}^{|\mathcal{A}|})$ as the set of all policies that are greedy w.r.t. a member of $\mathcal{F}^{|\mathcal{A}|}$, i.e., $\hat{\Pi}(\mathcal{F}^{|\mathcal{A}|}) = \{\hat{\pi}(\cdot; Q) : Q \in \mathcal{F}^{|\mathcal{A}|}\}$.

Theorem 14 *Let $(\hat{Q}^{(k)})_{k=0}^{K-1}$ be the solutions of the optimization problem (15)-(16) with the choice of*

$$\lambda_{h,n}^{(k)} = \lambda_{Q,n}^{(k)} = \left[\frac{1}{n\, J^2(Q^{\pi_k})} \right]^{\frac{1}{1+\alpha}}.$$

Let Assumptions A1–A5 hold; Assumptions A6 and A7 hold for any $\pi \in \hat{\Pi}(\mathcal{F}^{|\mathcal{A}|})$, and $\inf_{r \in [0,1]} C_{PI,\rho,\nu}(K;r) < \infty$. Then there exists $C_{LSPI}(\delta, K; \rho, \nu)$ such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have

$$\|Q^* - Q^{\pi_K}\|_{1,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} \left[C_{LSPI}(\delta, K; \rho, \nu) n^{-\frac{1}{2(1+\alpha)}} + \gamma^{K-1} R_{max} \right],$$

with probability at least $1 - \delta$.

In this theorem, the function $C_{\text{LSPI}}(\delta, K; \rho, \nu) = C_{\text{LSPI}}(\delta, K; \rho, \nu; L_R, L_P, \alpha, \beta, \gamma)$ is

$$C_{\text{LSPI}}(\delta, K; \rho, \nu; L_R, L_P, \alpha, \beta, \gamma) = C_{\text{I}}^{\frac{1}{2}}(\delta) \inf_{r \in [0,1]} \left\{ \left(\frac{1-\gamma}{1-\gamma^{K+1}} \right)^r \sqrt{\frac{1-(\gamma^{2r})^K}{1-\gamma^{2r}}} C_{\text{PI},\rho,\nu}^{\frac{1}{2}}(K;r) \right\},$$

with $C_{\mathrm{I}}(\delta)$ being defined as

$$C_{\mathrm{I}}(\delta) = \sup_{\pi \in \hat{\Pi}(\mathcal{F}^{|\mathcal{A}|})} \left[c_1 \left(1 + (\gamma L_P)^2\right) J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln\left(\frac{K}{\delta}\right) + c_2 \left(L_R^{\frac{2\alpha}{1+\alpha}} + \frac{L_R^2}{[J(Q^\pi)]^{\frac{2}{1+\alpha}}} \right) \right],$$

in which $c_1, c_2 > 0$ are universal constants.

Proof Fix $0 < \delta < 1$. For each iteration $k = 0, \ldots, K-1$, invoke Theorem 11 with the confidence parameter δ/K and take the supremum over all policies to upper bound the Bellman residual error $\|\varepsilon_k^{\mathrm{BR}}\|_\nu$ as

$$\left\| \hat{Q}^{(k)} - T^{\pi_k} \hat{Q}^{(k)} \right\|_\nu^2 \leq \underbrace{\sup_{\pi \in \hat{\Pi}(\mathcal{F}^{|\mathcal{A}|})} c\left(J(Q^\pi), L_R, L_P, \alpha, \beta, \gamma, \frac{\delta}{K}\right)}_{\triangleq c'} n^{-\frac{1}{1+\alpha}},$$

which holds with probability at least $1 - \frac{\delta}{K}$. Here $c(\cdot)$ is defined as in Theorem 11. For any $r \in [0, 1]$, we have

$$\begin{aligned} \mathcal{E}(\varepsilon_0^{\mathrm{BR}}, \ldots, \varepsilon_{K-1}^{\mathrm{BR}}; r) &= \sum_{k=0}^{K-1} a_k^{2r} \left\|\varepsilon_k^{\mathrm{BR}}\right\|_\nu^2 \leq c'\, n^{-\frac{1}{1+\alpha}} \sum_{k=0}^{K-1} a_k^{2r} \\ &= c'\, n^{-\frac{1}{1+\alpha}} \left(\frac{1-\gamma}{1-\gamma^{K+1}}\right)^{2r} \frac{1-(\gamma^{2r})^K}{1-\gamma^{2r}}, \end{aligned}$$

where we used the definition of a_k (32). We then apply Theorem 13 with the choice of $p = 1$ to get that with probability at least $1 - \delta$, we have

$$\|Q^* - Q^{\pi_K}\|_{1,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} \left[C_{\mathrm{LSPI}}(\rho, \nu; K) n^{-\frac{1}{1+\alpha}} + \gamma^{K-1} R_{\max} \right].$$

Here

$$C_{\mathrm{LSPI}}(\rho, \nu; K) = \sqrt{\sup_{\pi \in \hat{\Pi}(\mathcal{F}^{|\mathcal{A}|})} c\left(J(Q^\pi), L_R, L_P, \alpha, \gamma, \frac{\delta}{K}\right)} \inf_{r \in [0,1]} \left\{ \left(\frac{1-\gamma}{1-\gamma^{K+1}}\right)^r \sqrt{\frac{1-(\gamma^{2r})^K}{1-\gamma^{2r}}} C_{\mathrm{PI},\rho,\nu}^{\frac{1}{2}}(K; r) \right\}.$$

■

Theorem 14 upper bounds the performance loss and relates it to the number of samples n, the capacity of the function space quantified by α, the number of iterations K, the concentrability coefficients, and some other properties of the MDP such as L_R, L_P, and γ.

This theorem indicates that the behavior of the upper bound as a function of the number of samples is $O(n^{-\frac{1}{2(1+\alpha)}})$. This upper bound is notable because of its minimax optimality, as discussed in detail after Theorem 11.

The term C_{LSPI} has two main components. The first is $C_{\mathrm{PI},\rho,\nu}(\cdot; r)$, which describes the effect of the sampling distribution ν and the evaluation distribution ρ, as well as the

transition probability kernel of the MDP itself on the performance loss. This term has been thoroughly discussed by Farahmand et al. (2010), but briefly speaking it indicates that ν and ρ affect the performance through a weighted summation of $c_{\mathrm{PI}_1,\rho,\nu}$ (Definition 12). The concentrability coefficients $c_{\mathrm{PI}_1,\rho,\nu}$ is defined as the square root of the *expected* squared Radon-Nikodym of the future state-action distributions starting from ρ w.r.t. the sampling distribution ν. This may be much tighter compared to the previous results (e.g., Antos et al. 2008b) that depend on the *supremum* of the Radon-Nikodym derivative. One may also notice that Theorem 13 actually provides a stronger result than what is reported in Theorem 14: The effect of errors at earlier iterations on the performance loss is geometrically decayed. So one may potentially use a fewer number of samples in the earlier iterations of REG-LSPI (or any other API algorithm) to get the same guarantee on the performance loss. We ignore this effect to simplify the result.

The other important term is C_{I}, which mainly describes the effect of L_R, L_P, and $\sup_{\pi\in\hat{\Pi}(\mathcal{F}^{|\mathcal{A}|})} J(Q^\pi)$ on the performance loss. These quantities depend on the MDP, as well as the function space $\mathcal{F}^{|\mathcal{A}|}$. If the function space is "matched" with the MDP, these quantities would be small, otherwise they may even be infinity.

Note that C_{I} provides an upper bound on the constant in front of REG-LSTD procedure by taking supremum over all policies in $\hat{\Pi}(\mathcal{F}^{|\mathcal{A}|})$. This might be a conservative estimate as the actual encountered policies are the rather restricted random sequence $\pi_0, \pi_1, \ldots, \pi_{K-1}$ generated by the REG-LSPI procedure. One might expect that as the sequence $\hat{Q}^{(k-1)}$ converge to a neighbourhood of Q^*, the value function Q^{π_k} of the greedy policy $\pi_k = \hat{\pi}(\cdot; \hat{Q}^{(k-1)})$, which is the policy being evaluated, converges to a neighbourhood of Q^* too. Thus with certain assumptions, one might be able to show that its smoothness $J(Q^{\pi_k})$, the quantity that appears in the upper bound of Theorem 11, belongs to a neighbourhood of $J(Q^*)$. If $J(Q^*)$ is small, the value of $J(Q^{\pi_k})$ in that neighbourhood can be smaller than $\sup_{\pi\in\hat{\Pi}(\mathcal{F}^{|\mathcal{A}|})} J(Q^\pi)$. We postpone the analysis of this finer structure of the problem to future work.

Finally we note that the optimality of the error bound for the policy evaluation task, as shown by Theorem 11, does not necessarily imply that the REG-LSPI algorithm has the optimal sample complexity rate for the corresponding RL/Planning problem as well. The reason is that it is possible to get close to the optimal policy, which is the ultimate goal in RL/Plannning, even though the estimate of the action-value function is still inaccurate. To act optimally, it is sufficient to have an action-value function whose greedy policy is the same as the optimal policy. This can happen even if there is some error in the estimated action-value function. This is called the *action-gap phenomenon* and has been analyzed in the reinforcement learning context by Farahmand (2011a).

5.3.1 Comparison with Similar Statistical Guarantees

Theorem 14 might be compared with the results of Antos et al. (2008b), who introduced a BRM-based API procedure and studied its statistical properties, Lazaric et al. (2012), who analyzed LSPI with linear function approximators, Ávila Pires and Szepesvári (2012), who studied a regularized variant of LSTD, and Ghavamzadeh et al. (2011), who analyzed the statistical properties of Lasso-TD. Although these results address different algorithms, comparing them with the results of this work is insightful.

We first focus on Antos et al. (2008b). Their simplified upper bound for $\|Q^* - Q^{\pi_K}\|_{1,\rho}$ is $C_{\rho,\nu}^{1/2}\sqrt{V_{\mathcal{F}}\log(n) + \ln(K/\delta)}\, n^{-1/4}$, in which $V_{\mathcal{F}}$ is the "effective" dimension of $\mathcal{F}$ and is defined based on the pseudo-dimension of sub-graphs of $\mathcal{F}$ and the so-called "VC-crossing dimension" of $\mathcal{F}$; and $C_{\rho,\nu}$ is a concentrability coefficient and plays a similar rule to our $C_{\mathrm{PI},\rho,\nu}(K;r)$. In contrast, our simplified upper bound is $C_{\mathrm{LSPI}}(\delta)n^{-\frac{1}{2(1+\alpha)}}$, in which $C_{\mathrm{LSPI}}(\delta)$ can roughly be factored into $C_{\mathrm{PI},\rho,\nu}^{\frac{1}{2}}(K;r)\, C_1(J(Q^\pi), L_R, L_P)\sqrt{\ln(K/\delta)}$.

One important difference between these two results is that Antos et al. (2008b) considered parametric function spaces, which have finite effective dimension $V_{\mathcal{F}}$, while this work considers nonparametric function spaces, which essentially are infinite dimensional. The way they use the parametric function space assumption is equivalent to assuming that $\log \mathcal{N}_1(u, \mathcal{F}, x_{1:n}) \leq V_{\mathcal{F}} \log(\frac{1}{u})$ as opposed to $\log \mathcal{N}_\infty(u, \mathcal{F}_B, x_{1:n}) \leq C\left(\frac{R}{u}\right)^{2\alpha}$ of Assumption A4. Our assumption lets us describe the capacity of infinite dimensional function spaces $\mathcal{F}$. Disregarding this crucial difference, one may also note that our upper bound's dependence on the number of samples (i.e., $O(n^{-\frac{1}{2(1+\alpha)}})$) is much faster than theirs (i.e., $O(n^{-1/4})$). This is more noticeable when we apply our result to a finite dimensional function space, which can be done by letting $\alpha \to 0$ at a certain rate, to recover the error upper bound of $n^{-1/2}$.[18] This improvement is mainly because of more advanced techniques used in our analysis, i.e., the relative deviation tail inequality and the peeling device in this work in contrast with the uniform deviation inequality of Antos et al. (2008b).

The other difference is in the definition of concentrability coefficients ($C_{\mathrm{PI},\rho,\nu}(K)$ vs. $C_{\rho,\nu}$). In Definition 12, we use the expectation of Radon-Nikodym derivative of two distributions while their definition uses the supremum of a similar quantity. This can be a significant improvement in the multiplicative constant of the upper bound. For more information regarding this improvement, which can be used to improve the result of Antos et al. (2008b) too, refer to Farahmand et al. (2010).

Lazaric et al. (2012) analyzed unregularized LSTD/LSPI specialized for linear function approximators with finite number of basis functions (parametric setting). Their rate of $O(n^{-1/2})$ for $\|V^* - V^{\pi_K}\|_{2,\rho}$ is faster than the rate in the work of Antos et al. (2008b), and is comparable to our rate for $\|Q^* - Q^{\pi_K}\|_{1,\rho}$ when $\alpha \to 0$. The difference of their work with ours is that they focus on a parametric class of function approximators as opposed to the nonparametric class in this work. Moreover, because they formulate the LSTD as a fixed-point problem, in contrast to this work and that of Antos et al. (2008b), their algorithm and results are only applicable to on-policy sampling scenario.

Ávila Pires and Szepesvári (2012) studied a regularized version of LSTD in the parametric setting that works for both on-policy and off-policy sampling. Beside the difference between the class of function spaces with this work (parametric vs. nonparametric), another algorithmic difference is that they only use a regularizer for the projected Bellman error term, similar to (16), as opposed to using regularizers in both terms of REG-LSTD (15)-(16) (cf. Section 4.1). Also the weight used in their loss function, the matrix M in their paper, is not necessarily the one induced by data. Their result indicates $O(n^{-1/2})$ for the projected Bellman error, which is comparable, though with some subtle differences, to Lazaric et al.

18. For problems with finite state space, we have $\log \mathcal{N}_\infty(u, \mathcal{F}_R) \leq |\mathcal{X}| \log(\frac{Q_{\max}}{u})$, so with a similar $\alpha \to 0$ argument, we get $O(n^{-1/2})$ error upper bound (disregarding the logarithmic terms).

(2012). It is remarkable that they separate the error bound analysis to deterministic and probabilistic parts. In the deterministic part, they use perturbation analysis to relate the loss to the error in the estimation of certain parameters used by the algorithms. In the probabilistic part, they provide upper bounds on the error in estimation of the parameters. We conjecture that their proof technique, even though simple and elegant, cannot easily be extended to provide the right convergence rate for large function spaces because the current analysis is based on a uniform bound on the error of a noisy matrix. Providing a tight uniform bound for a matrix (or operator) for large state spaces might be difficult or impossible to achieve.

Ghavamzadeh et al. (2011) analyzed Lasso-TD, a policy evaluation algorithm that uses linear function approximators and enforces sparsity by the l_1-regularization, and provided error upper bounds w.r.t. the empirical measure (or what they call Markov design). Their error upper bound is $O([\|w^*\|_1^2 \log(p)]^{1/4} n^{-1/4})$, where w^* is the weight vector describing the projection of Q^π onto the span of p basis functions. With some extra assumptions on the Grammian of the basis functions, they obtain faster rate of $O(\sqrt{\|w^*\|_0 \log(p)}\, n^{-1/2})$. These results indicate that by using the sparsity-inducing regularizer, the dependence of the error bound on the number of features becomes logarithmic.

We conjecture that if one uses REG-LSTD with a linear function space (similar to Section 4.2.1) with $J^2(h) = \|u\|_1$ and $J^2(Q) = \|w\|_1$, the current analysis leads to the error upper bound $O(\|w^*\|_1^{1/2} n^{-1/4})$ with a logarithmic dependence on p. This result might be obtained using Corollary 5 of Zhang (2002) as Assumption A4. To get a faster rate of $O(n^{-1/2})$, one should make extra assumptions on the Grammian—as was done by Ghavamzadeh et al. (2011). We should emphasize that even with the choice of linear function approximators and the l_1-regularization, REG-LSTD would not be the same algorithm as Lasso-TD since REG-LSTD uses regularization in both optimization problems (15)-(16). Also note that the error upper bound of Ghavamzadeh et al. (2011) is on the empirical norm $\|\cdot\|_{2,\mathcal{D}_n}$ as opposed to the norm $\|\cdot\|_{2,\nu}$, which is w.r.t. the measure ν. This means that their result does not provide a generalization upper bound on the quality of the estimated value function over the whole state space, but provides an upper bound only on the training data.

Comparing this work with its conference version (Farahmand et al., 2009b), we observe that the main difference in the theoretical guarantees is that the current results are for more general function spaces than the Sobolev spaces considered in the conference paper. Assumption A4 specifies the requirement on the capacity of the function space, which is satisfied not only by the Sobolev spaces (with the choice of $\alpha = d/2k$ for $\mathbb{W}^k(\mathcal{X})$ with $\mathcal{X}$ being an open Euclidean ball in $\mathbb{R}^d$ and $k > d/2$; cf. Section A.5.6 alongside Lemma 6.21 of Steinwart and Christmann (2008), or Theorem 2.4 of van de Geer (2000) for $\mathcal{X} = [0,1]$ or Lemma 20.6 of Györfi et al. (2002) for $\mathcal{X} = [0,1]^d$), but also many other large function spaces including several commonly-used RKHS.

6. Conclusion and Future Work

We introduced two regularization-based API algorithms, namely REG-LSPI and REG-BRM, to solve RL/Planning problems with large state spaces. Our formulation was general and could incorporate many types of function spaces and regularizers. We specifically showed how these algorithms can be implemented efficiently when the function space is the

span of a finite number of basis functions (parametric model) or an RKHS (nonparametric model).

We then focused on the statistical properties of REG-LSPI and provided its performance loss upper bound (Theorem 14). The error bound demonstrated the role of the sample size, the complexity of function space to which the action-value function belongs (quantified by its metric entropy in Assumption A4), and the intrinsic properties of the MDP such as the behavior of concentrability coefficients and the smoothness-expansion property of the Bellman operator (Definition 12 and Assumption A7). The result indicated that the dependence on the sample size for the task of policy evaluation is optimal.

This work (and its conference (Farahmand et al., 2009b) and the dissertation (Farahmand, 2011b) versions) alongside the work on the Regularized Fitted Q-Iteration algorithm (Farahmand et al., 2008, 2009a) are the first that address the statistical performance of a *regularized* RL algorithm. Nevertheless, there have been a few other work that also used regularization for RL/Planning problems, most often without analyzing their statistical properties.

Jung and Polani (2006) studied adding regularization to BRM, but their solution is restricted to deterministic problems. The main contribution of that work was the development of fast incremental algorithms using the *sparsification* technique. The l_1-regularization has been considered by Loth et al. (2007), who were similarly concerned with incremental implementations and computational efficiency. Xu et al. (2007) provided a kernel-based, but not regularized, formulation of LSPI. They used sparsification to provide basis functions for the LSTD procedure. Sparsification leads to a selection of only a subset of data points to be used as the basis functions, thus indirectly controls the complexity of the resulting function space. This should be contrasted with a regularization-based approach in which the regularizer interacts with the empirical loss to jointly determine the subset of the function space to which the estimate belongs.

Kolter and Ng (2009) formulated an l_1-regularization fixed-point formulation LSTD, which is called Lasso-TD by Ghavamzadeh et al. (2011), and provided LARS-like algorithm (Efron et al., 2004) to compute the solutions. Johns et al. (2010) considered the same fixed-point formulation and cast it as a linear complementarity problem. The statistical properties of this l_1-regularized fixed-point formulation is studied by Ghavamzadeh et al. (2011), as discussed earlier. Lasso-TD has a fixed-point formulation, which looks different from our coupled optimization formulation (15)-(16), but under on-policy sampling scenario, it is equivalent to a particular version of REG-LSTD: If we choose a fixed linear function approximator (parametric), use the l_1-norm in the projection optimization problem (15), but do not regularize optimization problem (16) (i.e., $\lambda_{Q,n} = 0$), we get Lasso-TD. Geist and Scherrer (2012) suggested a different algorithm where the projection is not regularized (i.e., $\lambda_{h,n} = 0$), but the optimization problem (16) is regularized with the l_1-norm of the parameter weights. The choice of only regularizing (16) is the same as the one in the algorithm introduced and analyzed by Ávila Pires and Szepesvári (2012), except that the latter work uses the l_2-norm. Hoffman et al. (2012) introduced an algorithm similar to that of Geist and Scherrer (2012) with the difference that the projection optimization (15) uses the l_2-norm (so it is a mixed l_1/l_2-regularized algorithm). All these algorithms are parametric. Several TD-based algorithms and their regularized variants are discussed in a survey by Dann et al. (2014).

Taylor and Parr (2009) unified several kernelized reinforcement learning algorithms, and showed the equivalence of kernelized value function approximators such as GPTD (Engel et al., 2005), the work of Xu et al. (2007), and a few other methods with a model-based reinforcement learning algorithm that has certain regularization on the transition kernel estimator, reward estimator, or both. Their result was obtained by considering two separate regularized regression problems: One that predicts the reward function given the current state and the other that predicts the next-state kernel values given the current-state ones. Their formulation is different from our formulation that is stated as a coupled optimization problem in an RKHS.

Similar to other kernel-based algorithms (e.g., SVMs, Gaussian Process Regressions, Splines, etc.), devising a computationally efficient implementation of REG-LSPI/BRM is important to ensure that it is a practical algorithm for large-scale problems. A naive implementation of these algorithms requires the computation time of $O(n^3K)$, which is prohibitive for large sample sizes. One possible workaround is to reduce the effective number of samples by the sparsification technique (Engel et al., 2005; Jung and Polani, 2006; Xu et al., 2007). The other is to use elegant vector-matrix multiplication methods, which are used in iterative methods for matrix inversion, such as those based on the Fast Multipole Methods (Beatson and Greengard, 1997) and the Fast Gauss Transform (Yang et al., 2004). These methods can reduce the computational cost of vector-matrix multiplication from $O(n^2)$ to $O(n \log n)$, which results in computation time of $O(n^2 K \log n)$ for REG-LSPI/BRM, at the cost of some small, but controlled, numerical error. Another possibility is to use stochastic gradient-like algorithms similar to the works of Liu et al. (2012); Qin et al. (2014). The use of stochastic gradient-like algorithms is especially appealing in the light of results such as Bottou and Bousquet (2008); Shalev-Shwartz and Srebro (2008). They analyze the tradeoff between the statistical error and the optimization error caused by the choice of optimization method. They show that one might achieve lower generalization error by using a faster stochastic gradient-like algorithm, which processes more data points less accurately, rather than a slower but more accurate optimization algorithm, which can only process fewer data points. Designing scalable optimization algorithms for REG-LSPI/BRM is a topic for future work.

An important issue in the successful application of any RL/Planning algorithm, including REG-LSPI and REG-BRM, is the proper choice of parameters. In REG-BRM and REG-LSTD we are faced with the choice of $\mathcal{F}^{|\mathcal{A}|}$ and the corresponding regularization parameters $\lambda_{Q,n}$ and $\lambda_{h,n}$. The proper choice of these parameters, however, depends on quantities that are not known, e.g., $J(Q^\pi)$ and the choice of $\mathcal{F}^{|\mathcal{A}|}$ that "matches" with the MDP. This problem in the RL/Planning context has been addressed by Farahmand and Szepesvári (2011). They introduced a complexity-regularization-based model selection algorithm that allows one to design adaptive algorithms: Algorithms that perform almost the same as the one with the prior knowledge of the best parameters.

Another important question is how to extend these algorithms to deal with continuous action MDPs. There are two challenges: Computational and statistical. The computational challenge is finding the greedy action at each state in the policy improvement step. In general, this is an intractable optimization problem, which cannot be solved exactly or even with any suboptimality guarantee. To analyze this inexact policy improvement some parts of the theory, especially the error propagation result, should be modified. Moreover, we also have a statistical challenge: One should specifically control the complexity of the

policy space as the complexity of $\{\max_{a\in\mathcal{A}} Q(\cdot,a) : Q \in \mathcal{F}^{|\mathcal{A}|}\}$ might be infinity even though $\mathcal{F}^{|\mathcal{A}|}$ has a finite complexity (Antos et al., 2008a). A properly modified algorithm might be similar to the continuous-action extension of Farahmand et al. (2015), an API algorithm that explicitly controls the complexity of the policy space.

Finally an open theoretical question is to characterize the properties of the MDP that determine the function space to which action-value function belong. A similar question is how the values of L_P and L_R in Assumption A7 are related to the intrinsic properties of the MDP. We partially addressed this question for the convolutional MDPs, but analysis of more general MDPs is remained to be done.

Acknowledgments

We thank the members of the Reinforcement Learning and Artificial Intelligence (RLAI) research group at the University of Alberta and the Reasoning and Learning Lab at McGill University for fruitful discussions. We thank the anonymous reviewers and the editor for their helpful comments and suggestions, which improved the quality of the paper. We gratefully acknowledge funding from the National Science and Engineering Research Council of Canada (NSERC) and Alberta Innovates Centre for Machine Learning (AICML). Shie Mannor was partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 306638 (SUPREL).

Proofs and Auxiliary Results

In these appendices, we first prove Theorem 10, which provides the closed-form solutions for REG-LSTD and REG-BRM when the function space is an RKHS (Appendix A). We then attend to the proof of Theorem 11 (Policy Evaluation error for REG-LSTD). The main body of the proof for Theorem 11 is in Appendix B. To increase the readability and flow, the proofs of some of the auxiliary and more technical results are postponed to Appendices C, D, and E.

More specifically, we prove an extension of Theorem 21.1 of Györfi et al. (2002) in Appendix C (Lemma 15). We present a modified version of Theorem 10.2 of van de Geer (2000) in Appendix D. We then provide a covering number result in Appendix E (Lemma 20). The reason we require these results will become clear in Appendix B. Finally, we introduce convolutional MDPs as an instance of problems that satisfy Assumption A7 (Appendix F).

We would like to remark that the generic "constants" $c, c' > 0$ in the proofs, especially those related to the statistical guarantees, might change from line to line, if their exact value is not important in the bound. These values are constant as a function of important quantities of the upper bound (such as n, α, $J(Q^\pi)$, etc.), but may depend on $Q_{\max}$ or $|\mathcal{A}|$.

Appendix A. Proof of Theorem 10 (Closed-Form Solutions for RKHS Formulation of REG-LSTD/BRM)

Proof REG-BRM: First, notice that the optimization problem (28) can be written in the form $c_n(Q) + \lambda_{Q,n} \|Q\|_{\mathcal{H}}^2 \overset{Q}{\to} \min!$ with an appropriately defined functional c_n.[19] In order to apply the representer theorem (Schölkopf et al., 2001), we require to show that c_n depends on Q only through the data-points $Z_1, Z_1', \ldots, Z_n, Z_n'$. This is immediate for all the terms that define c_n except the term that involves $\hat{h}_n(\cdot; Q)$. However, since $\hat{h}_n$ is defined as the solution to the optimization problem (27), calling for the representer theorem once again, we observe that $\hat{h}_n$ can be written in the form

$$\hat{h}_n(\cdot; Q) = \sum_{t=1}^{n} \beta_t^* \mathrm{K}(Z_t, \cdot),$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_t^*)^\top$ satisfies

$$\boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^n} \left[\left\| \boldsymbol{K}_h \boldsymbol{\beta} - \hat{T}^\pi Q \right\|_n^2 + \lambda_{h,n} \boldsymbol{\beta}^\top \boldsymbol{K}_h \boldsymbol{\beta} \right].$$

Solving this minimization problem leads to

$$\boldsymbol{\beta}^* = (\boldsymbol{K}_h + n\lambda_{h,n} \boldsymbol{I})^{-1} (\hat{T}^{\pi_k} Q).$$

In both equations $(\hat{T}^\pi Q)$ is viewed as the n-dimensional vector

$$\Big((\hat{T}^\pi Q)(Z_1), \ldots, (\hat{T}^\pi Q)(Z_n) \Big)^\top = \big(R_1 + \gamma Q(Z_1'), \ldots, R_n + \gamma Q(Z_n') \big)^\top.$$

Thus, $\boldsymbol{\beta}^*$ depends on Q only through $Q(Z_1'), \ldots, Q(Z_n')$. Plugging this solution into (28), we get that $c_n(Q)$ indeed depends on Q through

$$Q(Z_1), Q(Z_1'), \cdots, Q(Z_n), Q(Z_n'),$$

and thus on data points $Z_1, Z_1', \cdots, Z_n, Z_n'$. The representer theorem then implies that the minimizer of $c_n(Q) + \lambda_{Q,n} \|Q\|_{\mathcal{H}}^2$ can be written in the form $Q(\cdot) = \sum_{i=1}^{2n} \tilde{\alpha}_i \mathrm{K}(\tilde{Z}_i, \cdot)$, where $\tilde{Z}_i = Z_i$ if $i \le n$ and $\tilde{Z}_i = Z_{i-n}'$, otherwise.

Let $\tilde{\boldsymbol{\alpha}} = (\alpha_1, \ldots, \alpha_n, \alpha_1', \ldots, \alpha_n')^\top$. Using the reproducing kernel property of K, we get the optimization problem

$$\|\boldsymbol{C}_1 \boldsymbol{K}_Q \tilde{\boldsymbol{\alpha}} - (\boldsymbol{r} + \gamma \boldsymbol{C}_2 \boldsymbol{K}_Q \tilde{\boldsymbol{\alpha}})\|_n^2 - \|\boldsymbol{B}(\boldsymbol{r} + \gamma \boldsymbol{C}_2 \boldsymbol{K}_Q \tilde{\boldsymbol{\alpha}})\|_n^2 + \lambda_{Q,n} \tilde{\boldsymbol{\alpha}}^\top \boldsymbol{K}_Q \tilde{\boldsymbol{\alpha}} \overset{\tilde{\boldsymbol{\alpha}}}{\to} \min!.$$

Solving this for $\tilde{\boldsymbol{\alpha}}$ concludes the proof for REG-BRM.

REG-LSTD: The first part of the proof that shows c_n depends on Q only through the data-points $Z_1, Z_1', \ldots, Z_n, Z_n'$ is exactly the same as the proof of REG-BRM. Thus, using the representer theorem, the minimizer of (30) can be written in the form $Q(\cdot) = \sum_{i=1}^{2n} \tilde{\alpha}_i \mathrm{K}(\tilde{Z}_i, \cdot)$, where $\tilde{Z}_i = Z_i$ if $i \le n$ and $\tilde{Z}_i = Z_{i-n}'$, otherwise. Let $\tilde{\boldsymbol{\alpha}} = (\alpha_1, \ldots, \alpha_n, \alpha_1', \ldots, \alpha_n')^\top$. Using the reproducing kernel property of K, we get the optimization problem

$$\|(\boldsymbol{C}_1 - \gamma \boldsymbol{E}\boldsymbol{C}_2) \boldsymbol{K}_Q \tilde{\boldsymbol{\alpha}} - \boldsymbol{E}\boldsymbol{r}\|_n^2 + \lambda_{Q,n} \tilde{\boldsymbol{\alpha}}^\top \boldsymbol{K}_Q \tilde{\boldsymbol{\alpha}} \overset{\tilde{\boldsymbol{\alpha}}}{\to} \min!.$$

Replacing $\boldsymbol{C}_1 - \gamma \boldsymbol{E}\boldsymbol{C}_2$ with $\boldsymbol{F}$ and solving for $\tilde{\boldsymbol{\alpha}}$ concludes the proof. ■

19. Here $f(Q) \overset{Q}{\to} \min!$ indicates that Q is a minimizer of $f(Q)$.

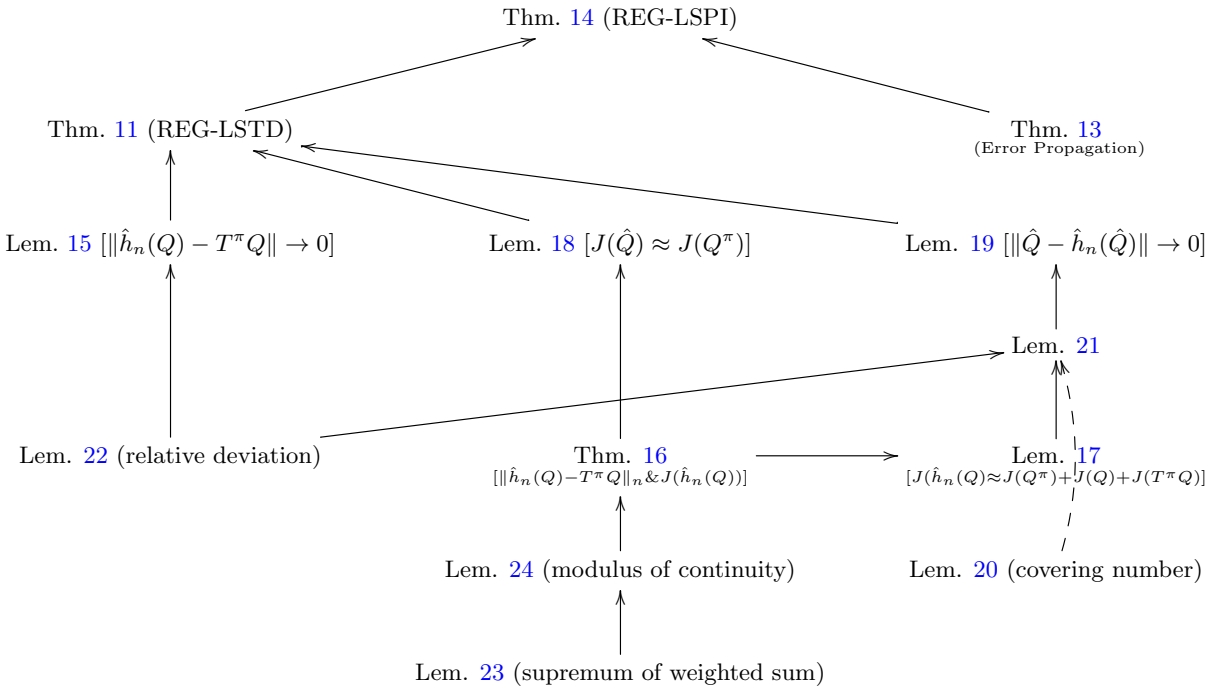

Figure 2: Dependencies of results used to prove the statistical guarantee for REG-LSPI (Theorem 14).

Appendix B. Proof of Theorem 11 (Statistical Guarantee for REG-LSTD)

The goal of Theorem 11 is to provide a finite-sample upper bound on the Bellman error $\|\hat{Q} - T^\pi \hat{Q}\|_\nu$ for REG-LSTD defined by the optimization problems (15) and (16). Since $\|\hat{Q} - T^\pi \hat{Q}\|_\nu^2 \le 2\|\hat{Q} - \hat{h}_n(\cdot; \hat{Q})\|_\nu^2 + 2\|\hat{h}_n(\cdot; \hat{Q}) - T^\pi \hat{Q}\|_\nu^2$, we may upper bound the Bellman error by upper bounding each term in the RHS. Recall from the discussion after Theorem 11 that the analysis is more complicated than the conventional supervised learning setting because the corresponding optimization problems are coupled: $\hat{h}_n(\cdot; \hat{Q})$ is a function of $\hat{Q}$ which itself is a function of $\hat{h}_n(\cdot; \hat{Q})$.

Theorem 11 is proven using Lemma 15, which upper bounds $\|\hat{h}_n(\cdot; \hat{Q}) - T^\pi \hat{Q}\|_\nu$, and Lemma 19, which upper bounds $\|\hat{Q} - \hat{h}_n(\cdot; \hat{Q})\|_\nu$. We also require to relate the smoothness $J(\hat{Q})$ to the smoothness $J(Q^\pi)$. Lemma 18 specifies this relation. The proof of these lemmas themselves require further developments, which will be discussed when we encounter them. Figure 2 shows the dependencies between all results that lead to the proof of Theorem 11 and consequently Theorem 14.

The following lemma controls the error behavior resulting from the optimization problem (15). This lemma, which is a result on the error upper bound of a regularized regression estimator, is similar to Theorem 21.1 of Györfi et al. (2002) with two main differences. First,

it holds uniformly over $T^\pi Q$ (as opposed to a fixed function $T^\pi Q$); second, it holds for function spaces that satisfy a general metric entropy condition (as opposed to the special case of the Sobolev spaces).

Lemma 15 (Convergence of $\hat{h}_n(\cdot; Q)$ to $T^\pi Q$) *For any random $Q \in \mathcal{F}^{|\mathcal{A}|}$, let $\hat{h}_n(Q)$ be defined according to (15). Under Assumptions A1–A5 and A7, there exist finite constants $c_1, c_2 > 0$ such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have*

$$\left\| \hat{h}_n(\cdot; Q) - T^\pi Q \right\|_\nu^2 \le 4\lambda_{h,n} J^2(T^\pi Q) + 2\lambda_{h,n} J^2(Q) + c_1 \frac{1}{n\lambda_{h,n}^\alpha} + c_2 \frac{\ln(1/\delta)}{n},$$

with probability at least $1 - \delta$.

Proof See Appendix C. ■

When we use this lemma to prove Theorem 11, the action-value function Q that appears in the bound is the result of the optimization problems defined in (16), that is $\hat{Q}$, and so is random. Lemma 18, which we will prove later, provides a deterministic upper bound for the smoothness $J(\hat{Q})$ of this random quantity.

It turns out that to derive our main result, we require to know more about the behavior of the regularized regression estimator than what is shown in Lemma 15. In particular, we need an upper bound on the empirical error of the regularized regression estimator $\hat{h}_n(\cdot; Q)$ (cf. (33) below). Moreover, we should bound the random smoothness $J(\hat{h}_n(\cdot; Q))$ by some deterministic quantities, which turns out to be a function of $J(T^\pi Q)$ and $J(Q)$. Theorem 16 provides us with the required upper bounds. This theorem is a modification of Theorem 10.2 by van de Geer (2000), with two main differences: 1) It holds uniformly over Q and 2) $\hat{h}_n(\cdot; Q)$ uses the same data $\mathcal{D}_n$ that is used to estimate Q itself.

We introduce the following notation: Let $w = (x, a, r, x')$ and define the random variables $w_i = (X_i, A_i, R_i, X_i')$ for $1 \le i \le n$. The data set $\mathcal{D}_n$ would be $\{w_1, \dots, w_n\}$. For a measurable function $g : \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, let $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n |g(w_i)|^2$. Consider the regularized least squares estimator:

$$\hat{h}_n(\cdot; Q) = \operatorname*{argmin}_{h \in \mathcal{F}^{|\mathcal{A}|}} \left[\left\| h - [R_i + \gamma Q(X_i', \pi(X_i'))] \right\|_n^2 + \lambda_{h,n} J^2(h) \right], \tag{33}$$

which is the same as (15) with π replacing π_k.

Theorem 16 (Empirical error and smoothness of $\hat{h}_n(\cdot; Q)$) *For a random function $Q \in \mathcal{F}^{|\mathcal{A}|}$, let $\hat{h}_n(\cdot, Q)$ be defined according to (33). Suppose that Assumptions A1–A5 and A7 hold. Then there exist constants $c_1, c_2 > 0$, such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have*

$$\left\| \hat{h}_n(\cdot; Q) - T^\pi Q \right\|_n \le c_1 \max \Bigg\{ \frac{Q_{max}^{1+\alpha} \sqrt{\frac{\ln(1/\delta)}{n}}}{\lambda_{h,n}^{\frac{\alpha}{2}}}, \\ Q_{max} \left(J(Q) + J(T^\pi Q) \right)^{\frac{\alpha}{1+\alpha}} \left(\frac{\ln(1/\delta)}{n} \right)^{\frac{1}{2(1+\alpha)}}, \\ \sqrt{\lambda_{h,n}} J(T^\pi Q) \Bigg\},$$

$$J(\hat{h}_n(\cdot;Q)) \leq c_2 \max\left\{ J(Q) + J(T^\pi Q), \frac{Q_{max}^{1+\alpha}\sqrt{\frac{\ln(1/\delta)}{n}}}{\lambda_{h,n}^{\frac{1+\alpha}{2}}} \right\},$$

with probability at least $1-\delta$.

Proof See Appendix D. ■

The following lemma, which is an immediate corollary of Theorem 16, indicates that with the proper choice of the regularization coefficient, the complexity of the regression function $\hat{h}_n(\cdot;Q)$ is in the same order as the complexities of Q, $T^\pi Q$, and Q^π. This result will be used in the proof of Lemma 21, which itself is used in the proof of Lemma 19.

Lemma 17 (Smoothness of $\hat{h}_n(\cdot;Q)$) *For a random $Q \in \mathcal{F}^{|\mathcal{A}|}$, let $\hat{h}_n(\cdot;Q)$ be the solution to the optimization problem (15) with the choice of regularization coefficient*

$$\lambda_{h,n} = \left[\frac{1}{n\, J^2(Q^\pi)}\right]^{\frac{1}{1+\alpha}}.$$

Let Assumptions A1–A5 and A7 hold. Then, there exits a finite constant $c > 0$, depending on Q_{max}, such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, the upper bound

$$J(\hat{h}_n(\cdot;Q)) \leq c\left(J(T^\pi Q) + J(Q) + J(Q^\pi)\sqrt{\ln(1/\delta)}\right)$$

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

Proof With the choice of $\lambda_{h,n} = \left[\frac{1}{nJ^2(Q^\pi)}\right]^{\frac{1}{1+\alpha}}$, Theorem 16 implies that there exist some finite constant $c_1 > 0$ as well as $c_2 > 0$, which depends on $Q_{\max}$, such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, the inequality

$$\begin{aligned} J(\hat{h}_n(\cdot;Q)) &\leq c_1 \max\left\{ J(Q) + J(T^\pi Q), \frac{Q_{\max}^{1+\alpha}\sqrt{\frac{\ln(1/\delta)}{n}}}{\left[\frac{1}{nJ^2(Q^\pi)}\right]^{-\frac{1}{2}}} \right\} \\ &\leq c_2\left(J(T^\pi Q) + J(Q) + J(Q^\pi)\sqrt{\ln(1/\delta)}\right) \end{aligned}$$

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

An intuitive understanding of this result might be gained if we consider $\hat{h}_n(\cdot;Q^\pi)$, which is the regression estimate for $T^\pi Q^\pi = Q^\pi$. This lemma then indicates that the smoothness of $\hat{h}_n(\cdot;Q^\pi)$ is comparable to the smoothness of its target function Q^π. This is intuitive whenever the regularization coefficients are chosen properly.

The following lemma relates $J(\hat{Q})$ and $J(T^\pi\hat{Q})$, which are random, to the complexity of the action-value function of the policy π, i.e., $J(Q^\pi)$. This result is used in the proof of Theorem 11.

Lemma 18 (Smoothness of $\hat{Q}$) *Let Assumptions A1–A7 hold, and let $\hat{Q}$ be the solution to (16) with the choice of*

$$\lambda_{h,n} = \left[\frac{1}{n\, J^2(Q^\pi)}\right]^{\frac{1}{1+\alpha}}.$$

Then, there exists a finite constant $c > 0$ such that for any $n \in \mathbb{N}$ and $0 < \delta < e^{-1}$, we have

$$\lambda_{Q,n} J^2(\hat{Q}) \le \lambda_{Q,n} J^2(Q^\pi) + c \frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln(1/\delta)}{n^{\frac{1}{1+\alpha}}},$$

with probability at least $1 - \delta$.

Proof By Assumption A6 we have $Q^\pi \in \mathcal{F}^{|\mathcal{A}|}$, so by the optimizer property of $\hat{Q}$ (cf. (16)), we get

$$\lambda_{Q,n} J^2(\hat{Q}) \le \left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n} J^2(\hat{Q}) \le \left\|Q^\pi - \hat{h}_n(\cdot;Q^\pi)\right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n} J^2(Q^\pi). \quad (34)$$

Since $Q^\pi = T^\pi Q^\pi$, we have $\|Q^\pi - \hat{h}_n(\cdot;Q^\pi)\|_{\mathcal{D}_n} = \|T^\pi Q^\pi - \hat{h}_n(\cdot;Q^\pi)\|_{\mathcal{D}_n}$. So Theorem 16 shows that with the choice of $\lambda_{h,n} = [\frac{1}{nJ^2(Q^\pi)}]^{\frac{1}{1+\alpha}}$, there exists a finite constant $c > 0$ such that for any $n \in \mathbb{N}$ and for $0 < \delta < e^{-1} \approx 0.3679$, we have

$$\left\|Q^\pi - \hat{h}_n(\cdot;Q^\pi)\right\|_{\mathcal{D}_n}^2 \le c_1 \left(1 \vee Q_{\max}^{2(1+\alpha)}\right) \frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln(1/\delta)}{n^{\frac{1}{1+\alpha}}}, \quad (35)$$

with probability at least $1 - \delta$. Chaining inequalities (34) and (35) finishes the proof. ■

The other main ingredient of the proof of Theorem 11 is an upper bound to $\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\|_\nu$, which is closely related to the optimization problem (16). This task is done by Lemma 19. In the proof of this lemma, we call Lemma 21, which shall be stated and proven right after this result.

Lemma 19 (Convergence of $\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\|_\nu$) *Let $\hat{Q}$ be the solution to the set of coupled optimization problems (15)–(16). Suppose that Assumptions A1–A7 hold. Then there exists a finite constant $c > 0$ such that for any $n \in \mathbb{N}$ and $0 < \delta < 2e^{-1}$ and with the choice of*

$$\lambda_{h,n} = \lambda_{Q,n} = \left[\frac{1}{n\, J^2(Q^\pi)}\right]^{\frac{1}{1+\alpha}},$$

we have

$$\left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_\nu^2 \le c \frac{(1+\gamma^2 L_P^2)^\alpha\, J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) \ln(1/\delta) + L_R^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}},$$

with probability at least $1 - \delta$.

Proof Decompose

$$\left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_\nu^2 = I_{1,n} + I_{2,n},$$

with

$$\begin{aligned}\frac{1}{2}I_{1,n} &= \left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n}J^2(\hat{Q}),\\ I_{2,n} &= \left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_\nu^2 - I_{1,n}.\end{aligned} \tag{36}$$

In what follows, we upper bound each of these terms.
$I_{1,n}$: Use the optimizer property of $\hat{Q}$ to get

$$\frac{1}{2}I_{1,n} = \left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n}J^2(\hat{Q}) \le \left\|Q^\pi - \hat{h}_n(\cdot;Q^\pi)\right\|_{\mathcal{D}_n}^2 + \lambda_{Q,n}J^2(Q^\pi).$$

To upper bound $\left\|Q^\pi - \hat{h}_n(\cdot;Q^\pi)\right\|_{\mathcal{D}_n}^2 = \left\|T^\pi Q^\pi - \hat{h}_n(\cdot;Q^\pi)\right\|_{\mathcal{D}_n}^2$, we evoke Theorem 16. For our choice of $\lambda_{Q,n}$, there exists a constant $c_1 > 0$ such that for any $n \in \mathbb{N}$ and $0 < \delta_1 < 1$, we have

$$\frac{1}{2}I_{1,n} \le \lambda_{Q,n}J^2(Q^\pi) + c_1\frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)\ln(1/\delta_1)}{n^{\frac{1}{1+\alpha}}}, \tag{37}$$

with probability at least $1 - \delta_1$.
$I_{2,n}$: With our choice of $\lambda_{Q,n}$ and $\lambda_{h,n}$, Lemma 21, which shall be proven later, indicates that there exist some finite constants $c_2, c_3, c_4 > 0$ such that for any $n \in \mathbb{N}$ and finite $J(Q^\pi)$, L_R, and L_P, and $0 < \delta_2 < 1$, we have

$$I_{2,n} \le c_2\frac{L_R^{\frac{2\alpha}{1+\alpha}} + [J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}}[\ln(1/\delta_2)]^{\frac{\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}} + c_3\frac{(1+\gamma^2L_P^2)^\alpha}{n\,\lambda_{Q,n}^\alpha} + c_4\frac{\ln(1/\delta_2)}{n}, \tag{38}$$

with probability at least $1 - \delta_2$. For $\delta_2 < e^{-1}$ and $\alpha \ge 0$, we have $[\ln(1/\delta_2)]^{\frac{\alpha}{1+\alpha}} \le \ln(1/\delta_2)$, and also

$$\frac{1}{n\,\lambda_{Q,n}^\alpha} = \frac{[J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}} \le \frac{[J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}}\ln(1/\delta_2). \tag{39}$$

With the right choice of constants, $\frac{\ln(1/\delta_2)}{n}$ can be absorbed into the other terms. Select $\delta_1 = \delta_2 = \delta/2$. Inequalities (37), (38), and (39) imply that with the specified choice of $\lambda_{Q,n}$ and $\lambda_{h,n}$, there exists a finite constant $c_5 > 0$ such that for any $0 < \delta < 2e^{-1}$, we have

$$\left\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\right\|_\nu^2 \le c_5\frac{(1+\gamma^2L_P^2)^\alpha\,J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)\ln(1/\delta) + L_R^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}},$$

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

To upper bound $I_{2,n}$, defined in (36), we simultaneously apply the peeling device (cf. Section 5.3 of van de Geer 2000) on two different, but coupled, function spaces (one to which

$\hat{Q}$ belongs and the other to which $\hat{h}_n(\cdot;\hat{Q})$ belongs). In each layer of peeling, we apply an exponential tail inequality to control the relative deviation of the empirical mean from the true mean (Lemma 22 in Appendix C). We also require a covering number result, which is stated as Lemma 20. The final result of this procedure is a tight upper bound on $I_{2,n}$, as stated in Lemma 21.

To prepare for the peeling argument, define the following subsets of $\mathcal{F}$ and $\mathcal{F}^{|\mathcal{A}|}$:

$$\mathcal{F}_\sigma \triangleq \left\{ f \,:\, f \in \mathcal{F}, J^2(f) \le \sigma \right\},$$
$$\mathcal{F}_\sigma^{|\mathcal{A}|} \triangleq \left\{ f \,:\, f \in \mathcal{F}^{|\mathcal{A}|}, J^2(f) \le \sigma \right\}.$$

Let

$$g_{Q,h}(x,a) \triangleq \sum_{j=1}^{|\mathcal{A}|} \mathbb{I}_{\{a=a_j\}} \left[Q_j(x) - h_j(x)\right]^2. \tag{40}$$

To simplify the notation, we use $z = (x,a)$ and $Z = (X,A)$ in the rest of this section. Define G_{σ_1,σ_2} as the space of $g_{Q,h}$ functions with $J(Q) \le \sigma_1$ and $J(h) \le \sigma_2$, i.e.,

$$G_{\sigma_1,\sigma_2} \triangleq \left\{ g_{Q,h} \,:\, \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}; Q \in \mathcal{F}_{\sigma_1}^{|\mathcal{A}|}, h \in \mathcal{F}_{\sigma_2}^{|\mathcal{A}|} \right\}. \tag{41}$$

The following lemma provides an upper bound on the covering numbers of G_{σ_1,σ_2}.

Lemma 20 (Covering Number) *Let Assumptions A3, A4, and A5 hold. Then there exists a constant $c_1 > 0$, independent of σ_1, σ_2, α, Q_{max}, and $|\mathcal{A}|$, such that for any $u > 0$ and all $((x_1,a_1),\dots,(x_n,a_n)) \in \mathcal{X} \times \mathcal{A}$, the empirical covering number of the class of functions G_{σ_1,σ_2} defined in (41) w.r.t. the empirical norm $\|\cdot\|_{2,z_{1:n}}$ is upper bounded by*

$$\log \mathcal{N}_2(u, G_{\sigma_1,\sigma_2}, (x,a)_{1:n}) \le c_1 |\mathcal{A}|^{1+\alpha} Q_{max}^{2\alpha} \left(\sigma_1^\alpha + \sigma_2^\alpha\right) u^{-2\alpha}.$$

Proof See Appendix E. ∎

Next, we state and prove Lemma 21, which provides a high probability upper bound on $I_{2,n}$.

Lemma 21 *Let $I_{2,n}$ be defined according to (36). Under Assumptions A1–A5 and A7 and with the choice of*

$$\lambda_{h,n} = \lambda_{Q,n} = \left[\frac{1}{n\, J^2(Q^\pi)}\right]^{\frac{1}{1+\alpha}},$$

there exist constants $c_1, c_2, c_3 > 0$, such that for any $n \in \mathbb{N}$, finite $J(Q^\pi)$, L_R, and L_P, and $\delta > 0$ we have

$$I_{2,n} \le c_1 \frac{L_R^{\frac{2\alpha}{1+\alpha}} + [J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}} [\ln(1/\delta)]^{\frac{\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}} + c_2 \frac{(1+\gamma^2 L_P^2)^\alpha}{n\, \lambda_{Q,n}^\alpha} + c_3 \frac{\ln(1/\delta)}{n},$$

with probability at least $1-\delta$.

Proof Let $Z = (X, A)$ be a random variable with distribution ν that is independent from $\mathcal{D}_n$. Without loss of generality, we assume that $Q_{\max} \geq 1/2$. We use the peeling device in conjunction with Lemmas 20 and 22 to obtain a tight high-probability upper bound on $I_{2,n}$. Based on the definition of $I_{2,n}$ in (36) we have

$$\mathbb{P}\left\{I_{2,n} > t\right\} = \mathbb{P}\left\{\frac{\mathbb{E}\left[g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z)|\mathcal{D}_n\right] - \frac{1}{n}\sum_{i=1}^n g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z_i)}{t + 2\lambda_{Q,n}J^2(\hat{Q}) + \mathbb{E}\left[g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z)|\mathcal{D}_n\right]} > \frac{1}{2}\right\}. \tag{42}$$

To benefit from the peeling device, we relate the complexity of $\hat{h}_n(\cdot;\hat{Q})$ to the complexity of $\hat{Q}$. For a fixed $\delta_1 > 0$ and some constant $c > 0$, to be specified shortly, define the following event:

$$\mathcal{A}_0 = \left\{\omega \,:\, J^2(\hat{h}_n(\cdot;\hat{Q})) \leq c\left(J^2(T^\pi\hat{Q}) + J^2(\hat{Q}) + J^2(Q^\pi)\ln(1/\delta_1)\right)\right\}.$$

Lemma 17 indicates that $\mathbb{P}\{\mathcal{A}_0\} \geq 1 - \delta_1$, where the constant c here can be chosen to be three times of the squared value of the constant in the lemma. We have $\mathbb{P}\{I_{2,n} > t\} = \mathbb{P}\left\{I_{2,n} > t, \mathcal{A}_0^C\right\} + \mathbb{P}\{I_{2,n} > t, \mathcal{A}_0\} \leq \delta_1 + \mathbb{P}\{I_{2,n} > t, \mathcal{A}_0\}$, so we focus on upper bounding $\mathbb{P}\{I_{2,n} > t, \mathcal{A}_0\}$.

Since $\hat{Q} \in \mathcal{F}^{|\mathcal{A}|}$, there exists $l \in \mathbb{N}_0$ such that $2^l t\mathbb{I}_{\{l\neq 0\}} \leq 2\lambda_{Q,n}J^2(\hat{Q}) < 2^{l+1}t$. Fix $l \in \mathbb{N}_0$. For any $Q \in \mathcal{F}^{|\mathcal{A}|}$, Assumption A7 relates $J(T^\pi Q)$ to $J(Q)$:

$$J^2(Q) \leq \frac{2^l t}{\lambda_{Q,n}} \Rightarrow J^2(T^\pi Q) \leq 2\left(L_R^2 + \gamma^2 L_P^2 \frac{2^l t}{\lambda_{Q,n}}\right).$$

Thus on the event $\mathcal{A}_0$, if $\hat{Q} \in \mathcal{F}^{|\mathcal{A}|}_{\sigma_1^l}$ where $\sigma_1^l = \frac{2^l t}{\lambda_{Q,n}}$, we also have $\hat{h}_n(\hat{Q}) \in \mathcal{F}^{|\mathcal{A}|}_{\sigma_2^l}$ with

$$\sigma_2^l = c\left[2\left(L_R^2 + (1+\gamma^2 L_P^2)\frac{2^l t}{\lambda_{Q,n}}\right) + J^2(Q^\pi)\ln(1/\delta_1)\right]. \tag{43}$$

Apply the peeling device on (42). Use (43) and note that if for an $l \in \mathbb{N}_0$ we have $2\lambda_{Q,n}J^2(\hat{Q}) \geq 2^l t\mathbb{I}_{\{l\neq 0\}}$, we also have $t + 2\lambda_{Q,n}J^2(\hat{Q}) \geq 2^l t$ to get

$$\begin{aligned}
\mathbb{P}\{I_{2,n} > t\} &= \mathbb{P}\left\{I_{2,n} > t, \mathcal{A}_0^C\right\} + \mathbb{P}\{I_{2,n} > t, \mathcal{A}_0\} \\
&\leq \delta_1 + \sum_{l=0}^{\infty}\mathbb{P}\Bigg\{\mathcal{A}_0, 2^l t\mathbb{I}_{\{l\neq 0\}} \leq 2\lambda_{Q,n}J^2(\hat{Q}) < 2^{l+1}t, \\
&\qquad\qquad \frac{\mathbb{E}\left[g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z)|\mathcal{D}_n\right] - \frac{1}{n}\sum_{i=1}^n g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z_i)}{t + 2\lambda_{Q,n}J^2(\hat{Q}) + \mathbb{E}\left[g_{\hat{Q},\hat{h}_n(\cdot;\hat{Q})}(Z)|\mathcal{D}_n\right]} > \frac{1}{2}\Bigg\} \\
&\leq \delta_1 + \sum_{l=0}^{\infty}\mathbb{P}\left\{\sup_{g_{Q,h}\in G_{\sigma_1^l,\sigma_2^l}} \frac{\mathbb{E}\left[g_{Q,h}(Z)|\mathcal{D}_n\right] - \frac{1}{n}\sum_{i=1}^n g_{Q,h}(Z_i)}{2^l t + \mathbb{E}\left[g_{Q,h}(Z)|\mathcal{D}_n\right]} > \frac{1}{2}\right\}.
\end{aligned} \tag{44}$$

Let us study the behavior of the l^{th} term of the above summation by verifying the conditions of Lemma 22 with the choice of $\varepsilon = \frac{1}{2}$ and $\eta = 2^l t$.

Condition (A1): Since all functions involved are bounded by $Q_{\max}$, it is easy to see that $|g_{Q,h}(x,a)| \le \sum_{j=1}^{|\mathcal{A}|} \mathbb{I}_{\{a=a_j\}} \left|[Q_j(x) - h_j(x)]^2\right| \le 4Q_{\max}^2$. Therefore, K_1, defined in Lemma 22, can be set to $K_1 = 4Q_{\max}^2$.

Condition (A2): We have $\mathbb{E}\left[\left|[Q(Z) - h(Z)]^2\right|^2\right] \le 4Q_{\max}^2 \mathbb{E}\left[[Q(Z) - h(Z)]^2\right]$. Therefore, K_2 can be set to $K_2 = 4Q_{\max}^2$.

Condition (A3): We should satisfy $\frac{\sqrt{2}}{4}\sqrt{n\eta} \ge 288 \max\{8Q_{\max}^2, \sqrt{8}Q_{\max}\}$. Since $\eta = 2^l t \ge t$, it is sufficient to have

$$t \ge \frac{c}{n}, \tag{C1}$$

in which c is a function of $Q_{\max}$ (we can choose $c = 2 \times 4608^2\, Q_{\max}^4$).

Condition (A4): We shall verify that for $\varepsilon' \ge \frac{1}{8}\eta = \frac{1}{8}2^l t$, and $\sigma_1 = \sigma_1^l$ and $\sigma_2 = \sigma_2^l$, the following holds:

$$\frac{\sqrt{n}(\frac{1}{2})(\frac{1}{2})\varepsilon'}{96\sqrt{2}\max\{K_1, 2K_2\}} \ge \int_{\frac{(\frac{1}{2})(\frac{1}{2})\varepsilon'}{16\max\{K_1,2K_2\}}}^{\sqrt{\varepsilon'}} \left(\log \mathcal{N}_2\left(u, \left\{ g \in G_{\sigma_1,\sigma_2} \,:\, \frac{1}{n}\sum_{i=1}^n g^2(z_i) \le 16\varepsilon' \right\}, z_{1:n}\right)\right)^{1/2} du. \tag{45}$$

Notice that there exists a constant $c > 0$ such that for any $u, \varepsilon' > 0$

$$\begin{aligned}\log \mathcal{N}_2\left(u, \left\{ g \in G_{\sigma_1,\sigma_2} \,:\, \frac{1}{n}\sum_{i=1}^n g^2(z_i) \le 16\varepsilon' \right\}, z_{1:n}\right) &\le \log \mathcal{N}_2\left(u, G_{\sigma_1,\sigma_2}, z_{1:n}\right) \\ &\le c\,(\sigma_1^\alpha + \sigma_2^\alpha) u^{-2\alpha},\end{aligned} \tag{46}$$

where we used Lemma 20 in the second inequality.

Plug (46) into (45) with the choice of $\sigma_1 = \sigma_1^l = \frac{2^l t}{\lambda_{Q,n}}$ and $\sigma_2 = \sigma_2^l = c[2(L_R^2 + (1+\gamma^2 L_P^2)\frac{2^l t}{\lambda_{Q,n}}) + J^2(Q^\pi)\ln(1/\delta_1)]$. Therefore, for some constant $c' = c'(Q_{\max}) > 0$, the inequality

$$c'\sqrt{n}\varepsilon' \ge \int_0^{\sqrt{\varepsilon'}} \left[\underbrace{\left(\frac{2^l t}{\lambda_{Q,n}}\right)^\alpha}_{(a)} + \underbrace{c\left[2\left(L_R^2 + (1+\gamma^2 L_P^2)\frac{2^l t}{\lambda_{Q,n}}\right) + J^2(Q^\pi)\ln(1/\delta_1)\right]^\alpha}_{(b)} \right]^{1/2} u^{-\alpha} du,$$

implies (45). Because $(a+b)^{\frac{1}{2}} \le (a^{\frac{1}{2}} + b^{\frac{1}{2}})$ for non-negative a and b, it suffices to verify the following two conditions:

(a) We shall verify that for $\varepsilon' \ge \frac{1}{8}2^l t$, we have

$$c\sqrt{n}\varepsilon' \ge \left(\frac{2^l t}{\lambda_{Q,n}}\right)^{\frac{\alpha}{2}} \varepsilon'^{\frac{1-\alpha}{2}} \Leftrightarrow c\frac{\sqrt{n}\varepsilon'^{\frac{1+\alpha}{2}}\lambda_{Q,n}^{\frac{\alpha}{2}}}{(2^l t)^{\frac{\alpha}{2}}} \ge 1$$

for some $c > 0$. Substituting ε' with $2^l t$, we see that it is enough if for some constant $c > 0$,

$$t \geq \frac{c}{2^l \, n \, \lambda_{Q,n}^{\alpha}}. \tag{D1}$$

(b) We should verify that for $\varepsilon' \geq \frac{1}{8} 2^l t$, the following is satisfied:

$$\sqrt{n}\varepsilon' \geq c \left[\underbrace{L_R^2}_{(b_1)} + \underbrace{(1+\gamma^2 L_P^2)\frac{2^l t}{\lambda_{Q,n}}}_{(b_2)} + \underbrace{J^2(Q^\pi)\ln(1/\delta_1)}_{(b_3)} \right]^{\alpha/2} \varepsilon'^{\frac{1-\alpha}{2}},$$

for some $c > 0$. After some manipulations, we get that the previous inequality holds if the following three inequalities are satisfied:

$$(b_1): \qquad t \geq c_1' \frac{L_R^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}}, \tag{D2}$$

$$(b_2): \qquad t \geq c_2' \frac{(1+\gamma^2 L_P^2)^\alpha}{n \lambda_{Q,n}^{\alpha}}, \tag{D3}$$

$$(b_3): \qquad t \geq c_3' \frac{[J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}} [\ln(1/\delta_1)]^{\frac{\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}}, \tag{D4}$$

for some constants $c_1', c_2', c_3' > 0$.

Fix $\delta > 0$ and let $\delta_1 = \delta/2$. Whenever (C1), (D1), (D2), (D3), and (D4) are satisfied, for some choice of constants $c, c' > 0$ we have

$$\begin{aligned} \mathbb{P}\{I_{2,n} > t\} &\leq \frac{\delta}{2} + \sum_{l=0}^{\infty} 60 \exp\left(- \frac{n(2^l t)(\frac{1}{4})(1-\frac{1}{2})}{128 \times 2304 \times \max\{16 Q_{\max}^4, 4 Q_{\max}^2\}} \right) \\ &\leq \frac{\delta}{2} + c \exp(-c' n \, t). \end{aligned}$$

Let the left-hand side be equal δ and solve for t. Considering all aforementioned conditions, we get that there exist constants $c_1, c_2, c_3 > 0$ such that for any $n \in \mathbb{N}$, finite $J(Q^\pi)$, L_R, and L_P, and $\delta > 0$, we have

$$I_{2,n} \leq c_1 \frac{L_R^{\frac{2\alpha}{1+\alpha}} + [J(Q^\pi)]^{\frac{2\alpha}{1+\alpha}} [\ln(1/\delta)]^{\frac{\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}} + c_2 \frac{(1+\gamma^2 L_P^2)^\alpha}{n \, \lambda_{Q,n}^{\alpha}} + c_3 \frac{\ln(1/\delta)}{n},$$

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

After developing these tools, we are ready to prove Theorem 11.

Proof [Proof of Theorem 11] We want to show that $\|\hat{Q} - T^\pi \hat{Q}\|_\nu$ is small. Since (15)-(16) minimize $\|\hat{h}_n(\cdot; \hat{Q}) - T^\pi \hat{Q}\|_\nu$ and $\|\hat{Q} - \hat{h}_n(\cdot; \hat{Q})\|_\nu$, we upper bound $\|\hat{Q} - T^\pi \hat{Q}\|_\nu$ in terms of these quantities as follows:

$$\left\| \hat{Q} - T^\pi \hat{Q} \right\|_\nu^2 \leq 2 \left\| \hat{Q} - \hat{h}_n(\cdot; \hat{Q}) \right\|_\nu^2 + 2 \left\| \hat{h}_n(\cdot; \hat{Q}) - T^\pi \hat{Q} \right\|_\nu^2. \tag{47}$$

Let us upper bound each of these two terms in the RHS. Fix $0 < \delta < 1$.

Bounding $\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi\hat{Q}\|_\nu$: Lemma 15 indicates that there exist constants $c_1, c_2 > 0$ such that for any random $\hat{Q} \in \mathcal{F}^{|\mathcal{A}|}$ and any fixed $n \in \mathbb{N}$, we have

$$\left\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi\hat{Q}\right\|_\nu^2 \le \lambda_{h,n}\left(2J^2(\hat{Q}) + 4J^2(T^\pi\hat{Q})\right) + c_1\frac{1}{n\lambda_{h,n}^\alpha} + c_2\frac{\ln(3/\delta)}{n}, \tag{48}$$

with probability at least $1 - \delta/3$. Note that $T^\pi\hat{Q} \in \mathcal{F}^{|\mathcal{A}|}$ is implied by Assumption A7 and $\hat{Q} \in \mathcal{F}^{|\mathcal{A}|}$.

Because $\hat{Q}$ is random itself, the terms $J(\hat{Q})$ and $J(T^\pi\hat{Q})$ in the upper bound of (48) are also random. In order to upper bound them, we use Lemma 18, which states that upon the choice of $\lambda_{h,n} = \lambda_{Q,n} = [\frac{1}{n\,J^2(Q^\pi)}]^{\frac{1}{1+\alpha}}$, there exists a constant $c_3 > 0$ such that for any $n \in \mathbb{N}$,

$$\lambda_{h,n}J^2(\hat{Q}) = \lambda_{Q,n}J^2(\hat{Q}) \le \lambda_{Q,n}J^2(Q^\pi) + c_3\frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)}{n^{\frac{1}{1+\alpha}}}\ln(3/\delta) \tag{49}$$

holds with probability at least $1 - \delta/3$. We use Assumption A7 to show that we have

$$\lambda_{h,n}J^2(T^\pi\hat{Q}) \le 2\lambda_{Q,n}L_R^2 + 2(\gamma L_P)^2\left(\lambda_{Q,n}J^2(Q^\pi) + c_3\frac{J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)}{n^{\frac{1}{1+\alpha}}}\ln(3/\delta)\right), \tag{50}$$

with the same probability. Plugging (49) and (50) into (48) and using the selected schedule for $\lambda_{Q,n}$ and $\lambda_{h,n}$, we get

$$\begin{aligned}
&\left\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi\hat{Q}\right\|_\nu^2 \le \\
&\left[\left(2 + c_1 + 8(\gamma L_P)^2\right)J^{\frac{2\alpha}{1+\alpha}}(Q^\pi) + c_3\left(2 + 8(\gamma L_P)^2\right)J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)\ln(3/\delta) + \frac{8L_R^2}{J^{\frac{2}{1+\alpha}}(Q^\pi)}\right]\frac{1}{n^{\frac{1}{1+\alpha}}} \\
&+ c_2\frac{\ln(3/\delta)}{n},
\end{aligned}$$

with probability at least $1 - \frac{2}{3}\delta$. By the proper choice of constants, the term $c_2 n^{-1}\ln(3/\delta)$ can be absorbed into $n^{\frac{-1}{1+\alpha}}\ln(3/\delta)$. Therefore, there exists a constant $c_4 > 0$ such that

$$\left\|\hat{h}_n(\cdot;\hat{Q}) - T^\pi\hat{Q}\right\|_\nu^2 \le \left[c_4\left[1 + (\gamma L_P)^2\right]J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)\ln(1/\delta) + \frac{8L_R^2}{[J(Q^\pi)]^{\frac{2}{1+\alpha}}}\right]\frac{1}{n^{\frac{1}{1+\alpha}}}, \tag{51}$$

with probability at least $1 - \frac{2}{3}\delta$.

Bounding $\|\hat{Q} - \hat{h}_n(\cdot;\hat{Q})\|_\nu$: With our choice of $\lambda_{Q,n}$ and $\lambda_{h,n}$, Lemma 19 states that there exists a constant $c_5 > 0$ such that for any $n \in \mathbb{N}$,

$$\left\|\hat{Q} - \hat{h}_n(\cdot;\cdot;\hat{Q})\right\|_\nu^2 \le c_5\frac{(1+\gamma^2L_P^2)^\alpha J^{\frac{2\alpha}{1+\alpha}}(Q^\pi)\ln(1/\delta) + L_R^{\frac{2\alpha}{1+\alpha}}}{n^{\frac{1}{1+\alpha}}}, \tag{52}$$

holds with probability at least $1 - \delta/3$.

Thus, inequality (47) alongside upper bounds (51) and (52) indicate that there exist constants $c_6, c_7 > 0$ such that for any $n \in \mathbb{N}$ and $\delta > 0$, we have

$$\left\|\hat{Q} - T^{\pi}\hat{Q}\right\|_{\nu}^{2} \leq \frac{c_6 \left[1 + (\gamma L_P)^2\right] J^{\frac{2\alpha}{1+\alpha}}(Q^{\pi}) \ln(1/\delta) + c_7 \left(L_R^{\frac{2\alpha}{1+\alpha}} + \frac{L_R^2}{[J(Q^{\pi})]^{\frac{2}{1+\alpha}}}\right)}{n^{\frac{1}{1+\alpha}}},$$

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

A careful study of the proof of Theorem 11 and the auxiliary results used in it reveals that one can indeed reuse a single data set in all iterations. Recall that at the k^{th} iteration of an API procedure such as REG-LSPI, the policy $\pi = \pi_k$ is the greedy policy w.r.t. $\hat{Q}^{(k-1)}$, so it depends on earlier data sets. This implies that a function such as $T^{\pi}\hat{Q} = T^{\hat{\pi}(\cdot;\hat{Q}^{(k-1)})}\hat{Q}$ is random with two sources of randomness: One source is the data set used in the current iteration, which defines the empirical loss functions. This directly affects $\hat{Q}$. The other source is $\hat{\pi}(\cdot;\hat{Q}^{(k-1)})$, which depends on the data sets in earlier iterations. When we assume that all data sets are independent from each other, the randomness of π does not cause any problem because we can work on the probability space conditioned on the data sets of the earlier iterations. Conditioned on that randomness, the policy π becomes a deterministic function. This is how we presented the statement of Theorem 11 by stating that π is fixed. Nonetheless, the proofs can handle the dependence with no change. Briefly speaking, the reason is that when we want to provide a high probability upper bounds on certain random quantities, we take the supremum over both $\hat{Q}$ and $T^{\pi}\hat{Q}$ and consider them as two separate functions, even though they are related through a random T^{π} operator.

To see this more clearly, notice that in the proof of Lemma 15, which is used in the proof of this theorem, we define the function spaces $\mathcal{G}_l$ that chooses the functions h, Q, and $T^{\pi}Q$ separately. We then take the supremum over all functions in $\mathcal{G}_l$. This means that for the probabilistic upper bound, the randomness of π in $T^{\pi}Q$ becomes effectively irrelevant as we are providing a uniform over $\mathcal{G}_l$ guarantee. In the proof of this theorem, we also use Lemma 19, which itself uses Theorem 16 and Lemma 21 that have a similar construct.

Appendix C. Proof of Lemma 15 (Convergence of $\hat{h}_n(\cdot;Q)$ to $T^{\pi}Q$)

The following lemma, quoted from Györfi et al. (2002), provides an exponential probability tail inequality for the relative deviation of the empirical mean from the true mean. A slightly modified version of this result was published as Theorem 2 of Kohler (2000). This result is used in the proof of Lemmas 15 and 21.

Lemma 22 (Theorem 19.3 of Györfi et al. 2002) *Let $Z, Z_1, \cdots, Z_n$ be independent and identically distributed random variables with values in $\mathcal{Z}$. Let $0 < \varepsilon < 1$ and $\eta > 0$. Assume that $K_1, K_2 \geq 1$ and let $\mathcal{F}$ be a permissible class of functions $f : \mathcal{Z} \to \mathbb{R}$ with the following properties:*

(A1) $\|f\|_{\infty} \leq K_1$,

(A2) $\mathbb{E}\left[f(Z)^2\right] \leq K_2 \mathbb{E}\left[f(Z)\right]$,

(A3) $\sqrt{n}\varepsilon\sqrt{1-\varepsilon}\sqrt{\eta} \geq 288 \max\{2K_1, \sqrt{2K_2}\}$,

(A4) *For all* $z_1, \cdots, z_n \in \mathcal{Z}$ *and all* $\delta \geq \eta/8$,

$$\frac{\sqrt{n}\varepsilon(1-\varepsilon)\delta}{96\sqrt{2}\max\{K_1, 2K_2\}} \geq \int_{\frac{\varepsilon(1-\varepsilon)\delta}{16\max\{K_1,2K_2\}}}^{\sqrt{\delta}} \sqrt{\log \mathcal{N}_2\left(u, \left\{f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} f^2(z_i) \leq 16\delta\right\}, z_{1:n}\right)}\, du.$$

Then

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \frac{\left|\mathbb{E}\left[f(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right|}{\eta + \mathbb{E}\left[f(Z)\right]} > \varepsilon\right\} \leq 60 \exp\left(-\frac{n\,\eta\,\varepsilon^2(1-\varepsilon)}{128 \times 2304 \max\{K_1^2, K_2\}}\right).$$

Let us now turn to the proof of Lemma 15. This proof follows similar steps to the proof of Theorem 21.1 of Györfi et al. (2002).

Proof [Proof of Lemma 15] Without loss of generality, assume that $Q_{\max} \geq 1/2$. Denote $z = (x, a)$ and let $Z = (X, A) \sim \nu$, $R \sim \mathcal{R}(\cdot|X, A)$, and $X' \sim P(\cdot|X, A)$ be random variables that are independent of $\mathcal{D}_n = \{(X_i, A_i, R_i, X_i')\}_{i=1}^n$. Define the following error decomposition

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{A}} \left|\hat{h}_n(z;Q) - T^\pi Q(z)\right|^2 d\nu(z) &= \mathbb{E}\left[\left|\hat{h}_n(Z;Q) - \left[R + \gamma Q(X', \pi(X'))\right]\right|^2 \middle| \mathcal{D}_n\right] - \\ &\quad \mathbb{E}\left[\left|T^\pi Q(Z) - \left[R + \gamma Q(X', \pi(X'))\right]\right|^2\right] \\ &= I_{1,n} + I_{2,n},\end{aligned}$$

with

$$\begin{aligned}\frac{1}{2} I_{1,n} &= \frac{1}{n}\sum_{i=1}^{n} \left|\hat{h}_n(Z_i;Q) - \left[R_i + \gamma Q(X_i', \pi(X_i'))\right]\right|^2 - \left|T^\pi Q(Z_i) - \left[R_i + \gamma Q(X_i', \pi(X_i'))\right]\right|^2 + \\ &\quad \lambda_{h,n}\left(J^2(\hat{h}_n(\cdot;Q)) + J^2(Q) + J^2(T^\pi Q)\right), \\ I_{2,n} &= \mathbb{E}\left[\left|\hat{h}_n(Z;Q) - \hat{T}^\pi Q(Z)\right|^2 - \left|T^\pi Q(Z) - \hat{T}^\pi Q(Z)\right|^2 \middle| \mathcal{D}_n\right] - I_{1,n}.\end{aligned}$$

By the optimizer property of $\hat{h}_n(\cdot;Q)$, we get the following upper bound by substituting $\hat{h}_n(\cdot;Q)$ with $T^\pi Q \in \mathcal{F}^{|\mathcal{A}|}$:

$$\begin{aligned}I_{1,n} &\leq 2\left[\frac{1}{n}\sum_{i=1}^{n} \left|T^\pi Q(Z_i) - \hat{T}^\pi Q(Z_i)\right|^2 - \left|T^\pi Q(Z_i) - \hat{T}^\pi Q(Z_i)\right|^2 + \right. \\ &\qquad \left. \lambda_{h,n}\left(J^2(T^\pi Q) + J^2(Q) + J^2(T^\pi Q)\right)\right] \\ &= 4\lambda_{h,n} J^2(T^\pi Q) + 2\lambda_{h,n} J^2(Q). \end{aligned} \tag{53}$$

We now turn to upper bounding $\mathbb{P}\{I_{2,n} > t\}$. Given a policy π and functions $h, Q, Q' \in \mathcal{F}^{|\mathcal{A}|}$, for $w = (x, a, r, x')$ define $g : \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ as

$$g_{h,Q,Q'}(w) = \left|h(z) - \left[r + \gamma Q(x', \pi(x'))\right]\right|^2 - \left|Q'(z) - \left[r + \gamma Q(x', \pi(x'))\right]\right|^2.$$

Note that $g_{\hat{h}_n(\cdot;Q),Q,T^\pi Q}$ is the function appearing in the definition of $I_{2,n}$. Define the following function spaces for $l = 0, 1, \ldots$:

$$\mathcal{G}_l \triangleq \left\{ g_{h,Q,Q'} : \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X} \to \mathbb{R} \,:\, h, Q, Q' \in \mathcal{F}^{|\mathcal{A}|}; J^2(h), J^2(Q), J^2(Q') \le \frac{2^l t}{\lambda_{h,n}} \right\}.$$

Denote $W = (X, A, R, X')$ and $W_i = (X_i, A_i, R_i, X_i')$. Apply the peeling device to get

$$\begin{aligned}
\mathbb{P}\{I_{2,n} > t\} \le \sum_{l=0}^{\infty} \mathbb{P}\Bigg(&\exists h, Q \in \mathcal{F}^{|\mathcal{A}|}, 2^l t\, \mathbb{I}_{\{l \neq 0\}} \le 2\lambda_{h,n}\left(J^2(h) + J^2(Q) + J^2(T^\pi Q)\right) < 2^{l+1} t; \\
&\text{s.t. } \frac{\mathbb{E}\left[g_{h,Q,T^\pi Q}(W)|\mathcal{D}_n\right] - \frac{1}{n}\sum_{i=1}^n g_{h,Q,T^\pi Q}(W_i)}{t + 2\lambda_{h,n}\left(J^2(h) + J^2(Q) + J^2(T^\pi Q)\right) + \mathbb{E}\left[g_{h,Q}(W)|\mathcal{D}_n\right]} > \frac{1}{2} \Bigg) \\
\le \sum_{l=0}^{\infty} \mathbb{P}\Bigg(&\sup_{g \in \mathcal{G}_l} \frac{\mathbb{E}\left[g(W)|\mathcal{D}_n\right] - \frac{1}{n}\sum_{i=1}^n g(W_i)}{2^l t + \mathbb{E}\left[g(W)|\mathcal{D}_n\right]} > \frac{1}{2} \Bigg).
\end{aligned}$$

Here we used the simple fact that if $2\lambda_{h,n}\left(J^2(h) + J^2(Q) + J^2(T^\pi Q)\right) < 2^{l+1} t$, then $J^2(h)$, $J^2(Q)$, and $J^2(T^\pi Q)$ are also less than $\frac{2^l t}{\lambda_{h,n}}$, so $g_{h,Q,T^\pi Q} \in \mathcal{G}_l$.

We study the behavior of the l^{th} term of the above summation by verifying the conditions of Lemma 22—similar to what we did in the proof of Lemma 21.

It is easy to verify that (A1) and (A2) are satisfied with the choice of $K_1 = K_2 = 4Q_{\max}^2$. Condition (A3) is satisfied whenever

$$t \ge \frac{c_1}{n}, \tag{54}$$

for some constant $c_1 > 0$ depending on $Q_{\max}$ (the constant can be set to $c_1 = 2 \times 4608^2 Q_{\max}^2$).

To verify condition (A4), we first require an upper bound on $\mathcal{N}_2(u, \mathcal{G}_l, w_{1:n})$ for any sequence $w_{1:n}$. This can be done similar to the proof of Lemma 20: Denote $\mathcal{F}_l = \{f : f \in \mathcal{F}, J^2(f) \le \frac{2^l t}{\lambda_{h,n}}\}$. For $g_{h_1,Q_1,T^\pi Q_1}, g_{h_2,Q_2,T^\pi Q_2} \in \mathcal{G}_l$ and any sequence $w_{1:n}$ we have

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^n \left|g_{h_1,Q_1,T^\pi Q_1}(w_i) - g_{h_2,Q_2,T^\pi Q_2}(w_i)\right|^2 \\
&\le 12(2+\gamma)^2 Q_{\max}^2 \frac{1}{n}\sum_{i=1}^n \Big[\left|h_1(z_i) - h_2(z_i)\right|^2 + 4\gamma^2 \left|Q_1(x_i', \pi(x_i')) - Q_2(x_i', \pi(x_i'))\right|^2 + \\
&\qquad\qquad \left|T^\pi Q_1(z_i) - T^\pi Q_2(z_i)\right|^2 \Big] \\
&\le 12(2+\gamma)^2 Q_{\max}^2 \frac{1}{n}\sum_{i=1}^n \sum_{a \in \mathcal{A}} \Big[\left|h_1(x_i, a) - h_2(x_i, a)\right|^2 + 4\gamma^2 \left|Q_1(x_i', a) - Q_2(x_i', a)\right|^2 + \\
&\qquad\qquad \left|T^\pi Q_1(x_i, a) - T^\pi Q_2(x_i, a)\right|^2 \Big].
\end{aligned}$$

With the same covering set argument as in the proof of Lemma 20, we get that for any $u > 0$,

$$\mathcal{N}_2(18\sqrt{2|\mathcal{A}|}Q_{\max}\, u, \mathcal{G}_l, w_{1:n}) \le \mathcal{N}_2(u, \mathcal{F}_l, x_{1:n})^{|\mathcal{A}|} \times \mathcal{N}_2(u, \mathcal{F}_l, x'_{1:n})^{|\mathcal{A}|} \times \mathcal{N}_2(u, \mathcal{F}_l, x_{1:n})^{|\mathcal{A}|}.$$

Evoke Assumption A4 to get

$$\log \mathcal{N}_2(u, \mathcal{G}_l, w_{1:n}) \le c(|\mathcal{A}|, Q_{\max}) \left(\frac{2^l t}{\lambda_{h,n}}\right)^{\alpha} u^{-2\alpha}.$$

Plugging this covering number result into condition (A4), one can verify that the condition is satisfied if

$$t \ge \frac{c_2}{n\lambda_{h,n}^{\alpha}}, \tag{55}$$

for a constant $c_2 > 0$, which is only a function of $Q_{\max}$ and $|\mathcal{A}|$. Therefore, Lemma 22 indicates that

$$\mathbb{P}\{I_{2,n} > t\} \le 60 \sum_{l=0}^{\infty} \exp\left(-\frac{n(2^l t)(1/4)(1/2)}{128 \times 2304 \times \max\{16Q_{\max}^4, 4Q_{\max}^2\}}\right) \le c_3 \exp(-c_4 n t). \tag{56}$$

for some constants $c_3, c_4 > 0$.

Combining (53), (54), (55), and (56), we find that there exist $c_5, c_6 > 0$ such that for any $n \in \mathbb{N}$ and $0 < \delta < 1$, we have

$$\left\|\hat{h}_n(Q) - T^{\pi}Q\right\|_{\nu}^2 \le 4\lambda_{h,n} J^2(T^{\pi}Q) + 2\lambda_{h,n} J^2(Q) + c_5 \frac{1}{n\lambda_{h,n}^{\alpha}} + c_6 \frac{\ln(1/\delta)}{n}.$$

Here, c_5 is only a function of $Q_{\max}$ and $|\mathcal{A}|$, and c_6 is a function of $Q_{\max}$. ■

Appendix D. Proof of Theorem 16 (Empirical Error and Smoothness of $\hat{h}_n(\cdot; Q)$)

To prove Theorem 16, which is a modification of Theorem 10.2 by van de Geer (2000), we first need to modify and specialize Lemma 3.2 by van de Geer (2000) to be suitable to our problem. The modification is required because Q in (33) is a random function in $\mathcal{F}^{|\mathcal{A}|}$ as opposed to being a fixed function as in Theorem 10.2 of van de Geer (2000).

Let us denote $z = (x, a) \in \mathcal{Z} = \mathcal{X} \times \mathcal{A}$ and $Z' = (x, a, R, X') \in \mathcal{Z}' = \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X}$ with $(R, X') \sim P(\cdot, \cdot|x, a)$. Let $\mathcal{D}_n$ denote the set $\{(x_i, a_i, R_i, X'_i)\}_{i=1}^n$ of independent random variables. We use z_i to refer to (x_i, a_i) and Z'_i to refer to (x_i, a_i, R_i, X'_i). Let P_n be the probability measure that puts mass $1/n$ on $z_1, \ldots, z_n$, i.e., $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$, in which δ_z is the Dirac's delta function that puts a mass of 1 at z.

Denote $\mathcal{G} : \mathcal{Z} \to \mathbb{R}$ and $\mathcal{G}' : \mathcal{Z}' \to \mathbb{R}^{3|\mathcal{A}|}$, which is defined as the set

$$\mathcal{G}' = \left\{ (Q, T^{\pi}Q, \mathbf{1}) \,:\, Q \in \mathcal{F}^{|\mathcal{A}|} \right\}$$

with $\mathbf{1} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{A}|}$ being a bounded constant function (and not necessarily equal to 1). We use $\|g\|_\infty$ to denote the supremum norm of functions in $\mathcal{G}$. The supremum norm of vector-valued functions in $\mathcal{G}'$ is defined by taking the supremum norm over the l_∞-norm of each vector. Similarly, the supremum norm of $(g, g') \in \mathcal{G} \times \mathcal{G}'$ is defined by $\|(g, g')\|_\infty \triangleq \max\{\|g\|_\infty, \|g'\|_\infty\}$.

For $g \in \mathcal{G}$, we define $\|g\|_{P_n} \triangleq [\frac{1}{n}\sum_{i=1}^n g^2(z_i)]^{1/2}$. To simplify the notation, we use the following definition of the inner product: Fix $n \in \mathbb{N}$. Consider $z_1, \ldots, z_n$ as a set of points in $\mathcal{Z}$, and a real-valued sequence $w = (w_1, \ldots, w_n)$. For a function $g \in \mathcal{G}$, define $\langle w, g \rangle_n \triangleq \frac{1}{n}\sum_{i=1}^n w_i g(z_i)$.

For any $g' = (Q, T^\pi Q, \mathbf{1}) \in \mathcal{G}'$, define the mapping $\bar{W}(g')(x, a, r, x') : \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ by $\bar{W}(g')(x, a, r, x') = r\,\mathbf{1} + \gamma Q(x', \pi(x')) - T^\pi Q(x, a)$. For any fixed $g' \in \mathcal{G}'$ and $i = 1, \ldots, n$, define the random variables $W_i(g') = \bar{W}(g')(Z_i')$ and let $W(g')$ denote the random vector $[W_1(g') \ldots W_n(g')]^\top$. Notice that $W_i(g')$ can be re-written as $W_i(g') = (R_i - r(z_i)) + \gamma(Q(X', \pi(X')) - (P^\pi Q)(z_i))$, thus for any fixed g', $\mathbb{E}[W_i(g')] = 0$ $(i = 1, \ldots, n)$. For notational simplification, we use $a \vee b = \max\{a, b\}$.

Lemma 23 (Modified Lemma 3.2 of van de Geer 2000) *Fix the sequence $(z_i)_{i=1}^n \subset \mathcal{Z}$ and let $(Z_i')_{i=1}^n \subset \mathcal{Z}'$ be the sequence of independent random variables defined as above. Assume that for some constants $0 < R \le L$, it holds that $\sup_{g\in\mathcal{G}} \|g\|_{P_n} \le R$, $\sup_{g'\in\mathcal{G}'} \|g'\|_\infty \le L$, and $|R_i| \le L$ $(1 \le i \le n)$ almost surely. There exists a constant C such that for all $0 \le \varepsilon < \delta$ satisfying*

$$\sqrt{n}(\delta - \varepsilon) \ge C\,L \left[\int_{\frac{\varepsilon}{28L}}^{R} [\log \mathcal{N}_\infty(u, \mathcal{G} \times \mathcal{G}')]^{1/2} \mathrm{d}u \ \vee R \right], \tag{57}$$

we have

$$\mathbb{P}\left\{ \sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \left| \frac{1}{n}\sum_{i=1}^n W_i(g') g(z_i) \right| \ge \delta \right\} \le 4 \exp\left(-\frac{n(\delta - \varepsilon)^2}{2^7 \times 3^5 (RL)^2} \right).$$

The main difference between this lemma and Lemma 3.2 of van de Geer (2000) is that the latter provides a maximal inequality for $\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n W_i g(z_i)$, with W_i being random variables that satisfy a certain exponential probability inequality, while our result is a maximal inequality for $\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{1}{n}\sum_{i=1}^n W_i(g') g(z_i)$, i.e., the random variables $W_i(g')$ are functions of an arbitrary $g' \in \mathcal{G}'$. The current proof requires us to have a condition on the metric entropy w.r.t. the supremum norm (cf. (57)) instead of w.r.t. the empirical L_2-norm used in Lemma 3.2 of van de Geer (2000). The possibility of relaxing this requirement is an interesting question. We now prove this result.

Proof First, note that for any $g_1, g_2 \in \mathcal{G}$, and $g'_1, g'_2 \in \mathcal{G}'$ (with the identification of g' with its corresponding Q and $T^\pi Q$), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} W_i(g'_1)g_1(z_i) - W_i(g'_2)g_2(z_i) = \\
&\frac{1}{n}\sum_{i=1}^{n} (R_i - r(z_i))(g_1(z_i) - g_2(z_i)) + \\
&\frac{1}{n}\sum_{i=1}^{n} \gamma\left[(Q_1(X'_i,\pi(X'_i)) - P^\pi Q_1(z_i)) - (Q_2(X'_i,\pi(X'_i)) - P^\pi Q_2(z_i))\right] g_1(z_i) + \\
&\frac{1}{n}\sum_{i=1}^{n} \gamma(Q_2(X'_i,\pi(X'_i)) - P^\pi Q_2(z_i))(g_1(z_i) - g_2(z_i)) \leq \\
&2L\left\|g_1 - g_2\right\|_{P_n} + \gamma R\left[\left\|Q_1 - Q_2\right\|_\infty + \left\|P^\pi Q_1 - P^\pi Q_2\right\|_\infty\right] + 3\gamma L\left\|g_1 - g_2\right\|_{P_n} = \\
&(2+3\gamma)L\left\|g_1 - g_2\right\|_{P_n} + \gamma R\left\|Q_1 - Q_2\right\|_\infty + R\left\|T^\pi Q_1 - T^\pi Q_2\right\|_\infty, \qquad (58)
\end{aligned}$$

where we used the boundedness assumptions, the definition of the supremum norm, the norm inequality $\frac{1}{n}\sum_{i=1}^n |g_1(z_i) - g_2(z_i)| \leq \|g_1 - g_2\|_{P_n}$, and the fact that $|\gamma Q(X'_i,\pi(X'_i)) - \gamma P^\pi Q(z_i)| = |r(z_i) + \gamma Q(X'_i,\pi(X'_i)) - T^\pi Q(z_i)| \leq (2+\gamma)L \leq 3L$ for any L-bounded Q and $T^\pi Q$ to get the inequality. We used $\left\|P^\pi Q^s - P^\pi Q^{s-1}\right\|_\infty = \gamma^{-1}\left\|T^\pi Q^s - T^\pi Q^{s-1}\right\|_\infty$ to get the last equality.

Let $\{(g^s_j, {g'}^s_j)\}_{j=1}^{N_s}$ with $N_s = \mathcal{N}_\infty(2^{-s}R, \mathcal{G}\times\mathcal{G}')$ be a minimal $2^{-s}R$-covering of $\mathcal{G}\times\mathcal{G}'$ w.r.t. the supremum norm. For any $(g, g') \in \mathcal{G}\times\mathcal{G}'$, there exists a $(g^s, (Q^s, T^\pi Q^s, \mathbf{1})) = (g^s, g'^s) \in \{(g^s_j, {g'}^s_j)\}_{j=1}^{N_s}$ such that $\|(g,g') - (g^s, g'^s)\|_\infty \leq 2^{-s}R$. This implies that both $\|Q^s - Q\|_\infty$ and $\|T^\pi Q^s - T^\pi Q\|_\infty$ are smaller than $2^{-s}R$ as well. Moreover, $\|g^s - g\|_{P_n} \leq \|g^s - g\|_\infty \leq 2^{-s}R$. By (58) we get

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n} W_i(g'^s)g^s(z_i) - W_i(g')g(z_i)\right| &\leq [(2+3\gamma)L + (1+\gamma)R](2^{-s}R) \leq (3+4\gamma)L(2^{-s}R) \\
&\leq 7RL\,2^{-s}.
\end{aligned}$$

Choose $S = \min\{s \geq 1 : 2^{-s} \leq \frac{\varepsilon}{7RL}\}$, which entails that for any $(g,g') \in \mathcal{G}\times\mathcal{G}'$, the covering set defined by $\{(g^S_j, {g'}^S_j)\}_{j=1}^{N_S}$ approximates the inner product of $[g(z_1)\cdots g(z_n)]^\top$ and $W(g')$ with an error less than ε. So it suffices to prove the exponential inequality for

$$\mathbb{P}\left\{\max_{j=1,\ldots,N_S}\left|\frac{1}{n}\sum_{i=1}^{n} W_i({g'}^S_j)g^S_j(z_i)\right| \geq \delta - \varepsilon\right\}.$$

We use the chaining technique (e.g., see van de Geer 2000) as follows (we choose $g^0 = 0$, so $W_i(g'^0)g^0(z_i) = 0$ for all $1 \le i \le n$):

$$\frac{1}{n}\sum_{i=1}^{n} W_i(g'^S)g^S(z_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\left(W_i(g'^s)g^s(z_i) - W_i(g'^{s-1})g^{s-1}(z_i)\right) =$$
$$\sum_{s=1}^{S}\Bigg[\frac{1}{n}\sum_{i=1}^{n}(R_i - r(z_i))(g^s(z_i) - g^{s-1}(z_i)) +$$
$$\frac{1}{n}\sum_{i=1}^{n}\gamma\left[(Q^s(X'_i,\pi(X'_i)) - (P^\pi Q^s)(z_i)) - (Q^{s-1}(X'_i,\pi(X'_i)) - (P^\pi Q^{s-1})(z_i))\right]g^s(z_i) +$$
$$\frac{1}{n}\sum_{i=1}^{n}\gamma(Q^{s-1}(X'_i,\pi(X'_i)) - (P^\pi Q^{s-1})(z_i))(g^s(z_i) - g^{s-1}(z_i))\Bigg].$$

Because each of these summations consists of bounded random variables with expectation zero, we may use Hoeffding's inequality alongside the union bound to upper bound them. To apply Hoeffding's inequality, we require an upper bound on the sum of squared values of random variables involved. To begin, we have $|g^s(z_i) - g^{s-1}(z_i)| = |g^s(z_i) - g(z_i) + g(z_i) - g^{s-1}(z_i)| \le 2^{-s}R + 2^{-(s-1)}R = 3 \times 2^{-s}R$. Similarly, both $\left\|Q^s - Q^{s-1}\right\|_\infty$ and $\left\|T^\pi Q^s - T^\pi Q^{s-1}\right\|_\infty$ are smaller than $3 \times 2^{-s}R$. As a result, for the first term we get

$$\frac{1}{n}\sum_{i=1}^{n}\left[(R_i - r(z_i))(g^s(z_i) - g^{s-1}(z_i))\right]^2 \le 36(RL)^2 2^{-2s}.$$

For the second term we have

$$\frac{1}{n}\sum_{i=1}^{n}\left|\gamma\left[(Q^s(X'_i,\pi(X'_i)) - (P^\pi Q^s)(z_i)) - (Q^{s-1}(X'_i,\pi(X'_i)) - (P^\pi Q^{s-1})(z_i))\right]g^s(z_i)\right|^2$$
$$\le 2\gamma^2\left[\left\|Q^s - Q^{s-1}\right\|_\infty^2 + \gamma^{-2}\left\|T^\pi Q^s - T^\pi Q^{s-1}\right\|_\infty^2\right]\|g^s\|_{P_n}^2$$
$$\le 2(1+\gamma^2)3^2(2^{-s}R)^2R^2 \le 36R^4 2^{-2s},$$

in which we used $\left\|P^\pi Q^s - P^\pi Q^{s-1}\right\|_\infty = \gamma^{-1}\left\|T^\pi Q^s - T^\pi Q^{s-1}\right\|_\infty$. And finally,

$$\frac{1}{n}\sum_{i=1}^{n}\left|\gamma(Q^{s-1}(X'_i,\pi(X'_i)) - (P^\pi Q^{s-1})(z_i))(g^s(z_i) - g^{s-1}(z_i))\right|^2 \le (3L)^2 3^2 (2^{-s}R)^2$$
$$= 9^2(RL)^2 2^{-2s},$$

where we used the fact that $|\gamma Q(X'_i,\pi(X'_i)) - \gamma P^\pi Q(z_i)| \le 3L$ for any L-bounded Q and $T^\pi Q$.

Let η_s be a sequence of positive real-valued numbers satisfying $\sum_{s=1}^{S}\eta_s \le 1$. We continue the chaining argument by the use of the union bound and the fact that $N_s N_{s-1} \le N_s^2$ to

get

$$
\begin{aligned}
P_1 &= \mathbb{P}\left\{ \sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \left| \frac{1}{n}\sum_{i=1}^n W_i(g')g(z_i) \right| \geq \delta \right\} \\
&\leq \mathbb{P}\left\{ \max_{j=1,\dots,N_S} \left| \frac{1}{n}\sum_{i=1}^n W_i({g'_j}^S){g_j}^S(z_i) \right| \geq \delta-\varepsilon \right\} \\
&\leq \sum_{s=1}^S \mathbb{P}\Bigg\{ \max_{\substack{(g^s,{g'}^s)\\(g^{s-1},{g'}^{s-1})}} \left| \frac{1}{n}\sum_{i=1}^n (R_i - r(z_i))(g^s(z_i)-g^{s-1}(z_i)) \right| \geq \frac{\eta_s(\delta-\varepsilon)}{3} \Bigg\} + \\
&\qquad \mathbb{P}\Bigg\{ \max_{\substack{(g^s,{g'}^s)\\(g^{s-1},{g'}^{s-1})}} \Bigg| \frac{1}{n}\sum_{i=1}^n \gamma\big[\big(Q^s(X'_i,\pi(X'_i)) - (P^\pi Q^s)(z_i)\big) - \\
&\qquad\qquad \big(Q^{s-1}(X'_i,\pi(X'_i)) - (P^\pi Q^{s-1})(z_i)\big)\big] g^s(z_i) \Bigg| \geq \frac{\eta_s(\delta-\varepsilon)}{3} \Bigg\} + \\
&\qquad \mathbb{P}\Bigg\{ \max_{\substack{(g^s,{g'}^s)\\(g^{s-1},{g'}^{s-1})}} \left| \frac{1}{n}\sum_{i=1}^n \gamma(Q^{s-1}(X'_i,\pi(X'_i)) - (P^\pi Q^{s-1})(z_i))(g^s(z_i)-g^{s-1}(z_i)) \right| \geq \\
&\qquad\qquad \frac{\eta_s(\delta-\varepsilon)}{3} \Bigg\} \\
&\leq \sum_{s=1}^S N_s N_{s-1} \exp\left(-\frac{2(\delta-\varepsilon)^2\eta_s^2 n}{4\times 9^2 (RL)^2 2^{-2s}} \right) + N_s N_{s-1} \exp\left(-\frac{2(\delta-\varepsilon)^2\eta_s^2 n}{4\times 9^2 R^4 2^{-2s}} \right) + \\
&\qquad N_s N_{s-1} \exp\left(-\frac{2(\delta-\varepsilon)^2\eta_s^2 n}{3\times 9^2 (RL)^2 2^{-2s}} \right) \\
&\leq \sum_{s=1}^S 3\exp\left(2\log N_s - \frac{2(\delta-\varepsilon)^2\eta_s^2 n}{4\times 9^2 (RL)^2 2^{-2s}} \right). \qquad (59)
\end{aligned}
$$

Here the $\max_{(g^s,{g'}^s)}$ is over the corresponding covering set $\{(g_j^s, {g'_j}^s)\}_{j=1}^{N_s}$, which has N_s elements (and the same for $s-1$).
Choose

$$
\eta_s = \frac{3^2\sqrt{6}RL2^{-s}(\log N_s)^{1/2}}{\sqrt{n}(\delta-\varepsilon)} \vee \frac{2^{-s}\sqrt{s}}{8}. \qquad (60)
$$

Take C in (57) sufficiently large such that

$$
\sqrt{n}(\delta-\varepsilon) \geq 2\times 3^2\sqrt{6}\, RL \sum_{s=1}^S 2^{-s}[\log \mathcal{N}_\infty(2^{-s}R, \mathcal{G}\times\mathcal{G}')]^{1/2} \ \vee 72\sqrt{6\log 4}\, RL. \qquad (61)
$$

It can be shown that by this choice of η_s and the condition (61), we have $\sum_{s=1}^S \eta_s \leq 1$.

From (60), we have $\log N_s \leq \frac{n(\delta-\varepsilon)^2\eta_s^2}{2\times 3^5 (RL)^2 2^{-2s}}$, so P_1 in (59) can be upper bounded as follows

$$
P_1 \leq \sum_{s=1}^S 3\exp\left(-\frac{n(\delta-\varepsilon)^2\eta_s^2}{2\times 3^5 (RL)^2 2^{-2s}} \right).
$$

Since $\eta_s \geq 2^{-s}\sqrt{s}/8$ too, we have

$$P_1 \leq 3\sum_{s=1}^{S} \exp\left(-\frac{n(\delta-\varepsilon)^2 2^{-2s} s}{2^7 \times 3^5 (RL)^2 2^{-2s}}\right) \leq 3\sum_{s=1}^{\infty} \exp\left(-\frac{n(\delta-\varepsilon)^2 s}{2^7 \times 3^5 (RL)^2}\right)$$
$$\leq \frac{3\exp\left(-\frac{n(\delta-\varepsilon)^2}{2^7\times 3^5 (RL)^2}\right)}{1-\exp\left(-\frac{n(\delta-\varepsilon)^2}{2^7\times 3^5 (RL)^2}\right)} \leq 4\exp\left(-\frac{n(\delta-\varepsilon)^2}{2^7 \times 3^5 (RL)^2}\right),$$

where in the last inequality we used the assumption that $\sqrt{n}(\delta-\varepsilon) \geq 72\sqrt{6\log 4}\, RL$ (cf. (61)).

One can show that (61) is satisfied if

$$\sqrt{n}(\delta-\varepsilon) \geq 36\sqrt{6}\, L \int_{\frac{\varepsilon}{28L}}^{R} [\log \mathcal{N}_\infty(u, \mathcal{G}\times\mathcal{G}')]^{1/2} \mathrm{d}u \ \vee \ 72\sqrt{6\log 4}\, RL,$$

so C can be chosen as $C = 72\sqrt{6\log 4}$. ■

The following lemma, which is built on Lemma 23, is a result on the behavior of the modulus of continuity and will be used in the proof of Theorem 16. This lemma provides a high-probability upper bound on $\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha} J^\alpha(g,g')}$. Here $J(g,g')$ is a regularizer that is defined on $\mathcal{G}\times\mathcal{G}'$ and is a pseudo-norm.

This result is similar in spirit to Lemma 8.4 of van de Geer (2000), with two main differences: The first is that here we provide an upper bound on

$$\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha} J^\alpha(g,g')},$$

whereas in Lemma 8.4 of van de Geer (2000), the upper bound is on

$$\sup_{g\in\mathcal{G}} \frac{|\langle W, g\rangle_n|}{\|g\|_{P_n}^{1-\alpha}}.$$

The normalization by $\|g\|_{P_n}^{1-\alpha} J^\alpha(g,g')$ instead of $\|g\|_{P_n}^{1-\alpha}$ is important to get the right error bound in Theorem 16. The other crucial difference is that here W are random variables that are functions of $g' \in \mathcal{G}'$, while the result of van de Geer (2000) is for independent W. The proof technique is inspired by Lemmas 5.13, 5.14, and 8.4 of van de Geer (2000).

Lemma 24 (Modulus of Continuity for Weighted Sums) *Fix the sequence $(z_i)_{i=1}^n \subset \mathcal{Z}$ and define $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$. Let $(Z_i')_{i=1}^n \subset \mathcal{Z}'$ be the sequence of independent random variables defined as before. Assume that for some constant $L > 0$, it holds that $\sup_{g\in\mathcal{G}} \|g\|_{P_n} \leq L$, $\sup_{g'\in\mathcal{G}'} \|g'\|_\infty \leq L$, and $|R_i| \leq L$ $(1 \leq i \leq n)$ almost surely. Furthermore, suppose that there exist $0 < \alpha < 1$ and a finite constant A such that for all $u > 0$,*

$$\log \mathcal{N}_\infty\left(u, \left\{ (g,g') \in \mathcal{G}\times\mathcal{G}' \,:\, J(g,g') \leq B \right\}\right) \leq A\left(\frac{B}{u}\right)^{2\alpha}.$$

Then there exists a constant $c > 0$ such that for any $0 < \delta < 1$, we have

$$\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha} J^\alpha(g,g')} \le cL^{1+\alpha}\sqrt{\frac{\ln(\frac{1}{\delta})}{n}},$$

with probability at least $1-\delta$.

Proof The proof uses double-peeling, i.e., we peel on both $J(g,g')$ and $\|g\|_{P_n}$. Without loss of generality, we assume that $L \ge 1$. We use $c_1, c_2, \ldots > 0$ as constants. First, we start by peeling on $J(g,g')$:

$$\begin{aligned}
\delta &\triangleq \mathbb{P}\left\{\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha} J^\alpha(g,g')} \ge t\right\} \\
&\le \sum_{s=0}^{\infty} \mathbb{P}\left\{\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha}} \ge \underbrace{t.2^{\alpha s}}_{\triangleq\tau_s},\ 2^s\mathbb{I}_{\{s\neq 0\}} \le J(g,g') < 2^{s+1}\right\}. \qquad (62)
\end{aligned}$$

Let us denote each term in the RHS by δ_s. To upper bound δ_s, notice that by assumption $\|g\|_{P_n} \le L$. For each term, we peel again, this time on $\|g\|_{P_n}$, and apply Lemma 23:

$$\begin{aligned}
\delta_s &\le \sum_{r\ge 0} \mathbb{P}\Bigg\{\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha}} \ge \tau_s, 2^s\mathbb{I}_{\{s\neq 0\}} \le J(g,g') < 2^{s+1}, \\
&\qquad\qquad 2^{-(r+1)}L < \|g\|_{P_n} \le 2^{-r}L\Bigg\} \\
&\le \sum_{r\ge 0} \mathbb{P}\left\{\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} |\langle W(g'), g\rangle_n| \ge \tau_s\left(2^{-(r+1)}L\right)^{1-\alpha}, J(g,g') < 2^{s+1}, \|g\|_{P_n} \le 2^{-r}L\right\} \\
&\le \sum_{r\ge 0} 4\exp\left(-\frac{n\left[\tau_s\left(2^{-(r+1)}L\right)^{1-\alpha}\right]^2}{2^7\times 3^5\left(2^{-r}L\right)^2 L^2}\right) \\
&= \sum_{r\ge 0} 4\exp\left(-\frac{2^{2r\alpha}\, n\tau_s^2}{2^7\times 3^5\times 2^{2(1-\alpha)}L^{2(1+\alpha)}}\right) = c_2\exp\left(-\frac{c_1 n\tau_s^2}{L^{2(1+\alpha)}}\right). \qquad (63)
\end{aligned}$$

The last inequality holds only if the covering number condition in Lemma 23 is satisfied, which is the case whenever

$$\sqrt{n}\left(\tau_s(2^{-(r+1)}L)^{1-\alpha}\right) \ge C\,L\left[\int_0^{2^{-r}L}\sqrt{A}\left(\frac{2^{s+1}}{u}\right)^\alpha \mathrm{d}u \ \vee 2^{-r}L\right].$$

Substituting $\tau_s = 2^{\alpha s}t$ and solving the integral, we get that the condition is

$$\sqrt{n}t\,2^{\alpha s}(2^{-(r+1)}L)^{1-\alpha} \ge CL\sqrt{A}\left[(2^{s+1})^\alpha(2^{-r}L)^{1-\alpha} \vee 2^{-r}L\right],$$

which would be satisfied for

$$t \ge \frac{CL\sqrt{A}2^{1+\alpha}}{\sqrt{n}} \vee \frac{2^{1-\alpha}CL^{1+\alpha}}{\sqrt{n}} = c_3\frac{L^{1+\alpha}}{\sqrt{n}}. \qquad (64)$$

Plug (63) into (62) to get that

$$\delta \leq \sum_{s=0}^{\infty} c_2 \exp\left(-\frac{c_1 n\, t^2\, 2^{2\alpha s}}{L^{2(1+\alpha)}}\right) = c_4 \exp\left(-\frac{c_1 n\, t^2}{L^{2(1+\alpha)}}\right).$$

Solving for δ, we have $t \leq c_5 L^{1+\alpha}\sqrt{\frac{\ln(1/\delta)}{n}}$ with probability at least $1-\delta$. This alongside the condition (64) lead to the desired result. ■

Let us turn to the proof of Theorem 16. The proof is similar to the proof of Theorem 10.2 by van de Geer (2000), but with necessary modifications in order to get a high probability upper bound that holds uniformly over Q. We discuss the differences in more detail after the proof.

Proof [Proof of Theorem 16] Recall that in the optimization problem, we use $w_i = (X_i, A_i, R_i, X_i')$ $(i = 1, \ldots, n)$ to denote the i^{th} elements of the data set $\mathcal{D}_n = \{(X_i, A_i, R_i, X_i')\}_{i=1}^n$. Also for a measurable function $f : \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, we denote $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n |f(w_i)|^2$. We also let $(X, A) \sim \nu$, $R \sim \mathcal{R}(\cdot|X, A)$, and $X' \sim P(\cdot|X, A)$ be random variables that are independent of $\mathcal{D}_n$.

For any $Q \in \mathcal{F}^{|\mathcal{A}|}$ and the corresponding $T^\pi Q \in \mathcal{F}^{|\mathcal{A}|}$, define the mapping $\bar{W}(Q, T^\pi Q, \mathbf{1}) : \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X} \to \mathbb{R}$ by $\bar{W}(Q, T^\pi Q, \mathbf{1})(X, A, R, X') = R\,\mathbf{1} + \gamma Q(X', \pi(X')) - T^\pi Q(X, A)$, in which $\mathbf{1} \in \mathcal{F}^{|\mathcal{A}|}$ is the constant function defined on $\mathcal{X} \times \mathcal{A}$ with the value of one. For any fixed Q and $i = 1, \ldots, n$, define the random variables $W_i(Q) = \bar{W}(Q, T^\pi Q, \mathbf{1})(X_i, A_i, R_i, X_i')$ and let $W(Q)$ denote the random vector $[W_1(Q) \ldots W_n(Q)]^\top$. Notice that $|W_i(Q)| \leq 3Q_{\max}$, and we have $\mathbb{E}[W_i(Q) \mid Q] = 0$ $(i = 1, \ldots, n)$.

From the optimizer property of $\hat{h}_n = \hat{h}_n(\cdot, Q)$, we have

$$\begin{aligned}&\left\|\hat{h}_n(Q) - [R + \gamma Q(X_i', \pi(X_i'))]\right\|_n^2 + \lambda_{h,n} J^2(\hat{h}_n(Q)) \leq \\ &\left\|T^\pi Q - [R + \gamma Q(X_i', \pi(X_i'))]\right\|_n^2 + \lambda_{h,n} J^2(T^\pi Q).\end{aligned}$$

After expanding and rearranging, we get

$$\left\|\hat{h}_n(Q) - T^\pi Q\right\|_n^2 + \lambda_{h,n} J^2(\hat{h}_n(Q)) \leq 2\left\langle W(Q)\,,\, \hat{h}_n(Q) - T^\pi Q\right\rangle_n + \lambda_{h,n} J^2(T^\pi Q). \quad (65)$$

We evoke Lemma 24 to upper bound $\left|\left\langle W(Q)\,,\, \hat{h}_n(Q) - T^\pi Q\right\rangle_n\right|$. The function spaces $\mathcal{G}$ and $\mathcal{G}'$ in that lemma are set as $G : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $\mathcal{G}' : \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X} \to \mathbb{R}^3$ with

$$\begin{aligned}\mathcal{G} &= \left\{ h - T^\pi Q \,:\, h, Q \in \mathcal{F}^{|\mathcal{A}|} \right\},\\ \mathcal{G}' &= \left\{ (Q, T^\pi Q, \mathbf{1}) \,:\, Q, T^\pi Q \in \mathcal{F}^{|\mathcal{A}|} \right\}.\end{aligned}$$

All functions in $\mathcal{F}^{|\mathcal{A}|}$ are $Q_{\max}$-bounded, so the functions in $\mathcal{G}$ and $\mathcal{G}'$ are bounded by $2Q_{\max}$ and $(Q_{\max}, Q_{\max}, 1)$, respectively. Moreover for any $g \in \mathcal{G}$, $\frac{1}{n}\sum_{i=1}^n |g(X_i, A_i)|^2 \leq 4Q_{\max}^2$. So by setting L equal to $2Q_{\max}$ in that lemma, all boundedness conditions are satisfied.

Define $J(g, g') = J(h) + J(Q) + J(T^\pi Q)$ and denote

$$(\mathcal{G} \times \mathcal{G}')_B = \left\{ (g, g') \in \mathcal{G} \times \mathcal{G}' \,:\, J(g, g') \le B \right\}.$$

Lemma 24 requires an upper bound on $\log \mathcal{N}_\infty (u, (\mathcal{G} \times \mathcal{G}')_B)$. We relate the metric entropy of this space to that of $\mathcal{F}_B = \{ f \in \mathcal{F} \,:\, J(f) \le B \}$, which is specified by Assumption A4.

Notice that if $J(g, g') \le B$, each of $J(h)$, $J(Q)$, and $J(T^\pi Q)$ is also less than or equal to B. So we have

$$(\mathcal{G} \times \mathcal{G}')_B = \left\{ (h - T^\pi Q, Q, T^\pi Q, \mathbf{1}) \,:\, h, Q, T^\pi Q \in \mathcal{F}^{|\mathcal{A}|},\ J(h) + J(Q) + J(T^\pi Q) \le B \right\} \subset$$
$$\left\{ h - T^\pi Q \,:\, h, T^\pi Q \in \mathcal{F}^{|\mathcal{A}|},\ J(h) + J(T^\pi Q) \le B \right\} \times \left\{ Q \,:\, Q \in \mathcal{F}^{|\mathcal{A}|},\ J(Q) \le B \right\} \times$$
$$\left\{ T^\pi Q \,:\, T^\pi Q \in \mathcal{F}^{|\mathcal{A}|},\ J(T^\pi Q) \le B \right\} \times \{\mathbf{1}\}.$$

Because $J(\cdot)$ is a pseudo-norm, we have $J(h - T^\pi Q) \le J(h) + J(T^\pi Q)$, so

$$\left\{ h - T^\pi Q \,:\, h, T^\pi Q \in \mathcal{F}^{|\mathcal{A}|}, J(h) + J(T^\pi Q) \le B \right\} \subset \left\{ Q \,:\, Q \in \mathcal{F}^{|\mathcal{A}|}, J(Q) \le B \right\}.$$

As a result $(\mathcal{G} \times \mathcal{G}')_B$ is a subset of the product space $\left\{ Q \in \mathcal{F}^{|\mathcal{A}|} \,:\, J(Q) \le B \right\}^3$. Therefore by the usual covering argument, we get that

$$\log \mathcal{N}_\infty \left(u, (\mathcal{G} \times \mathcal{G}')_B\right) \le 3 \log \mathcal{N}_\infty \left(u, \left\{ Q \in \mathcal{F}^{|\mathcal{A}|} \,:\, J(Q) \le B \right\}\right).$$

It is easy to see that for finite $|\mathcal{A}|$, if $\log \mathcal{N}_\infty(u, \{f \in \mathcal{F} : J(f) \le B\}) \le C(\frac{B}{u})^{2\alpha}$, then $\log \mathcal{N}_\infty(u, \{(f_1, \dots, f_{|\mathcal{A}|}) \in \mathcal{F}^{|\mathcal{A}|} \,:\, J\left((f_1, \dots, f_{|\mathcal{A}|})\right) \le B\}) \le C_1(\frac{B}{u})^{2\alpha}$ (we benefit from the condition $J(Q(\cdot, a)) \le J(Q)$ in Assumption A3; the proof is similar to the proof of Lemma 20 in Appendix E). Here the constant C_1 depends on $|\mathcal{A}|$. This along with the previous inequality show that for some constant $A > 0$, we have

$$\log \mathcal{N}_\infty \left(u, (\mathcal{G} \times \mathcal{G}')_B\right) \le A \left(\frac{B}{u}\right)^{2\alpha}.$$

We are ready to apply Lemma 24 to upper bound the inner product term in (65). Fix $\delta > 0$. To simplify the notation, denote $L_n = \|\hat{h}_n(Q) - T^\pi Q\|_n$, set $t_0 = \sqrt{\frac{\ln(1/\delta)}{n}}$, and use $\hat{h}_n$ to refer to $\hat{h}_n(Q)$. There exists a constant $c > 0$ such that with probability at least $1 - \delta$, we have

$$L_n^2 + \lambda_{h,n} J^2(\hat{h}_n) \le 2cL^{1+\alpha} L_n^{1-\alpha} \left(J(\hat{h}_n) + J(Q) + J(T^\pi Q)\right)^\alpha t_0 + \lambda_{h,n} J^2(T^\pi Q). \tag{66}$$

Either the first term in the RHS is larger than the second one or the second term is larger than the first. We analyze each case separately.

Case 1. $2cL^{1+\alpha} L_n^{1-\alpha} (J(\hat{h}_n) + J(Q) + J(T^\pi Q))^\alpha t_0 \ge \lambda_{h,n} J^2(T^\pi Q)$. In this case we have

$$L_n^2 + \lambda_{h,n} J^2(\hat{h}_n) \le 4cL^{1+\alpha} L_n^{1-\alpha} \left(J(\hat{h}_n) + J(Q) + J(T^\pi Q)\right)^\alpha t_0. \tag{67}$$

Again, two cases might happen:

Case 1.a. $J(\hat{h}_n) > J(Q) + J(T^\pi Q)$: From (67) we have $L_n^2 \leq 2^{2+\alpha} c L^{1+\alpha} L_n^{1-\alpha} J^\alpha(\hat{h}_n) t_0$. Solving for L_n, we get that $L_n \leq 2^{\frac{2+\alpha}{1+\alpha}} c^{\frac{1}{1+\alpha}} L [J(\hat{h}_n)]^{\frac{\alpha}{1+\alpha}} t_0^{\frac{1}{1+\alpha}}$. From (67) we also have $\lambda_{h,n} J^2(\hat{h}_n) \leq 2^{2+\alpha} c L^{1+\alpha} L_n^{1-\alpha} J^\alpha(\hat{h}_n) t_0$. Plugging-in the recently obtained upper bound on L_n and solving for $J(\hat{h}_n)$, we get that

$$J(\hat{h}_n) \leq \frac{2^{2+\alpha} c L^{1+\alpha} t_0}{\lambda_{h,n}^{\frac{1+\alpha}{2}}}. \tag{68}$$

Substituting this in the upper bound on L_n, we get that

$$L_n \leq \frac{2^{2+\alpha} c L^{1+\alpha} t_0}{\lambda_{h,n}^{\frac{\alpha}{2}}}. \tag{69}$$

Case 1.b. $J(\hat{h}_n) \leq J(Q) + J(T^\pi Q)$: The upper bound on $J(\hat{h}_n)$ is obvious. From (67) we have $L_n^2 \leq 2^{2+\alpha} c L^{1+\alpha} L_n^{1-\alpha} \left(J(Q) + J(T^\pi Q)\right)^\alpha t_0$. Solving for L_n, we obtain

$$L_n \leq 2^{\frac{2+\alpha}{1+\alpha}} c^{\frac{1}{1+\alpha}} L \left(J(Q) + J(T^\pi Q)\right)^{\frac{\alpha}{1+\alpha}} t_0^{\frac{1}{1+\alpha}}. \tag{70}$$

Case 2. $2 c L^{1+\alpha} L_n^{1-\alpha} (J(\hat{h}_n) + J(Q) + J(T^\pi Q))^\alpha t_0 < \lambda_{h,n} J^2(T^\pi Q)$. In this case we have $L_n^2 + \lambda_{h,n} J^2(\hat{h}_n) \leq 2\lambda_{h,n} J^2(T^\pi Q)$, which implies that

$$L_n \leq \sqrt{2\lambda_{h,n}} J(T^\pi Q), \tag{71}$$

$$J(\hat{h}_n) \leq \sqrt{2} J(T^\pi Q). \tag{72}$$

By (69), (70), and (71) for L_n and (68), (72), and the condition $J(\hat{h}_n) \leq J(Q) + J(T^\pi Q)$ in Case 1.b. for $J(\hat{h}_n)$, we have that for any fixed $0 < \delta < 1$, with probability at least $1 - \delta$, the following inequalities hold:

$$\left\| \hat{h}_n(Q) - T^\pi Q \right\|_n \leq \max \left\{ \frac{2^{2+\alpha} c L^{1+\alpha} \sqrt{\frac{\ln(1/\delta)}{n}}}{\lambda_{h,n}^{\frac{\alpha}{2}}}, \right.$$
$$2^{\frac{2+\alpha}{1+\alpha}} c^{\frac{1}{1+\alpha}} L \left(J(Q) + J(T^\pi Q)\right)^{\frac{\alpha}{1+\alpha}} \left(\frac{\ln(1/\delta)}{n}\right)^{\frac{1}{2(1+\alpha)}},$$
$$\left. \sqrt{2\lambda_{h,n}} J(T^\pi Q) \right\},$$

$$J(\hat{h}_n(Q)) \leq \max \left\{ \frac{2^{2+\alpha} c L^{1+\alpha} \sqrt{\frac{\ln(1/\delta)}{n}}}{\lambda_{h,n}^{\frac{1+\alpha}{2}}}, J(Q) + J(T^\pi Q), \sqrt{2} J(T^\pi Q) \right\}.$$

■

Comparing this proof with that of Theorem 10.2 by van de Geer (2000), we see that here we do not normalize the function space $\mathcal{G} \times \mathcal{G}'$ to ensure that $J(g, g') \leq 1$ and then

use their Lemma 8.4, which provides a high-probability upper bound on $\sup_{g\in\mathcal{G}} \frac{|\langle W, g\rangle_n|}{\|g\|_{P_n}^{1-\alpha}}$. Instead we directly apply Lemma 24, which upper bounds $\sup_{(g,g')\in\mathcal{G}\times\mathcal{G}'} \frac{|\langle W(g'), g\rangle_n|}{\|g\|_{P_n}^{1-\alpha} J^\alpha(g,g')}$, on the (unnormalized) function space $\mathcal{G}\times\mathcal{G}'$. If we went through the former approach, in which the normalization is global, the first term in the RHS of (66) would be $L_n^{1-\alpha}(J(\hat{h}_n)+J(Q)+J(T^\pi Q))^{1+\alpha}t_0$ instead of $L_n^{1-\alpha}(J(\hat{h}_n)+J(Q)+J(T^\pi Q))^{\alpha}t_0$ of here, which is obtained by local normalization. This extra $J(\hat{h}_n)+J(Q)+J(T^\pi Q)$ would prevent us from getting proper upper bounds on L_n and $J(\hat{h}_n)$ in Case 1.a above. The reason that the original proof does not work is that here $W(g')$ is a function of $g'\in\mathcal{G}'$.

Appendix E. Proof of Lemma 20 (Covering Number of G_{σ_1,σ_2})

Here we prove Lemma 20, which relates the covering number of G_{σ_1,σ_2} to the covering number of $\mathcal{F}_{\sigma_1}$ and $\mathcal{F}_{\sigma_2}$.

Proof [Proof of Lemma 20] Let $g_{Q_1,h_1}, g_{Q_2,h_2}\in G_{\sigma_1,\sigma_2}$. By the definition of G_{σ_1,σ_2} (41), the functions Q_1 and h_1 corresponding to g_{Q_1,h_1} satisfy $Q_1\in\mathcal{F}_{\sigma_1}^{|\mathcal{A}|}$ and $h_1\in\mathcal{F}_{\sigma_2}^{|\mathcal{A}|}$ (and similarly for Q_2 and h_2). Set $z_i=(x_i,a_i)$. We have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n |g_{Q_1,h_1}(z_i)-g_{Q_2,h_2}(z_i)|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left[(Q_1(z_i)-h_1(z_i))^2-(Q_2(z_i)-h_2(z_i))^2\right]^2 \\
&\le 16Q_{\max}^2\frac{1}{n}\sum_{i=1}^n \left[(Q_1(z_i)-Q_2(z_i))+(h_1(z_i)-h_2(z_i))\right]^2 \\
&\le 32Q_{\max}^2\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{|\mathcal{A}|}\left[(Q_{1,j}(x_i)-Q_{2,j}(x_i))^2+(h_{1,j}(x_i)-h_{2,j}(x_i))^2\right].
\end{aligned}
$$

Assumption A3 implies that for $Q_1,Q_2\in\mathcal{F}_{\sigma_1}^{|\mathcal{A}|}$, the functions $Q_{1,j},Q_{2,j}$ are in $\mathcal{F}_{\sigma_1}$ and for $h_1,h_2\in\mathcal{F}_{\sigma_2}^{|\mathcal{A}|}$, the functions $h_{1,j},h_{2,j}$ are in $\mathcal{F}_{\sigma_2}$—for all $j=1,\cdots,|\mathcal{A}|$. Therefore the previous inequality shows that u-covers on $Q_j\in\mathcal{F}_{\sigma_1}$ and $h_j\in\mathcal{F}_{\sigma_2}$ (for $j=1,\cdots,|\mathcal{A}|$) w.r.t. the empirical norms $\|\cdot\|_{x_{1:n}}$ define an $8Q_{\max}\sqrt{|\mathcal{A}|}$ u-cover on G_{σ_1,σ_2} w.r.t. $\|\cdot\|_{z_{1:n}}$. Thus,

$$
\mathcal{N}_2\left(8Q_{\max}\sqrt{|\mathcal{A}|}u, G_{\sigma_1,\sigma_2},(x,a)_{1:n}\right)\le \mathcal{N}_2\left(u,\mathcal{F}_{\sigma_1},x_{1:n}\right)^{|\mathcal{A}|}\times\mathcal{N}_2\left(u,\mathcal{F}_{\sigma_2},x_{1:n}\right)^{|\mathcal{A}|}.
$$

Assumption A4 then implies that for a constant c_1, independent of u, $|\mathcal{A}|$, $Q_{\max}$, and α, and for all $((x_1,a_1),\ldots,(x_n,a_n))\in\mathcal{X}\times\mathcal{A}$ we have

$$
\log\mathcal{N}_2(u,G_{\sigma_1,\sigma_2},(x,a)_{1:n})\le c_1|\mathcal{A}|^{1+\alpha}Q_{\max}^{2\alpha}\left(\sigma_1^\alpha+\sigma_2^\alpha\right)u^{-2\alpha}.
$$

■

Appendix F. Convolutional MDPs and Assumption A7

In this appendix, we show that Assumption A7 holds for a certain class of MDPs. This class is defined by one dimensional MDPs in which the increment of the next X' compared to the current state X is a function of chosen action only, i.e., $X' - X \sim W(\pi(X))$.

Proposition 25 *Suppose that $\mathcal{X} = [-\pi, \pi]$ is the unit circle and $\mathcal{F}$ is the Sobolev space $\mathcal{W}^k(\mathcal{X}) = \mathcal{W}^{k,2}(\mathcal{X})$ and $J(\cdot)$ is defined as the corresponding norm $\|\cdot\|_{\mathcal{W}^{k,2}}$. For a function $f \in \mathcal{F}$, let $\tilde{f}(n)$ be the n^{th} Fourier coefficient, i.e., $\tilde{f}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-jnx}dx$. Consider the MDPs that have the* convolutional *transition probability kernel, that is, for any policy π and $V \in \mathcal{F}$, there exists $K_\pi(x, y) = K_\pi(x - y)$ such that*

$$\int_{\mathcal{X}} P(dy|x, \pi(x))V(y) = \int_{\mathcal{X}} K_\pi(x - y)V(y)dy = K_\pi * V.$$

Moreover, assume that $K_\pi, V \in L_1(\mathcal{X})$. For a given policy π, let $r^\pi(x) = r(x, \pi(x))$ ($x \in \mathcal{X}$). Assumption A7 is then satisfied with the choice of $L_R = \sup_\pi \|r^\pi\|_{\mathcal{W}^{k,2}}$ and $L_P = \sup_\pi \max_n |\tilde{K}_\pi(n)|$.

Proof By the convolution theorem, $\widetilde{(K_\pi * V)}(n) = \tilde{K}_\pi(n)\,\tilde{V}(n)$. It is also known that for $V \in \mathcal{F}$, we have $\|V\|^2_{\mathcal{W}^{k,2}} = \sum_{n=-\infty}^{\infty} \left(1 + |n|^2\right)^k |\tilde{V}(n)|^2$. Thus,

$$\begin{aligned}\|K_\pi * V\|^2_{\mathcal{W}^{k,2}} = \sum_{n=-\infty}^{\infty} (1 + |n|^2)^k |\tilde{K}_\pi(n)|^2 |\tilde{V}(n)|^2 &\le \left[\max_n |\tilde{K}_\pi(n)|^2\right] \sum_{n=-\infty}^{\infty} (1 + |n|^2)^k |\tilde{V}(n)|^2 \\ &= \left[\max_n |\tilde{K}_\pi(n)|^2\right] \|V\|^2_{\mathcal{W}^{k,2}}.\end{aligned}$$

Therefore, $\|T^\pi V\|_{\mathcal{W}^{k,2}} \le \|r^\pi\|_{\mathcal{W}^{k,2}} + \gamma \left[\max_n |\tilde{K}_\pi(n)|\right] \|V\|_{\mathcal{W}^{k,2}}$. Taking supremum over all policies finishes the proof. ■

The interpretation of $\max_n |\tilde{K}_\pi(n)|$ is the maximum gain of the linear filter K_π that is applied to a value function V. The gain here is explicitly written in the frequency domain.

Appendix G. The Metric Entropy and the Covering Number

Definition 26 (Definition 9.3 of Györfi et al. 2002) *Let $\varepsilon > 0$, $\mathcal{F}$ be a set of real-valued functions defined on $\mathcal{X}$, and $\nu_\mathcal{X}$ be a probability measure on $\mathcal{X}$. Every finite collection of $N_\varepsilon = \{f_1, \dots, f_{N_\varepsilon}\}$ defined on $\mathcal{X}$ with the property that for every $f \in \mathcal{F}$, there is a function $f' \in N_\varepsilon$ such that $\|f - f'\|_{p,\nu_\mathcal{X}} < \varepsilon$ is called an* ε-cover *of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\nu_\mathcal{X}}$. Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu_\mathcal{X}})$ be the size of the smallest ε-cover of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\nu_\mathcal{X}}$. If no finite ε-cover exists, take $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu_\mathcal{X}}) = \infty$. Then $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu_\mathcal{X}})$ is called an* ε-covering number *of $\mathcal{F}$ and $\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{p,\nu_\mathcal{X}})$ is called the* metric entropy *of $\mathcal{F}$ w.r.t. the same norm.*

The ε-covering of $\mathcal{F}$ w.r.t. the supremum norm $\|\cdot\|_\infty$ is denoted by $\mathcal{N}_\infty(\varepsilon, \mathcal{F})$. For $x_{1:n} = (x_1, \dots, x_n) \in \mathcal{X}^n$, one may also define the empirical measure $\nu_{\mathcal{X},n}(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{x_i \in A\}}$ for $A \subset \mathcal{X}$. This leads to the *empirical covering number* of $\mathcal{F}$ w.r.t. the empirical norm $\|\cdot\|_{p,x_{1:n}}$ and is denoted by $\mathcal{N}_p(\varepsilon, \mathcal{F}, x_{1:n})$. If $X_{1:n} = (X_1, \dots, X_n)$ is a sequence of random variables, the covering number $\mathcal{N}_p(\varepsilon, \mathcal{F}, X_{1:n})$ is a random variable too.

References

András Antos, Rémi Munos, and Csaba Szepesvári. Fitted Q-iteration in continuous action-space MDPs. In *Advances in Neural Information Processing Systems (NIPS - 20)*, pages 9–16. MIT Press, 2008a. 33

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008b. 6, 8, 10, 13, 19, 25, 26, 28, 29

Bernardo Ávila Pires and Csaba Szepesvári. Statistical linear estimation with penalized estimators: an application to reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012. 2, 13, 15, 28, 29, 31

Philip Bachman, Amir-massoud Farahmand, and Doina Precup. Sample-based approximate regularization. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *JMLR: W & CP*, 2014. 12

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, pages 30–37. Morgan Kaufmann, 1995. 8

André M.S. Barreto, Doina Precup, and Joelle Pineau. Reinforcement learning using kernel-based stochastic factorization. In *Advances in Neural Information Processing Systems (NIPS - 24)*, pages 720–728. 2011. 2

André M.S. Barreto, Doina Precup, and Joelle Pineau. On-line reinforcement learning using incremental kernel-based stochastic factorization. In *Advances in Neural Information Processing Systems (NIPS - 25)*, pages 1484–1492. 2012. 2

Rick Beatson and Leslie Greengard. A short course on fast multipole methods. In *Wavelets, Multilevel Methods and Elliptic PDEs*, pages 1–37. Oxford University Press, 1997. 32

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research (JMLR)*, 7:2399–2434, 2006. 12

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems (NIPS - 19)*, pages 137–144. MIT Press, 2006. 8

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. 8

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, 1978. 3, 5

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996. 3

Wendelin Böhmer, Steffen Grünewälder, Yun Shen, Marek Musial, and Klaus Obermayer. Construction of approximation spaces for reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 14:2067–2118, 2013. 2

Leon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS - 20)*, pages 161–168. MIT Press, 2008. 32

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996. 9

Corinna Cortes, Mehryar Mohri, and Andres Muñoz Medina. Adaptation algorithm and theory based on generalized discrepancy. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 169–178, 2015. 8

Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research (JMLR)*, 15:809–883, 2014. 31

Ronald A. Devore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998. 12

Paul Doukhan. *Mixing: Properties and Examples*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 1994. 6, 19

Bradley Efron, Trevor Hastie, Iain M. Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. 31

Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine learning (ICML)*, pages 201–208. ACM, 2005. 2, 32

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

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 1999. 12

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS - 24)*, pages 172–180. Curran Associates, Inc., 2011a. 28

Amir-massoud Farahmand. *Regularization in Reinforcement Learning*. PhD thesis, University of Alberta, 2011b. 2, 3, 25, 31

Amir-massoud Farahmand and Doina Precup. Value pursuit iteration. In *Advances in Neural Information Processing Systems (NIPS - 25)*, pages 1349–1357. Curran Associates, Inc., 2012. 2, 3

Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine Learning*, 85(3):299–332, 2011. 19, 32

Amir-massoud Farahmand and Csaba Szepesvári. Regularized least-squares regression: Learning from a β-mixing sequence. *Journal of Statistical Planning and Inference*, 142 (2):493 – 505, 2012. 6, 19, 24

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted Q-iteration: Application to planning. In *Recent Advances in Reinforcement Learning, 8th European Workshop (EWRL)*, volume 5323 of *Lecture Notes in Computer Science*, pages 55–68. Springer, 2008. 31

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 2009a. 2, 31

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration. In *Advances in Neural Information Processing Systems (NIPS - 21)*, pages 441–448. MIT Press, 2009b. 1, 2, 3, 13, 30, 31

Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems (NIPS - 23)*, pages 568–576. 2010. 8, 25, 26, 28, 29

Amir-massoud Farahmand, Doina Precup, Mohammad Ghavamzadeh, and André M.S. Barreto. Classification-based approximate policy iteration. *IEEE Transactions on Automatic Control*, 60(11):2989–2993, November 2015. 33

Matthieu Geist and Bruno Scherrer. ℓ_1-penalized projected Bellman residual. In *Recent Advances in Reinforcement Learning*, volume 7188 of *Lecture Notes in Computer Science*, pages 89–101. Springer Berlin Heidelberg, 2012. 2, 13, 31

Alborz Geramifard, Finale Doshi, Joshua Redding, Nicholas Roy, and Jonathan How. Online discovery of feature dependencies. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 881–888. ACM, 2011. 2

Alborz Geramifard, Thomas J. Walsh, Nicholas Roy, and Jonathan P. How. Batch iFDD: A scalable matching pursuit algorithm for solving MDPs. In *29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013. 3

Mohammad Ghavamzadeh, Alessandro Lazaric, Rémi Munos, and Matthew Hoffman. Finite-sample analysis of Lasso-TD. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1177–1184. ACM, 2011. 2, 13, 28, 30, 31

Grigori K. Golubev and Michael Nussbaum. A risk bound in Sobolev class regression. *The Annals of Statistics*, 18(2):758–778, 1990. 24

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Verlag, New York, 2002. 2, 12, 21, 23, 30, 33, 35, 45, 46, 59

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 1970. 15

Matthew W. Hoffman, Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Regularized least squares temporal difference learning with nested ℓ_1 and ℓ_2 penalization. In *Recent Advances in Reinforcement Learning*, volume 7188 of *Lecture Notes in Computer Science*, pages 102–114. Springer Berlin Heidelberg, 2012. 2, 31

Jeff Johns. *Basis Construction and Utilization for Markov Decision Processes using Graphs*. PhD thesis, University of Massachusetts Amherst, 2010. 3

Jeff Johns, Christopher Painter-Wakefield, and Ronald Parr. Linear complementarity for regularized policy evaluation and improvement. In *Advances in Neural Information Processing Systems (NIPS - 23)*, pages 1009–1017. 2010. 2, 31

Tobias Jung and Daniel Polani. Least squares SVM for least squares TD learning. In *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI)*, pages 499–503, 2006. 2, 31, 32

Michael Kohler. Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference*, 89:1–23, 2000. 45

J. Zico Kolter and Andrew Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 521–528. ACM, 2009. 2, 13, 31

George Konidaris, Sarah Osentoski, and Philip Thomas. Value function approximation in reinforcement learning using the Fourier basis. In *AAAI Conference on Artificial Intelligence*, 2011. 16

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research (JMLR)*, 4:1107–1149, 2003. 8, 9

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research (JMLR)*, 13: 3041–3074, October 2012. 28, 29

Bo Liu, Sridhar Mahadevan, and Ji Liu. Regularized off-policy TD-learning. In *Advances in Neural Information Processing Systems (NIPS - 25)*, pages 845–853, 2012. 32

Manuel Loth, Manuel Davy, and Philippe Preux. Sparse temporal difference learning using LASSO. In *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 352–359, 2007. 2, 31

Sridhar Mahadevan and Bo Liu. Basis construction from power series expansions of value functions. In *Advances in Neural Information Processing Systems (NIPS - 23)*, pages 1540–1548. 2010. 2

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 (JMLR)*, 8:2169–2231, 2007. 2

Stéphane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. 3

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009. 8

Mahdi Milani Fard, Yuri Grinberg, Amir-massoud Farahmand, Joelle Pineau, and Doina Precup. Bellman error based feature generation using random projections on sparse spaces. In *Advances in Neural Information Processing Systems (NIPS - 26)*, pages 3030–3038, 2013. 2

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 560–567, 2003. 8

Rémi Munos. Performance bounds in L_p norm for approximate value iteration. *SIAM Journal on Control and Optimization*, pages 541–561, 2007. 25, 26

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems (NIPS - 22)*, pages 1330–1338, 2009. 14

Michael Nussbaum. Spline smoothing in regression models and asymptotic efficiency in L_2. *Annals of Statistics*, 13(3):984–997, 1985. 24

Michael Nussbaum. Minimax risk: Pinsker bound. In *Encyclopedia of Statistical Sciences*, volume 3, pages 451–460. 1999. 23, 24

Dirk Ormoneit and Saunak Sen. Kernel-based reinforcement learning. *Machine Learning*, 49:161–178, 2002. 2

Christopher Painter-Wakefield and Ronald Parr. Greedy algorithms for sparse reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012. 3

Ronald Parr, Christopher Painter-Wakefield, Lihong Li, and Michael Littman. Analyzing feature generation for value-function approximation. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 737 – 744. ACM, 2007. 2

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. 3

Marek Petrik. An analysis of Laplacian methods for value function approximation in MDPs. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2574–2579, 2007. 2

Zhiwei Qin, Weichang Li, and Firdaus Janoos. Sparse reinforcement learning via convex optimization. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *JMLR: W & CP*, 2014. 32

Stéphane Ross, Geoffrey Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 627–635, 2011. 19

Paul-Marie Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000. 19

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *COLT '01/EuroCOLT '01: Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, pages 416–426. Springer-Verlag, 2001. 17, 34

Paul J. Schweitzer and Abraham Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110: 568–582, 1985. 8

Shai Shalev-Shwartz and Nathan Srebro. SVM optimization: Inverse dependence on training set size. In *Proceedings of the 25th international conference on Machine learning (ICML)*, pages 928–935. ACM, 2008. 32

Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(1):17–41, 2003. 21

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008. 13, 20, 21, 23, 30

Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *in Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009. 24

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053, 1982. 23

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998. 3

Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 993–1000. ACM, 2009. 13

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan Claypool Publishers, 2010. 3

Gavin Taylor and Ronald Parr. Kernelized value function approximation for reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1017–1024. ACM, 2009. 2, 31

Hans Triebel. *Theory of Function Spaces III*. Springer, 2006. 12

Alexander B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. 23, 24

Sara A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000. 12, 19, 20, 23, 24, 30, 33, 36, 39, 48, 49, 51, 53, 55, 57

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, chapter 5, pages 210–268. 2012. 19

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2007. 2

Ronald J. Williams and Leemon C. Baird. Tight performance bounds on greedy policies based on imperfect value functions. In *Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems*, 1994. 8

Xin Xu, Dewen Hu, and Xicheng Lu. Kernel-based least squares policy iteration for reinforcement learning. *IEEE Transactions on Neural Networks*, 18:973–992, 2007. 31, 32

Changjiang Yang, Ramani Duraiswami, and Larry Davis. Efficient kernel machines using the improved fast Gauss transform. In *Advances in Neural Information Processing Systems (NIPS - 17)*, pages 1561–1568. MIT Press, 2004. 32

Yuhong Yang and Andrew R. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. 23

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, January 1994. 6, 19

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research (JMLR)*, 2(527 – 550), 2002. 30

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3): 739–767, 2002. 20

Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49:1743–1752, 2003. 20